The role of telomere length and associated polymorphisms in type 2 diabetes in sub-Saharan African populations

Sebentile Hleli Mthimkulu

Student number: 1521476

Supervisor: Professor Michèle Ramsay

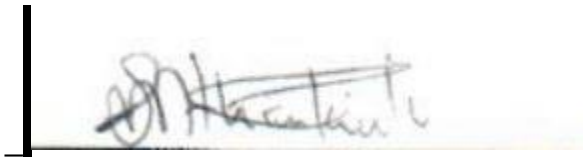
Co-supervisor: Ms Cassandra Soo

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

November 2019

Declaration

I Sebentile Hleli Mthimkulu declare that this Dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



_____

(Signature of candidate)

21st day of November 2019 in Johannesburg

Dedications

This is for my late grandmother, who ultimately believed in me. To my mother, father and brother for your unwavering support and love. To my cousin and aunt who would always cheer me on when all seemed bleak and most importantly God, who made it all possible.

Abstract

Telomeres are the capping repeats of eukaryotic chromosomes, which function in maintaining genome stability. With each successive cell replication, telomeres shorten until a critical threshold is reached and the cell undergoes apoptosis. States of heightened inflammation and oxidative stress accelerate the rate of telomere attrition as has been observed in non-communicable diseases such as type 2 diabetes. The prevalence of type 2 diabetes is rising, especially in low-to-middle income countries causing a significant health burden. Type 2 diabetes has previously been associated with a decrease in telomere length. The aim of this project was to determine the association between telomere length and type 2 diabetes in sub-Saharan African populations, and to explore the association of known telomere-associated variants in these African populations. A matched case-control association approach was used. Black male and female African adults, aged 40 years and above were studied: 836 (418 diabetic and 418 non-diabetic) for the telomere assay, and 770 (385 cases and 385 controls) for the genomic association study. A quantitative Real-Time polymerase chain reaction assay was used to measure telomere length. PLINK was used to explore the association of 28 previously identified SNP variants associated with telomere length between our selected cases and controls with age, sex, and BMI status, as potential confounders. The telomere length assay proved problematic as it resulted in high assay variability, non-specific amplification, primer-dimer formation, contamination of the “no sample negative control”, and non-amplification. In the genetic association study, none of the 28 variants previously associated with telomere length were associated with type 2 diabetes in this study.

Acknowledgements

To my supervisors Prof Michèle Ramsay and Ms Cassandra Soo, for believing in me even when I could not. For all the tireless hours you put in for this project to be completed, I am forever grateful.

To my friends Senanile Dlamini, Hlengiwe Nkhambule, Lombe Kampamba, Busisiwe Mthembu, Romould Boua, and Evans Mathebula for your unwavering support.

The SBIMB family for everything.

Lastly to the National Research Foundation (NRF), Prof Michèle Ramsay, and Ms Cassandra Soo for funding.

Table of Contents

Abbreviations

36B4	Acidic ribosomal phosphoprotein PO
AWI-Gen	Africa Wits INDEPTH partnership for genomic studies
BMI	Body Mass Index
BP	Blood pressure
CHD	Coronary heart disease
CI	Confidence interval
CRP	C-reactive protein
CVD	Cardiovascular disease
DMSO	Dimethylsulfoxide
DTT	Dithiothreitol
EDTA	Ethylenediaminetetra-acetic acid
GAPDH	Glyceraldehydes-3-phosphate dehydrogenase
GWAS	Genome-wide Association Study
H3Africa	Human Heredity and Health in Africa
HRM	High resolution melt
IDF	International Diabetes Federation
INDEPTH	International Network for the Demographic Evaluation of Populations and their Health in low- and middle-income countries
LD	Linkage disequilibrium
LTL	Leukocyte telomere length

LMICs	Low to middle income countries
MAF	Minor allele frequency
MMQPCR	Multiplex monochrome quantitative PCR
MR	Mendelian Randomisation
NCDs	Non-communicable diseases
NTC	No Template Control
OR	Odds Ratio
P	P-value
PCR	Polymerase Chain Reaction
QC	Quality control
qPCR	Quantitative polymerase chain reaction
ROS	Reactive oxygen species
RT-PCR	Real-time polymerase chain reaction
SCG	Single copy gene
SD	Standard Deviation
SE	Standard error
SNP	Single Nucleotide Polymorphism
SSA	Sub-Saharan Africa
T/S ratio	Relative concentration of Telomere repeats / relative concentration of Single copy gene copies
T2D	Type 2 diabetes
T1D	Type 1 diabetes

TE	Tris-EDTA
TERC	Telomerase RNA component
TERT	Telomerase reverse transcriptase
TL	Telomere Length
TRF	Terminal restriction fragment
TRF1	Telomeric repeat-binding factor 1
TRF2	Telomeric repeat-binding factor 2
WC	Waist Circumference
WHO	World Health Organisation
GC	Guanine-Cytosine

List of Figures

List of Tables

Chapter 1: Introduction and literature review

Over the last century, global life expectancy has substantively increased (Lunenfeld and Stratton, 2013). Although populations are longer lived, this increase in longevity has not resulted in an increase of health span along with this population greying (Vaiserman and Lushchack, 2017). The increase in the number of the aged within populations has resulted in an upsurge of chronic, non-communicable, and age-related disorders like type 2 diabetes (T2D), as increased age is the primary risk factor for most non-communicable diseases (NCDs) (Krasnienkov *et al.,* 2018). NCDs have a complex aetiology, and are influenced by both genetic variation and the environment (Vaiserman and Lushchack, 2017). Infectious diseases; like HIV/AIDS and tuberculosis are currently the leading causes of death in Africa, however it is predicted that NCDs in Africa will surpass the associated premature mortality of infectious diseases (Kushitor *et al.,* 2018). This change can largely be attributed to improved healthcare and treatment, and to transitioning economies in low-to-middle income countries (LMICs) (Pillay-van Wyk *et al.,* 2016). With the increase of NCDs in sub-Saharan Africa (SSA) the financial burden on the economy will also increase, especially in poverty-stricken rural communities, which are experiencing a complex epidemiological and demographic transition characterised by high population growth coupled with population greying (Thorogood *et al*., 2014; Pillay-van Wyk *et al.,* 2016).

T2D has become a global epidemic with 9% of adults (415 million people) estimated to be living with this condition (Jaacks *et al.,* 2016). Its prevalence is expected to increase to 642 million in the next decade, with the majority of this increase expected to occur amongst those residing in LMICs (Chatterjee *et al.,* 2017). The risk factors involved in the onset of T2D include: an unhealthy diet, sedentary lifestyle, and genetic susceptibility (Silva *et al.,* 2017). Impaired beta-cell function, insulin insensitivity, insulin resistance, and derailed glucose metabolism, are all pathophysiological traits of T2D (Skyler *et al.,* 2017). T2D is characterised by hyperglycaemia which results in states of increased oxidative stress and heightened inflammation (Tangvarasittichai *et al.,* 2015; Tamura *et al*., 2016). There is evidence that oxidative stress affects telomere length (TL) through the double-stranded breaks

that occur in the fragile Guanine-Cytosine (GC) complexes, which are characteristic of telomeres (Salpea and Humphries, 2010).

Telomeres are the capping ends of eukaryotic chromosomes characterised by TTAGGG hexamer repeats (Pillay-van Wyk *et al*., 2016). They protect the genome from chromosome fusions and breaks with associated nucleoproteins termed the "Shelterin complex" (Greider and Blackburn, 1989) Telomeres shorten with each successive cell replication due to a phenomenon called "the end replication problem" (Hayflick, 1998). This shortening of the telomeres leads to cell senescence, and eventually cell death, which is the main mechanism for biological ageing (Hayflick, 1998). Telomeres are characterised by regions rich in GC complexes, making them more vulnerable to DNA damage (Tzanetakoua *et al.,* 2012). This vulnerability is augmented by oxidative stress and inflammation in T2D affected individuals, which is suspected to accelerate telomere attrition (Tamura *et al.,* 2016a). These heightened stressful conditions caused by hyperglycaemia in diabetes may be the key components in the pathophysiology of T2D (Tamura *et al.,* 2016a). Many epidemiological studies have associated shorter TL with T2D, onset of T2D, development of insulin resistance, pre-diabetic conditions, and glycated haemoglobin levels (Adaikalakoteswari *et al.,* 2007; Salpea *et al.,* 2010; Zhao *et al.,* 2014; Tamura et al., 2016; Verhulst *et al.,* 2016).

As a complex trait, T2D also has a genetic component to its aetiology. Various genome-wide association studies (GWAS) have discovered variants that predispose one to diabetes, as well as variants that are associated with TL and T2D (Zee *et al*., 2011; Hara *et al.,* 2014). GWAS performed to elucidate the association between T2D and TL have found various significant associations with single nucleotide polymorphisms (SNPs), particularly in the genes implicated in telomere biology (Codd *et al.,* 2010; Levy *et al.,* 2010; Lee *et al.,* 2014;). One association in the *TERC* gene, rs12696304 was found to be associated with T2D in a Chinese population (Delgado *et al.,* 2018). However, the majority of SNPs associated with T2D are common variants, each contributing a small effect size to disease susceptibility (Hara *et al.,* 2014). Variants associated with TL are usually found in, or near, genes that encode proteins involved in; telomere maintenance, cell turnover, embryonic development, apoptosis, transcription, and cell cycle progression (Gorenjak *et al.,*

2018). TL varies in different populations; studies have shown that individuals of African descent, on average, have longer telomeres when compared to individuals from other populations (Hunt *et al.,* 2008; Zhu *et al.,* 2011; Needham *et al.,* 2013). Little research has been conducted on the association of TL with T2D and the underlying genetic variants in an African setting (Diez Roux *et al.,* 2009). Belonging to a certain race, and ethnic group, may predispose one to different exposures that could result in longer or shorter TL, and these differences may affect one's susceptibility to disease (Zeiger *et al.,* 2018). With the changing health dynamics in SSA, further investigations into these chronic conditions are crucial for the improvement of health.

In SSA the largest cardiometabolic studies were conducted by the Human Heredity and Health in Africa (H3Africa) consortium (Ramsay and Sankoh, 2016). The Africa Wits-INDEPTH partnership for genomic studies (AWI-Gen) falls under the H3Africa consortium. This Masters project was one of the AWI-Gen sub-studies. The AWI-Gen study enrolled over 10 000 individuals from four different countries (South Africa, Kenya, Ghana, and Burkina-Faso). The aim was to elucidate the genetic variants that drive the progression of various NCDs, such as T2D, hypertension, atherosclerosis, obesity, hypercholesterolemia, and others (Ali *et al.,* 2018). Genetic diversity across different African population groups was also explored, so as to elucidate the differences in disease progression in different African ethnic groups (Ramsay and Sankoh, 2016). Participants from this study were chosen from diabetic participants of the AWI-Gen study, from the six different study sites (Agincourt, Dikgale, Soweto (South Africa), Nairobi (Kenya), Nanoro (Burkina-Faso) and Navrongo (Ghana)) (Ali *et al.,* 2018).

1.1 Diabetes

Diabetes is defined as the failure of beta pancreatic cells to secrete adequate insulin to fulfil the body's physiological needs (Elmarakby *et al*., 2007). As an endocrine disease, diabetes is distinguished into two main types; diabetes mellitus type 1 and diabetes mellitus type 2 (Cole *et al*., 2016). Diabetes is mainly characterised by hyperglycaemia (a higher than normal concentration of fasting glucose defined as more than 7 mmol/l or more than 11 mmol/l after a meal, according to the World Health Organisation (WHO) guidelines,(www.who.int) and glucose intolerance

(metabolic conditions which result in hyperglycaemia) (Salpea and Humphries, 2010). T2D is often accompanied by comorbidities such as hypertension and abnormal lipoprotein metabolism (Verges *et al*., 2015).

1.1.1 Type 1 diabetes

Type 1 diabetes mellitus (T1D) is less prevalent than T2D and it has a higher incidence in younger individuals, as the age of onset is much earlier than that of T2D (Paschou *et al.,* 2018). Thus it is often referred to as juvenile, auto-immune, or insulin-mediated diabetes, which is induced by beta-cell failure due to an auto-immune reaction resulting in a destruction of beta pancreatic cells. This leads to reduced insulin production and patients usually require exogenous insulin replacement as treatment (Atkinson *et al.,* 2012). Like other NCDs, T1D has genetic, environmental, epigenetic, and immunologic risk factors. T1D is considered to be more severe than T2D, and has a larger genetic component (Paschou *et al.,* 2018). There has recently been an exponential increase of T1D prevalence and incidence globally, as is seen with the other NCDs (Atkinson *et al.,* 2012). This may be due to new gene-environment interactions, like the change in lifestyle and diets to sedentary and calorie-high diets. In SSA countries, T1D has not benefitted as much as T2D from medical advancements and research, as focus and policies of eradication seem to focus on the more common type of diabetes. With the increase in prevalence of T1D, steps need to be taken in order to motivate that further research be conducted to better enhance knowledge and possibly provide public health intervention and advice in this regard (Bahendeka, 2017).

1.1.2 T2D

T2D is often referred to as: adult-onset or non-insulin dependent diabetes, for the reason that age of onset is usually 40 years and above. However, cases below this age were rare but are increasing (IDF, 2016). Treatment does not entirely depend on exogenous administration of insulin (Olokoba *et al*., 2012), but is usually treated and controlled by practising a healthy lifestyle (weight loss, adequate exercise, and healthy diet) rather than with glucose lowering drugs (De Fronzo and Tripathy, 2009). Hyperglycaemia in T2D is largely due to beta-cell failure and insulin receptor insensitivity on target tissue cells (Kahn *et al.,* 2014). The initiation and progression

of T2D is attributed to beta-cell dysfunction which is brought about by an excess of glucose and fatty acids in the body (Poitout *et al.,* 2010). States of heightened glucose and fatty acids increase the metabolic rate of mitochondria, which in turn increases the concentration of reactive oxygen species (ROS) produced by the cells and eventually leads to protein misfolding of the insulin receptors (Muoio and Newgard, 2008).

1.1.3 Pathogenesis of T2D

Insulin resistance and beta-cell dysfunction are both crucial mechanisms in the pathophysiology of T2D (DeFronzo and Tripathy, 2009). Beta-cell dysfunction being more detrimental than insulin resistance as it occurs in the later stages of disease progression (Khodabandehloo *et al.,* 2016). Insulin resistance has been observed to precede beta-cell failure in the onset of T2D (Kasuga, 2006). However, insulin resistance on its own cannot initiate diabetes, both the sensitivity and secretion of the hormone need to be derailed (DeFronzo and Tripathy, 2009). Insulin resistance is defined as the lowered response of insulin to target tissues (muscle, liver and adipose) to effectively react to circulating concentrations of glucose (Khodabandehloo *et al.,* 2016). Insulin resistance initiates first in skeletal muscle, and thereafter progresses to other tissues (Muoio and Newgard, 2008). In skeletal muscle cells, the insulin signal prompts the uptake of glucose for glycogen synthesis after the ingestion of food. Insulin resistance is molecularly triggered by high levels of glucose, free fatty acids, and cytokines (De Fronzo and Tripathy, 2009).

Beta-cell failure is defined by the reduction in number of beta pancreatic cells accompanied by a decrease in function of the beta-cells to produce adequate insulin for target tissues (Muoio and Newgard, 2008). The mechanisms underlying beta-cell failure are not clearly understood, and existing research proposes that glycolipotoxity, which is the state of heightened free fatty acids in combination with elevated glucose levels, is the predominant mechanism that results in failure of beta pancreatic cells (Poitout *et al.,* 2010).

1.1.4 Risk factors

T2D risk factors may be categorised into three distinct groups: demographic, behavioural, and environmental factors (Jaacks *et al.,* 2016). The most common

predisposers are demographic and behavioural factors. These are age, obesity, a sedentary lifestyle, and a high calorie diet (Silva *et al.,* 2017). Factors which significantly increase susceptibility are; advanced age, obesity, ethnicity (non-white), family history of T2D, smoking, and a high calorie diet (Jaacks *et al.,* 2016). Recent evidence has found that a nutrient deficit and an excessively nutrient rich diet in-utero can predispose one to T2D in adulthood (Salpea and Humphries, 2010).

1.1.5 Genetics of T2D

T2D is a multifactorial trait resulting from the interplay between environmental and genetic components (Van Tilburg *et al.,* 2001). Genetic susceptibility of T2D is progressive and develops with time. Evolutionary processes drive most of these changes, which result from environmental changes through time. The thrifty gene hypothesis states, that certain genetic variations have occurred in populations as a means to adapt to certain environmental conditions like instances where there is a shortage of food supply, but when circumstances change and there is a surplus of nutrients, these variants now influence insulin sensitivity and resistance to promote the progression of diabetes (Zeggini *et al.,* 2007). There have been a plethora of studies that have been conducted on the genetic susceptibility of T2D. However, few genetic variants have been discovered but they have added to the knowledge of T2D genetics. These genetic variants were discovered through family studies, candidate gene approaches, and linkage analysis; it was not only until the era of GWAS, that more genetic variants were discovered. Although more variants have been discovered, they confer minor effects to disease pathophysiology and are generally small effect variants which when accompanied by behavioural risk factors result in disease progression (McCarthy and Zeggini, 2009). There is also the case of ‘missing’ heritability, where one individual has inherited from parents the genes predisposing T2D but does not express the disease; hence the heritability component is lost through conception, as it is with most NCDs (Salpea and Humphries, 2010). T2D-associated variants have diverse physiological functions. They are involved in beta-cell function, insulin secretion, and cell cycle maintenance (Zeggini *et al*., 2007; McCarthy and Zeggini, 2009; Bonnefond *et al.,* 2010). In order to better understand T2D onset, progression, and pathogenesis, more research into

copy number variants, SNPs, epigenetic modifications, and TL variation needs to be conducted (Bonnefond *et al.,* 2010; Salpea and Humphries, 2010).

1.1.6 Complications of T2D

T2D is a chronic disease resulting in various complications during its course. These complications have the greatest impact on patients and healthcare providers as they are difficult to treat and may lead to premature death (Skyler *et al.,* 2017). T2D complications are divided into macrovascular and microvascular complications. Microvascular complications include nephropathy, retinopathy, and neuropathy (Chatterjee *et al.,* 2017). Macrovascular complications lead to organ damage and are usually the cause of death in patients. These are cerebrovascular diseases, ischaemic heart disease, and peripheral artery disease (Papa *et al.,* 2013). Complications of T2D do not manifest when the glucose level in patients is controlled, as they are driven by hyperglycaemia (Skyler *et al.,* 2017). Patients usually present for treatment when the symptoms have advanced, and molecularly the disease has advanced such that the effects cannot be reversed (Chatterjee *et al.,* 2017).

1.2 Telomere biology

Telomeres are made of repetitive TTAGGG sequences found at both ends of all chromosomes. In humans their length may reach 15,000 base pairs, ending with a single-stranded overhang on the 3’ end which ranges from 75 to 200 base pairs in length as shown in Figure 1. The telomeres form a complex with proteins such as the enzyme telomerase, which functions in elongating the telomeric sequence, and other telomere binding proteins to ultimately form the shelterin complex. In this complex, telomeres form two loop structures called the T-loop and the D-loop. The T-loop is double-stranded while the D-loop is single-stranded. These loops protect the telomeric sequences from fusing with each other, and also from the DNA repair mechanisms that would recognise telomeres as double-stranded DNA breaks and attempt to correct them (Butt *et al.,* 2010).

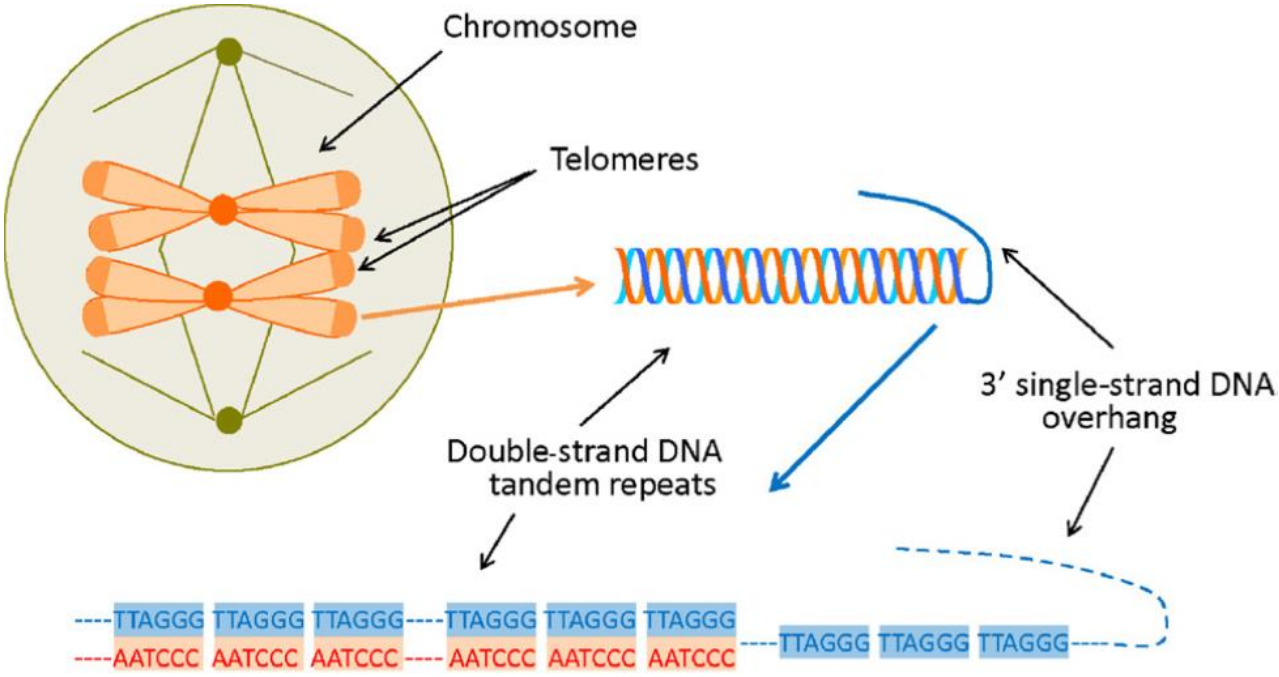


Figure 1.1: Telomere structure (Salpea and Humphries, 2010)

Telomeres play a vital role in genome stability by preventing chromosome end-to-end fusions and chromosome erosions (Sanders and Newman, 2013). Telomeres shorten with each successive cell replication due to a trend known as the "end replication problem" where DNA polymerase does not fully replicate the 5' lagging strand of the template due to the RNA primer preceding the enzyme (Blackburn *et al.,* 2015). ROS and inflammation also contribute to the attrition of telomeres. Oxidative stress has been shown to affect human health, having an impact on physical health resulting in greater telomere loss per cell division in humans (Harris *et al.,* 2012). With each successive cell replication approximately 30-100 nucleotides of telomeric DNA is lost due to the end replication problem (Zee *et al.,* 2010). TL is extremely variable between individuals, chromosomes in the same cell, tissues in one organism, between the sexes, and in individuals of the same age (Factor-litvak *et al.,* 2016). However, at birth TL appears to be more consistent suggesting that these differences are acquired later in life (Calado *et al.,* 2009). There is a strong association between TL in fathers and their children, suggesting that TL is paternally inherited (Butt *et al.,* 2010). Telomeres are typically longer in females, and the male telomeric attrition rate is higher than that of the female attrition rate (Factor-litvak *et al.,* 2016). Studies have found that the hormone oestrogen confers a certain

protective effect against telomere attrition (Cawthon, 2002; Calado *et al.,* 2009; Butt *et al.,* 2010).

The heterogeneity of TL in somatic cells was proved by Elizabeth Blackburn and Carol Greider, when they discovered the enzyme telomerase that elongates telomeres within cells. The discovery of telomerase was made in the ciliated protozoa *Tetrahymena thermophile*, where it was seen that telomerase elongates the 3” end of chromosomes by the addition of specialised telomere hexamer repeats cells. The telomerase enzyme is a ribonucleoprotein complex that consists of two core proteins, the telomerase reverse transcriptase (*TERT*) and the telomerase RNA component (*TERC*). However, telomerase is only up-regulated in certain highly proliferating cells including the germ cells, and down-regulated in differentiated somatic cells which consequently cannot regenerate TL. In contrast, telomerase is highly active in cancer cells (Greider and Blackburn, 1985; Greider and Blackburn, 1989). Studies have found that longer telomeres are a characteristic of cancerous cells as they have the ability to indefinitely elongate their telomeres with the aid of telomerase, hence longer telomeres may indicate a predisposition to certain types of cancers (Hunt *et al.,* 2008).

Ethnic differences in TL have also been observed in epidemiological studies comparing TL in individuals from different ethnic groups. A study by Hasen *et al.,*2016, found that TL in SSAs was significantly shorter than that of individuals of European descent and African Americans, the difference in base pairs was on average 0.6 kb shorter in Europeans ($P < 0.0001$) and 0.4 kb shorter in African Americans ($P <0.0001$) (Hansen *et al.,* 2016). The reason for these differences is unknown, and Hasen and colleagues have suggested that this may be due to polygenetic adaptation. Based on the hypothesis that melanoma is more prevalent in populations of European descent than populations of African descent and that evolutionary influences have maintained the long telomere in Africans as they were less likely to progress to the cancer of the skin. It is well-known that long telomeres predispose one to cancer and thus in Europeans evolution influenced the shortening of telomeres as a protective mechanism for melanoma (Hansen *et al.,* 2016).

In contrast a study conducted in South Africa, on black and white teachers ($n = 161$ and 180, respectively), found that white teachers had longer TLs than black teachers

(P <0.001) (Von Kanel *et al.,* 2015). The authors suggested that the difference could be due to the different genetic determinants in the two groups brought about by ethnic differences as Africans have more genetic diversity when compared to individuals of European descent (Von Kanel *et al.,* 2015).

1.2.1 Genetics of TL

The consistent average TL differences between species, and chromosome arm differences within species, indicate that mean TL is genetically determined (de Pauw *et al*. 2005). Telomeres have chromosome-specific lengths that are very similar between homologous chromosomes, but are different among individuals. It has been suggested that this chromosome-specific pattern of TLs is determined in the zygote. Twin and family studies have provided evidence showing that heritability of the inter-individual variation in TL ranges from 44% to 80% (Slagboom *et al*. 1994; Vasa-Nicotera *et al*., 2005; Njajou *et al*., 2009).

1.2.2 Heritability of TL

TL is a highly heritable trait, when either parents have relatively long and short telomeres, the offspring would have TL similar to that of the parents (Tedone *et al.,* 2014). One study found that offspring of long-lived individuals who had long telomeres also tend to have long telomeres (Njajou *et al.,* 2009). A study by Honig and colleagues on long-lived families revealed that TL has a heritability (h^2) of 0.54 (Honig *et al.,* 2015). Various studies found differences in TL when it came to sex, and ethnicity (Zhu *et al.,* 2011). TL is longer in women, but this may be attributed to the differences in hormone levels (Barrett and Richardson, 2011; Zhu *et al.,* 2011; Shaffer *et al.,* 2012; Honing et al., 2015). The heritability of TL differs from study to study; some studies found that there is a greater degree of heritability from the father (Andrew *et al.,* 2006; Deelen *et al.,* 2011). Others found that the mother had a higher degree of heritability (Codd *et al.,* 2010; Soerensen *et al.,* 2012). The high heritability however appears to remain constant, regardless of the numeric heritability index, or whether the mother or father confers a greater degree of heritability.

1.2.3 Variants associated with TL

Recent GWAS have identified SNPs affecting TL on chromosome 18q12.2 (Mangino *et al*., 2009) and chromosome 3q26 at a locus which includes *TERC*, the gene encoding the telomerase RNA component (Codd *et al*., 2010). Most studies on genetic variants discovered in TL GWAS are predominant in Caucasian and Asian populations. All but one GWAS investigating TL was performed in African Americans (Zeiger at al., 2018).The SNPs and genes discovered in TL GWAS mostly function in telomere maintenance, and to some extent attempt to explain the inter-individual variability of TL (Gorenjak *et al.,* 2018). Moreover, only a small portion of TL genetic variability has been discovered, the variance obtained from associated SNPs is approximately 1%, for each locus (Levy *et al.,* 2010; Mangino *et al.,* 2012; Codd et al., 2013; Liu *et al.,* 2014). Appendix C shows the TL-associated SNPs that have been discovered, and reached genome-wide significance with the role of the gene and SNP biological significance included. Biological significance is ambiguous in most studies, and the underlying disease is not clear in some. Appendix C depicts genes and SNPs that reached genome-wide significance in various studies. The studies evaluate the association of TL with various cardiometabolic conditions and cancers. The Table lists the gene, its function, the SNP identifier, and the association found together with the biological significance of the genetic variants from the respective studies.

1.2.4 Methods for measuring TL

With the increase in telomere research, several methods have been devised to measure TL. These methods include different molecular techniques such as Southern blotting, fluorescent *in situ* hybridisation (FISH), flow cytometry, and polymerase chain reaction (PCR). Moreover, these methods quantify different aspects of TL in cell populations and chromosomes (Montpetit *et al*., 2015).

All the available TL measurement methods have their varying advantages and limitations. When selecting a method to use in a study, the type of sample, quantity, and quality of the sample, the number of samples to be measured, how accurate the measurement needs to be, and the limit of detection, should all be considered (Lin *et al*., 2018). Furthermore, the study design, research question, and available

resources, dictate which method a researcher should choose for their specific study. Currently no available method can absolutely measure TL accurately and in a high-throughput setting (Nussey *et al*., 2014).

Terminal restriction fragment analysis (TRF) is considered the gold standard for TL measurement, and it utilizes Southern blotting as a molecular technique (Kimura *et al*., 2010). This method involves the digestion of genomic DNA by using a mixture of restriction enzymes that have no restriction sites within the telomeric region. Thereafter, the intact telomere strands are resolved on an agarose gel, based on their size. Telomere-specific probes are then used to visualize telomeric fragments by either Southern blotting, or in-gel hybridisation. A DNA ladder is used to compare the sizes of the varying telomeres that present as a smear (Kimura *et al*., 2010). The TRF method is laborious, time-consuming, and is not high-throughput, thus is limited for use in large epidemiological studies. Moreover, it requires large quantities of DNA (0.5 – 10 µg per sample) (Nussey *et al*., 2014).

Quantitative fluorescent *in-situ* hybridisation (Q-FISH), on the other hand requires the use of replicating cells at the metaphase stage of mitosis in order to measure telomeric repeats from individual chromosomes (Montpetit *et al*., 2015). This method is highly sensitive and uses labelled oligonucleotide probes, which hybridise to sequences in telomere repeats. This method cannot be used for high-throughput epidemiological studies. Furthermore, it is labour intensive, requires specialised equipment and is time-consuming, requiring a week to process only 10 samples (Vera and Blasco, 2012).

Flow-FISH is another TL measurement method that combines flow cytometry and FISH, resulting in a medium-throughput method that measures the average population of cells. The principle of FISH and flow-FISH is similar, the only difference being that instead of fluorescent microscopy; flow cytometry detection and analysis are used in flow-FISH. Additionally flow-FISH is the only method that is used clinically for the diagnosis of *dyskeratosis congenita* (Alter *et al*., 2007).

Currently the only high-throughput method available to measure TL is the quantitative polymerase chain reaction (qPCR). It measures relative average TL and is the preferred method for large epidemiological studies. This method is the

extension of the traditional PCR where fluorescent dyes enable the relative quantification of target DNA. The measurement of TL by the PCR technique, was first achieved by Cawthon in 2002 where he designed specific mis-matched primers that could align to complementary telomere repeats and not self-anneal. The amplification of the telomere repeats is then compared to the amplification of a single copy gene, which is then used to create a ratio which would represent the relative average TL for that particular sample. In addition, this method uses small quantities of DNA (less than 50 ng of DNA) (Cawthon *et al*., 2002). The greatest limitation of the qPCR method is amplification efficiencies of the target DNA. Small changes in efficiency in earlier cycles result in large changes in subsequent cycles. This may be caused by a large number of things; such as the mis-priming of primers, changes in thermal cycling temperatures, and reagents creating bias in the results. Moreover, the fluorescent dye used in the qPCR method (SYBR Green) has been demonstrated to produce artificial results due to its non-specific nature (Jodczyk, 2011).

1.3 TL in cardiovascular disease and T2D

1.3.1 TL in cardiovascular disease

Vascular ageing begins at the cellular level, when the balance between senescent cells and vascular progenitor cells is tipped, such that there is a decrease in progenitor cells and an increase in senescent cells, which in turn decreases the vascular tissues capacity for regeneration (Erusalimsky, 2009). Vascular endothelial cells that have become senescent have alterations with regard to gene expression, resulting in impaired cell function, resulting in atherogenesis (Samani *et al*., 2001).

Telomere attrition initiates the cascade for atherogenesis, as it has been shown that in vascular endothelial cells that are susceptible to atherosclerosis, there is marked telomere shortening (Allsopp *et al.,* 1995). As a potential biomarker of ageing, telomere attrition has been shown to play a role in atherosclerotic plaque formation and this ultimately leads to cardiovascular disease (CVD) development (Minamino *et al.,* 2002).

Studies investigating TL, longevity, and atherosclerosis, have found that individuals who are deemed healthy and long-lived have longer TLs than individuals who have

arterial stiffness and atherosclerosis (Samani *et al.,* 2001 and Terry *et al.,* 2012). Other degenerative disorders such as Alzheimer's, dementia, and stroke, also have marked telomere attrition, proving that telomere attrition is a common factor in the pathogenesis of many ageing disorders (Benetos *et al.,* 2001 and Honig *et al.,* 2006). When compared to healthy individuals, atherosclerotic patients were found to have TLs 11 years older than healthy individuals, indicative of vascular wall ageing (Brouilette *et al.,* 2003).

1.3.2 TL in T2D

Obesity causes a state characterised by low-grade systemic inflammation and increased oxidative stress which have been found to further decrease TL (Garcı´a-Calzo´n *et al.,* 2014). A study by Laimer *et al.,* 2015, confirmed that obesity may accelerate ageing as participants with a high BMI and high waist-to-hip ratio (WHR) had significantly shorter telomeres when compared to their healthy counterparts. Longitudinal studies, some in the form of interventions, have been conducted to track the attrition rate of telomeres in diseased and healthy people. Contrasting results were found by different studies. In some, TL attrition rate differed between the two groups, whereas others found no difference at all. In the health intervention groups, it was found that TL increased in some, and not in others. Reasons for these discordant results may be due to the method of telomere measurement across studies, the challenges in accurately measuring average TL, the sample size of the study, and the time from baseline in the longitudinal studies. Most studies found significant associations between obesity and shortened TL (Mundstock *et al.,* 2015). The systemic review conducted included 63 association studies, of these 24 did not find a significant association between TL and obesity (Mundstock *et al.,* 2015).

The same mechanism of action in telomere attrition was present in diabetic patients, the only difference being that in diabetes this is tissue-specific, showing more specific and aggressive telomere shortening in pancreatic beta-cells. Telomere attrition has also been associated with comorbid complications of diabetes such as diabetic nephropathy, microalbuminuria, and epithelial cancers (Sampson *et al*., 2006; Verzola *et al.,* 2008; Tentolouris *et al.,* 2009; Salpea and Humphries, 2010).

1.4 Mendelian randomisation studies and TL

Mendelian randomisation (MR) is a recent tool used in epidemiological studies, for underpinning causality of disease (Smith *et al.,* 2006). The concept lies in the use of genetic variants as instruments for modifiable risk factors (Zhan *et al.,* 2015). MR is an advanced step from observational studies that infers causality of disease more precisely (Burgess *et al.,* 2012). In telomere MR studies, associations with diseases have not found concrete causal genetic variants (Zhan *et al.,* 2015). The genetic variants which serve as proxies for the risk factor also function in randomising the study as it is seen with clinical trials (Burgess *et al.,* 2012). This concept is based on the rule of random assortment of genes during conception (Allman *et al.,* 2018). A study that investigated the association of C-reactive protein (CRP) with coronary heart disease (CHD) found that the genetic variant rs11206510 is associated with low-density lipoprotein (LDL) cholesterol concentration and CHD; hence the association between CRP and CHD which was previously known was disputed through the use of the MR approach (Burgess *et al.,* 2012). The genetic variants used in MR studies have to meet three criterions (Zhou *et al.,* 2016). These criterions are as follows; 1) The SNPs should be associated with the risk factor of interest (TL), 2) The SNPs should result in the disease outcome through the risk factor and not through any other biological mechanism, and lastly 3) The SNPs should stratify the study population in the same way that participants are stratified in randomised clinical trials (Burgess *et al.,* 2012). The principle aim of MR studies is to strengthen the association found between genetic variants, risk factors and disease, in order to infer accurate causality (Allman *et al.,* 2018).

1.5 Hypothesis, aim and objectives

The hypothesis of this study was that TL would be shorter in African individuals with T2D diabetes compared to controls matched according to age, sex, ethnicity, and BMI. The state of heightened inflammation and oxidative stress in the diabetic cases results in accelerated telomere attrition when compared to healthy controls.

For this project we aimed to examine the association between TL and T2D across different SSA populations and to assess whether or not TL-associated variants are associated with T2D in SSA populations.

The following were, therefore, the specific objectives of the project:

1. To identify and categorise study participants into groups of T2D cases and unaffected controls, matching the two groups for age, sex, ethnicity, and BMI.

2. To ascertain whether there is an association between mean TL and T2D considering age, sex, ethnicity, and BMI as potential confounders.

3. To select SNPs previously associated with TL in order to test for allelic association with T2D.

Chapter 2: Methods and Materials

2.1 Chapter Outline

This chapter describes the patient cohort, study site, all materials, reagents and equipment, used for this study. Data and statistical analysis employed were also explained. All reagents used in this project were obtained from reliable suppliers. The DNA was already extracted at the outset of the project by myself, as an intern at the Sydney Brenner Institute for Molecular Bioscience (SBIMB) Biobank. I also participated in the normalisation of the samples that were genotyped using the H3A array, also while still an intern. For my masters project, I normalised the telomere assay samples, optimised the qPCR assay, performed the qPCR, conducted the troubleshooting, and thereafter performed all the genetic association study analysis of the samples run on the H3A array by Illumina.

2.2 Population and sampling

This study was conducted in the first instance as a matched case-control study to investigate the association between TL and T2D in SSA populations. Furthermore, SNP variants previously associated with TL were assessed for genetic association in this cohort. The study was nested in the larger Africa Wits-INDEPTH partnership for genomic studies on genetic and environmental contributions to obesity and cardio-metabolic diseases in Africans (AWI-Gen), which is an H3Africa Consortium study.

Black African participants (male and female) aged 40 years and above were recruited in the AWI-Gen study (Ali *et al*., 2018). Migrants, related individuals, and pregnant women were excluded from the AWI-Gen study (Ali *et al*., 2018). For my research study 600 T2D cases and 600 matched controls were chosen from the AWI-Gen database. The 600 diabetic cases were selected from those available across the 6 different AWI-Gen sites. According to the AWI-Gen definition, the diabetic participants were those with a fasting plasma glucose level of 7 mmol/l and higher, and/or those who had self-reported as having been told by a healthcare practitioner that they were diabetic. The controls were chosen by matching each individual case to a control by: age, sex, site, and BMI. The matching was done as closely as possible to the case, according to the demographic and phenotypic data of selected cases. This study included participants from the 6 AWI-Gen study sites

namely; Soweto, Agincourt, and Dikgale (South Africa); Navrongo (Ghana); Nanoro (Burkina Faso), and lastly Nairobi (Kenya) (Ali *et al*., 2018) (AWI-Gen study Ethics approval number M121029 and renewed M170880). Ethics approval for my study was granted by the University of the Witwatersrand Human Research Ethics Committee (Medical), certificate number M1403107 (Appendix D).

2.3 General Laboratory Protocols

2.3.1 Extraction and Quantification of genomic DNA

Genomic DNA used for this project was extracted prior to the initiation of the study. DNA was extracted from buffy coat samples, and then stored in SBIMB Biobank. Working aliquots of samples were stored at 4°C for use. Buffy coat samples were extracted using both the sodium chloride manual precipitation method (salting out method) (Miller *et al*., 1988) and the automated DNA extraction robot, the QIAsymphony SP (Qiagen N.V., Vento, Netherlands). The salting out precipitation method was modified specifically by the SBIMB Biobank, and the automated method was performed according to manufacturer's instructions. The detailed protocol for the salting out method is provided in Appendix A and the protocol for quantification of genomic DNA is provided in Appendix B. Before storage, samples were quantified using the NanoDrop Lite UV-Vis Spectrophotometer (Thermo Fisher Scientific, Watham, MA, USA) at the absorbance of 260nm. The genomic DNA samples were eluted in Tris Ethylene diaminetetraacetic acid (TE); the same solution was used as the blank solution for the NanoDrop and as the eluent for the extracted genomic DNA. The A260/A280 ratios were recorded to assess the purity and protein contamination of genomic DNA. Thereafter agarose gel electrophoresis was performed to assess the integrity of extracted DNA in 0.8 % agarose gel using ethidium bromide as the visualisation agent (Thermo Fisher Scientific, Watham, MA, USA).

2.4 PCR

PCR is a biological technique where minute quantities of DNA are amplified using custom designed flanking forward and reverse primers. It makes use of the enzymes *Taq* Polymerase and DNA ligase in order to exponentially amplify the desired strand of DNA. PCR is performed by exposing the target DNA and PCR solution to

repetitive thermal cycling in which the DNA strand is first denatured, then the primers anneal and then the DNA is extended. This forms the basis of the assay Cawthon, (2002) designed to amplify telomeric ends and a SCG. In order to calculate TL the relative intensity of the telomere PCR products compared to the intensity of the SCG.

2.4.1 *ALB*umin (*ALB*) versus Beta-Globin (*HBB*) versus Acidic ribosomal phosphoprotein P0 (*36B4*)

For the selection of the reference gene, three reference genes were tested; albumin (*ALB*), beta-globin (*HBB*), and (*36B4*). The first two were evaluated in Cawthon's original protocol but *36B4* was selected as it has been used in many published studies utilising the qPCR assay for TL measurement. All three genes were assessed *in silico* and were checked on the UCSC Genome Browser Human (GRCh37 hg19) (http://genome.ucsc.edu/) to confirm the specificity of the PCR primers and whether amplified fragments would uniquely map to the desired gene. All gene fragments were found to match to the required genes, ruling out mis-annealing of primers and hence failure of product amplification. The final decision was to use the *36B4* gene as the single copy reference gene as results using this gene had been more widely published than the other two genes (Cawthon, 2002).

2.4.2 Assay establishment

The initial stages of assay establishment were to first replicate the original published qPCR protocol by Cawthon (2002), using the primers for the single copy reference gene of my choice (*36B4*) and the telomere amplification primers form the original protocol. Firstly, the assay was tested on the standard curve of both the *36B4* reaction and the telomere reaction. Once that proved successful the assay was implemented on the study samples.

2.4.3 Primer Design and Storage

The design of all primers used for the PCR assay were obtained from a protocol published by Cawthon (2002). Table 2.1: lists the primer sequences and their references. Primers were ordered from Life Technologies (Thermo Fisher Technologies, Carlsbad, CA, USA). PCR grade water, purchased from Roche Diagnostics (Risch-Rotkreuz, Switzerland) was used to reconstitute the lyophilised

dry primers to a concentration of 100 µM to make the master PCR stock solution. After overnight reconstitution, the master stock solution was diluted into a series of working stock solutions of 10 µM, which were stored at -20 °C.

Table: 2.1: Oligonucleotide primer sequences

Primer name	Sequence (5' – 3')	Reference
tel 1	GGTTTTTGAGGGTGAGGGTGAGGGTGAGGGTGAGGGT	Cawthon, 2002
tel 2	TCCCGATATCCCTATCCCTATCCCTATCCCTATCCCTA	Cawthon, 2002
36B4u	CAGAAGTGGGAAGGTGTAATCC	Cawthon, 2002
36B4d	CCCATTCTATCATCAACGGGTACAA	Cawthon, 2002

2.4.4 Thermal cycling equipment

All equipment used for the amplification of genomic DNA were bought from Life Technologies (Thermo Fisher Scientific, Watham, MA, USA). Reactions were aliquotted into the MicroAmp Optical 96-well reaction plates from Thermo Fisher Scientific. The QuantStudio™5 real time PCR system was used for all qPCR assays.

2.4.5 Reaction Set-up

For prevention of contamination, all pre-PCR preparations were performed in a laminar flow cabinet, with ultra-violet (UV) light treated filter tips, and pipettes, 30 minutes prior the commencement of PCR assay preparation. The PCR was performed in a separate designated PCR laboratory.

2.4.6 Agarose gel electrophoresis

Agarose gel electrophoresis was initially used to characterise the integrity of the genomic DNA samples by the SBIMB Biobank. An agarose gel concentration of 0.8 % was used for genomic DNA integrity checks. More importantly, gel electrophoresis

was used to resolve qPCR assay products after successful amplification. A 2 % agarose gel was used for the latter.

2.4.7 Gel preparation

Two grams of molecular grade agarose (Thermo Fisher Scientific) was boiled with 100 ml of 1X Tris Borate EDTA (TBE) buffer, in a microwave, to make 2 % agarose TBE/agarose solution, where 2 µl of ethidium bromide was added prior to pouring the agarose into the casting tray with a twelve toothed comb. Once the gel had set, the comb was removed and the casting tray was submerged in 1X TBE buffer in a gel tank.

2.4.8 Sample preparation and electrophoresis

Following the agarose gel preparation, 5 µl of PCR product, and 2 µl of loading dye were mixed together and loaded into the wells of the gel. PCR product samples were run adjacent to a 1KB+ ladder and a negative water control. A voltage of 80 V was applied for all the gel electrophoresis experiments, each experiment was run for approximately 40-45 minutes, or until the dye front was two thirds down the length of the gel. Following electrophoresis, the gel was placed in the BioRad Gel Doc™ EZ system, for visualisation under UV light and annotation. Gel images were saved as .JPEG files.

2.5 Quantitative PCR

The method used to measure TL was adapted from the original article by Cawthon, 2002. With minor alterations to the protocol, relative TL was measured using two separate PCR reactions; the SCG reaction, and the telomere reaction respectively. This method is usually referred to as the two-plate method. The assay relies on the comparison of specifically designed primers that amplify telomere repeats (T) in a genomic sample, relative to the amount of the SCG (S), and expressed as a T/S ratio. For this PCR the SCG, acidic ribosomal phosphoprotein PO (*36B4*) gene was used. The T/S ratio calculated post-PCR is essentially the relative TL for that particular sample. Samples were prepared in 96-well optical reaction plates as referred to in section 2.4.4 above. The samples for both the telomere and SCG

reactions were located in the same wells on both plates, in order to maintain as much similarity as possible.

2.6 Assay Protocol

2.6.1 Assay preparation

Tables 2.2 and 2.3 list the reagents and their required volumes used to make the master mix (MM) for the assay. The *36B4* gene and the telomere reactions were prepared separately. The volume of each MM was dependant on the number of samples to be assayed, and a 10 % increase in volume was allowed for pipetting error. After preparation, the MM was vortexed gently, then added to reaction plates and mixed with genomic DNA, by gentle aspiration and expiration during pipetting. 6.36 µl of MM was used and 3.63 µl of sample DNA was added for each reaction. The reaction plate was the covered with an optical plate seal and centrifuged for 2 minutes at 2800 g in an Allegra x-12 (Beckman Coulter, Indianapolis, USA). This was done in order to remove bubbles and to ensure that the samples were at the bottom of each well. The plate was the inserted into the QuantStudio™ 5 (Thermo Fisher Scientific, Watham, MA, USA). It is important to note that the two reactions were run separately as they had different cycling conditions, which are shown in Tables 2.4 and 2.5.

Table 2.2: Telomere reaction master mix

Reagents	Stock concentration	Final concentration	Volume for 1 reaction	Supplier
DTT	100 mM	2.0 mM	0.2 µl	Thermo Fisher Scientific, Watham, MA, USA
Tel 1 Primer	10 µM	270 nM	0.27 µl	Thermo Fisher Scientific, Watham, MA, USA
Tel 2 Primer	10 µM	900 nM	0.9 µl	Thermo Fisher Scientific, Watham, MA, USA
Power Up SYBR® Green Dye	2 X	1 X	5 µl	Thermo Fisher Scientific, Watham, MA, USA
DNA			3.63 µl	

Table 2.3: *36B4* reaction master mix

Regents	Stock concentration	Final concentration	Volume for 1 reaction	Supplier
***36B4*u Primer**	10 µM	300 nM	0.3 µl	Thermo Fisher Scientific, Watham, MA, USA
***36B4*d Primer**	10 µM	500 nM	0. 5 µl	Thermo Fisher Scientific, Watham, MA, USA
Power Up SYBR® Green Dye	2 X	1 X	5 µl	Thermo Fisher Scientific, Watham, MA, USA
PCR Grade water			0.57 µl	Roche Diagnostics
DNA			3.63 µl	

Table 2.4: Telomere reaction cycling conditions

Cycles	Temperature	Time
1	95 °C	10 minutes
40	95 °C	15 seconds (Data Collection)
40	54 °C	2 minutes (Data Collection)

Table 2.5: *36B4* reaction cycling conditions

Cycles	Temperature	Time
1	95 °C	10 minutes
40	95 °C	15 seconds (Data Collection)
40	45 °C	1 minute and 10 seconds (Data Collection)

2.6.2 Quality control and quality assurance for the telomere measurement assay.

A rigorous quality control (QC) routine was followed from selection of samples to assay data analysis. Sample QC was adopted from a publication by Seeker *et al.,* 2016, to ensure uniform and consistent results. Figure 2.1 depicts the overall workflow of the implemented QC. Samples that failed QC were removed from further analysis. The sample QC included the following criteria; a DNA yield of 50 ng/µl and more on the NanoDrop, a 260/280 ratio of less than 2.0 but more than 1.8 and a 260/230 ratio of between 2.0 and 1.8. If samples, passed this criterion they were cleared for PCR amplification.

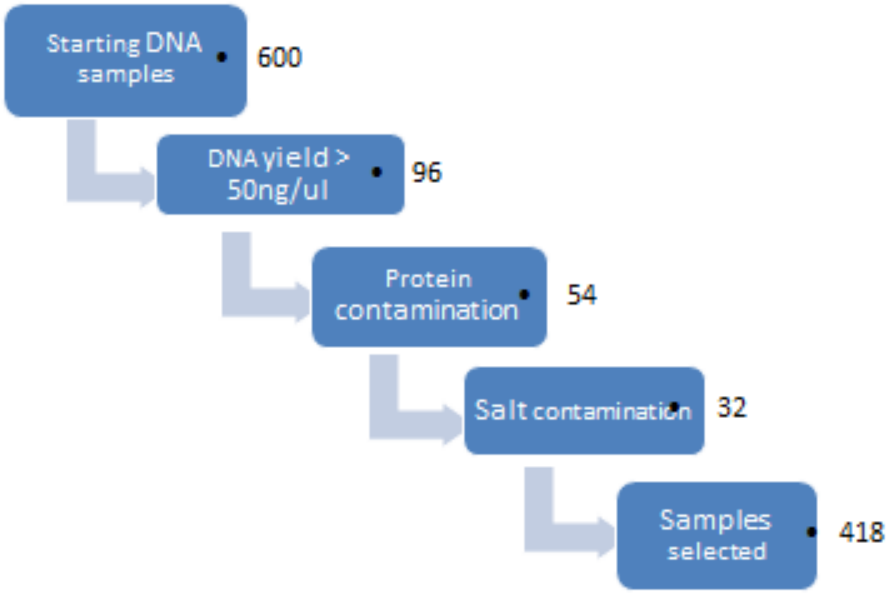


Figure 2.1: A total of 600 diabetic samples were identified from the AWI-Gen cohort data. Of these 600 cases, only 418 passed quality control, 96 samples had a DNA yield of less than 50 ng/µl, 54 were removed due to protein contamination and 32 were removed to salt contamination. The matched controls were removed with their cases in order to maintain the same number of cases and controls in the study.

For every plate a reference (golden) DNA sample and a negative control (water and master mix only), were run in triplicate to ensure quality assurance. Only 10% of the study samples were assayed in triplicate; and upon arriving at minimum variation, the remaining samples were assayed only once. The assay did not include a positive control sample, as there were no samples with a known TL. Moreover, the TL from a positive control sample had to be determined by the Terminal Restriction Fragment (TRF) analysis, which is considered the gold standard of telomere measurement. Since our laboratory has neither the equipment, nor the capacity to perform TRF analysis, a validated positive control was not included in the assay.

As specified by Cawthon (2002) in his initial protocol, a standard curve for each reaction, telomere and *36B4*, was required. This served the purpose of assessing

the amplification efficiency of each PCR run. To guarantee amplification efficiency; the standard curve efficiency, calculated by the QuantStudio5™ Design and Analysis software (v.1.4.3) had to fall within a percentage range of 90% - 110% (Bustin and Nolan, 2004). Moreover, the T/S ratio of every point within the standard curve had to be the same value to ensure that the assay was sensitive enough to detect differences in initial sample concentration. Lastly, following sample amplification on telomere and *36B4* reactions, the melting point of the PCR products was analysed. This was performed on the QuantStudio5™ design and analysis software (v.1.4.3). Due to the different PCR product sizes from the two reactions, the two reactions have different melting temperatures, thus different melting curves. The telomere amplicon was ~79 base pairs (this would vary according to average TL), whilst the *36B4* amplicon was 106 base pairs. On the melt-curve all samples had to have a melting temperature a degree less of above the designated melting point, in order to qualify for analysis. The melt-curve analysis also effectively tested for non-specific amplification and primer dimerisation.

2.7 Variant selection and genotyping of TL associated variants

With the recent rise in epidemiological studies of telomere length, the number of GWAS in this field has also increased. To date, several genetic variants implicated in TL variation and disease have been uncovered. The opportunity for performing a TL association in this cohort arose at the SBIMB after the genotyping data from the AWI-Gen study became available. The samples were genotyped on the H3A SNP genotyping array. The data from the AWI-Gen study for the samples used in my study were made available to me. For this part of my study I first selected published genetic variants that showed an association with TL and then used these variants as a proxy for TL. For the SNPs that I had available data, I tested the association of TL and T2D.

2.7.1 Variant selection

The variants used for this association were previously published to have an association with TL and that reached a genome-wide level of significance ($p<5x10-8$). Statistical power which is determined by sample size, allele frequency and effect size were not tested prior to analysis as it was not clear what values to add for allele

frequency and effect size. However a post hoc statistical power analysis for the study was run with a sample size of 770 cases and controls. The post hoc power calculation is explained in detail in section 2.8.2. A total of 53 telomere-associated SNPs were initially selected from the literature but only 28 of the SNPs were on the H3A SNP array. The association analysis was therefore run on 770 samples and 28 SNPs (the remainder of the 836 samples mentioned above did not have DNA results).

2.7.2 Genotyping and GWAS QC parameters

The H3A SNP array was specifically designed to be enriched for common African variants in order to be more appropriate for research in African populations. This African genotyping platform was used for determining the genotypes of ~2.3 million variants per study participant in this African cohort (Mulder *et al*., 2018). DNA samples were normalised to between 30 and 100 ng/µl and aliquotted into Illumina Midi plates as per service provider guidelines. The array was performed by Illumina at their laboratory in San Diego, California (USA), according to their in-house procedures. Data generated was then transferred to the SBIMB via secure download. All samples were run on the array irrespective of whether or not they passed Illumina's original QC check to ensure maximum output and minimise loss of samples. The SBIMB analysis team was then responsible for performing the QC on the samples and SNPs when the genomic data was received.

H3A Bionet developed a nextflow pipeline for the H3A Consortium to use for all of the genome-wide association studies (https://github.com/h3abionet/h3agwas). The AWI-Gen GWAS analysis team at the SBIMB optimised the pipeline for analysis of the QC of the AWI-Gen dataset. This meant that original cut-off values were modified for each of the QC steps both in terms of sample and SNP QC. Please refer to the Figures 2.2 and 2.3 below, which is a flow diagram which shows QC parameter cut-offs used during each stage of the QC process.

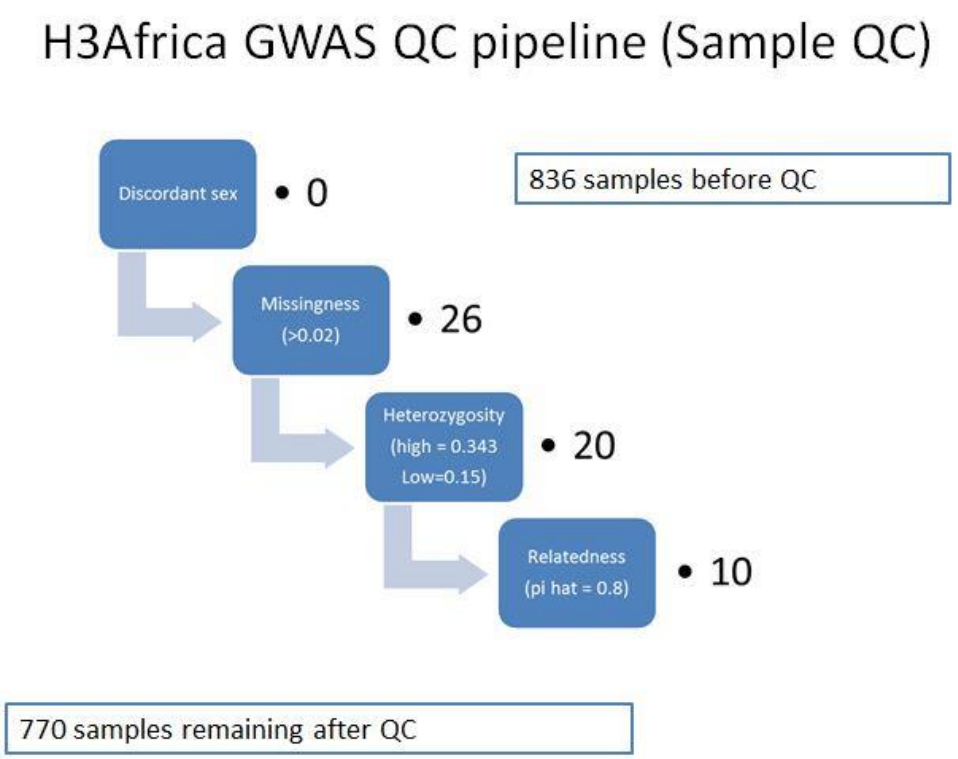


Figure 2.2: H3Africa GWAS QC pipeline parameters for sample QC and the number of samples that failed QC at each step. Sixty-six samples failed QC, 770 progressed for further analysis.

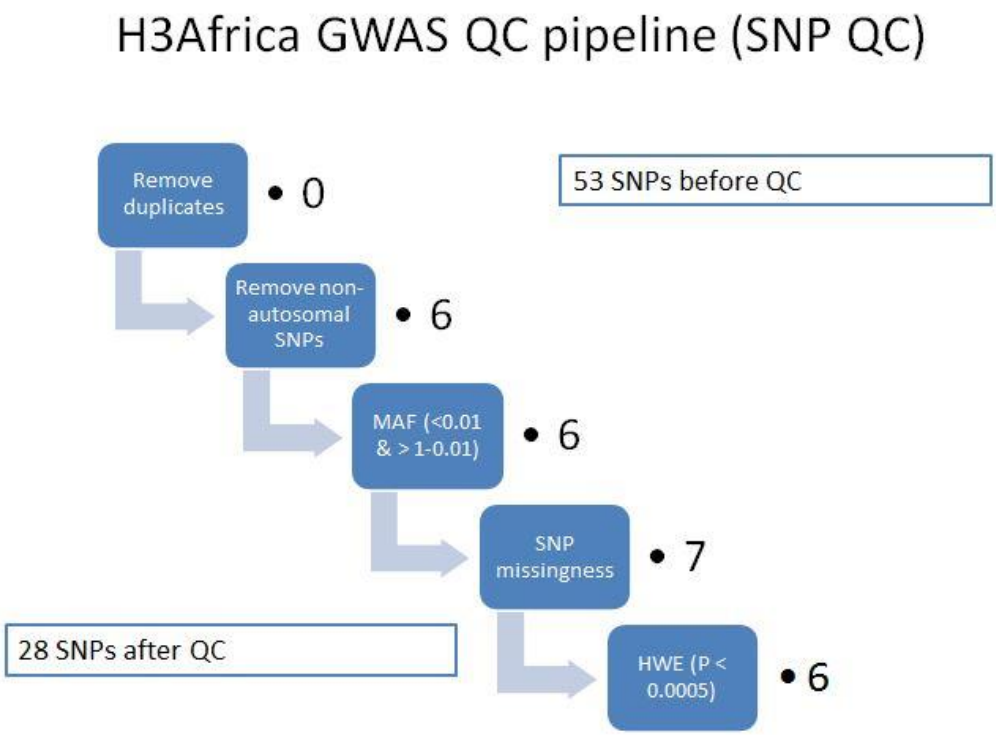


Figure 2.3: H3Africa GWAS QC pipeline parameters for SNP QC and the number of SNPs that failed QC at each step. Twenty-five SNPs failed QC, 28 SNPs progressed for further analysis.

2.8 Data analysis

Descriptive statistics for the study participants was performed on the latest version of the statistical software package for the social sciences (IBM SPSS Statistics, 2015). Sex and site (as a proxy for ethnicity) were assessed as nominal (categorical) variables, while age and BMI were assessed as continuous variables. A chi square test was done to determine whether sex was equally represented between cases and controls. Two-sample Wilcoxon rank-sum tests were carried out to determine whether or not BMI and age were significantly different between the case and control groups.

2.8.1 Data analysis for TL assay

If the amplification of the telomere and *36B4* gene reactions were successful, the Ct values, amplification efficiencies, and standard curves would have been analysed on the QuantStudio5™ Design and Analysis software (v1.4.3). Thereafter the results

obtained would have been exported to Microsoft Excel, where relative T/S ratios would be calculated using the Pfaffl method (Pfaffl, 2001), which is a relatively new method for analysing relative quantification. A standard curve is required for this method of analysis, as it takes into account the respective amplification efficiencies.

2.8.2 Post-test power calculation

Power was assessed using Quanto (http://biostats.usc.edu/Quanto.html) (Sexena *et al*., 2017 and Laufer *et al*., 2018) for separate case-control approaches with 770 participants. For example, with a sample size of 770 individuals the study is 80% powered to observe an OR (effect size) of 1.4, given an allele frequency of at least 0.12. The Figure 2.4 below shows the different OR and MAF parameters that would result in the different statistical power percentages from zero to a hundred (0 – 1).

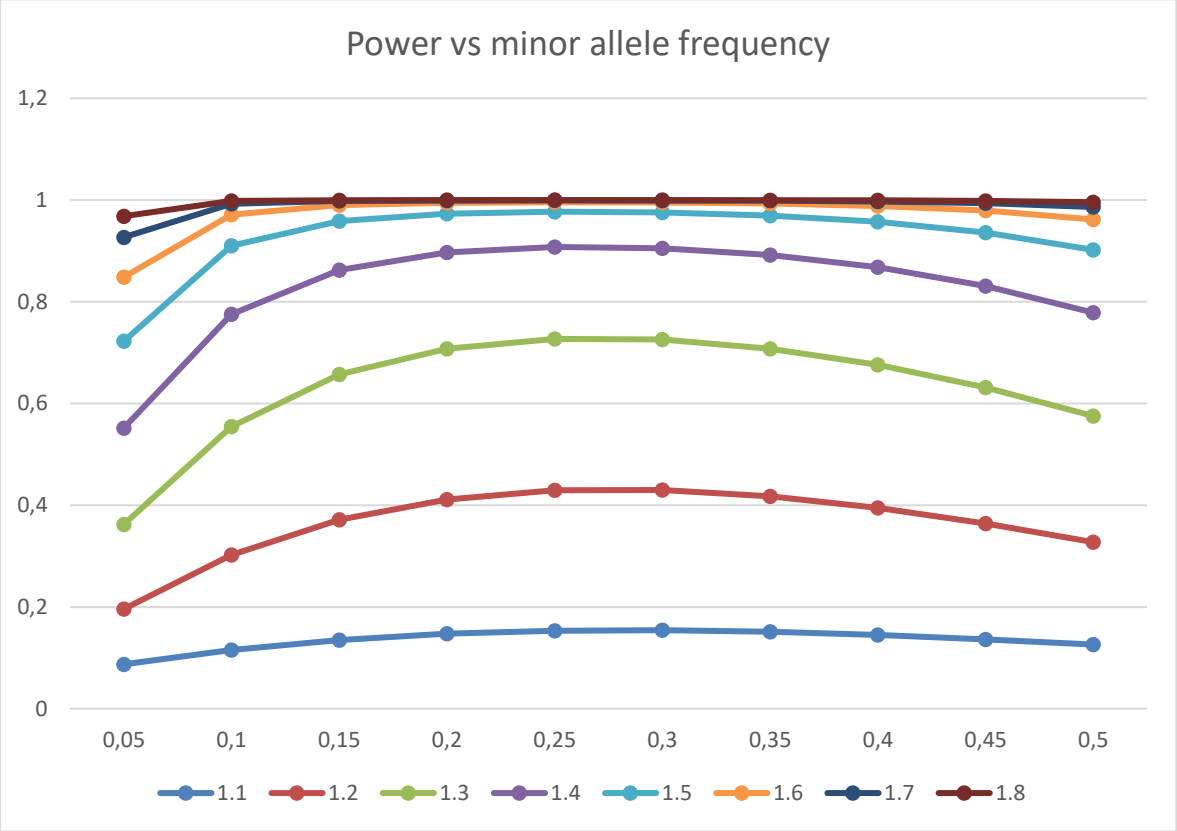


Figure 2.4: Power versus minor allele frequency graph. The graph shows the minor allele frequency (x-axis) at which a certain odds ratio (1.1 – 1.8) represented as the different coloured lines, reaches at certain power percentage (y-axis).

2.8.3 Data analysis for association

PLINK software (version 1.9, 2005 – 2018) was used for the association of the chosen TL variants with diabetic cases and controls. Files that are compatible with PLINK (.map and .ped) were generated from the raw data received from the H3Africa array. These were then converted into .bim, .bam and .fam files in PLINK for further association analyses. The data were separated into diabetic and non-diabetic participant categories so that a case-control type of analysis could be carried out. Diabetic participants were tagged as cases, and non-diabetic participants as controls. Allele and genotype frequencies were calculated from the genotype data for all the SNPs. A logistic regression analysis was used to test for association. Logistic regression is predictive regression analysis carried out on a dichotomous dependent variable (T2D and non-T2D). Logistic regression gives an OR that is the estimated measure of association between an exposure (variant) and outcome (T2D). The OR gives the odds of an outcome occurring if an exposure is present, compared to odds of the outcome occurring without the exposure variable. It gives a log odds increase of the outcome. An OR >1 signifies that the minor allele (A1) is associated with increased levels of T2D. In contrast, an OR < 1 means that the minor allele is associated with decreased levels of T2D. This logistic regression model was run with age, sex, and BMI as confounders.

Chapter 3: Results

3.1 Distribution of T2D cases and controls across AWI-gen sites

The inclusion criteria for the AWI-Gen participants were that they could be male or female participants of SSA origin, needed to be aged between 40 and 80 years, and not be pregnant if female (Ali *et al*., 2018).

In total, 600 participants from the AWI-Gen database were selected for the case group and they were then used to select appropriate controls. At the time of participant selection, the recruitment for AWI-Gen study had not been completed and only T2D cases already in the database could be assessed for this sub-study. It was therefore not possible to have the study participants evenly distributed across all the AWI-Gen sites. As such, the distribution of participants from each region (South, West and East) was skewed. A “group matching strategy” for age, sex and site (a proxy for ethnicity) was used and 600 controls were identified. After the quality control measures had been applied for the telomere assay, as explained in the methods chapter, only 418 cases remained and were matched to 416 controls, making a total of 836 participants who could progress to further analysis. The number was further reduced to 770 participants (385 cases and 385 controls) for the logistic regression analysis, as not all of the 836 participants above had genotyping data. The distribution of the 836 cases and controls is shown in Table 3.1. This also shows the number of participants selected from South, West, and East Africa. The majority of the participants were from South Africa.

Table 3.1: Distribution of 836 cases and controls across the different AWI-Gen sites

Distribution of T2D cases and controls per AWI-Gen sites stratified by sex							
	Diabetic cases (n = 418)		Matched controls (n = 418)				
AWI-Gen Site	Male	Female	Male	Female	Total per site	Region	Total per region
Agincourt	66	88	66	88	308	South Africa	608
Dikgale	26	72	26	72	196		
Soweto	16	39	16	39	104		
Nairobi	27	65	27	65	184	East Africa	184
Nanoro	4	2	4	2	12	West Africa	44
Navrongo	11	5	11	5	32		
Total by sex	146	272	146	272	836		

A principal component analysis (PCA) was performed in order to examine possible genetic substructure between the cases and controls selected from the AWI-Gen samples. The PCA plot of the cases and controls clustered as expected. With participants from South, East, and West Africa segregating into distinct clusters, irrespective of whether they were cases or controls. Furthermore, additional PC analysis with non-African populations (Figure 3.2) provided evidence that the samples used in this study were essentially non-admixed African samples (i.e. did not have substantive European or Asian ancestry).

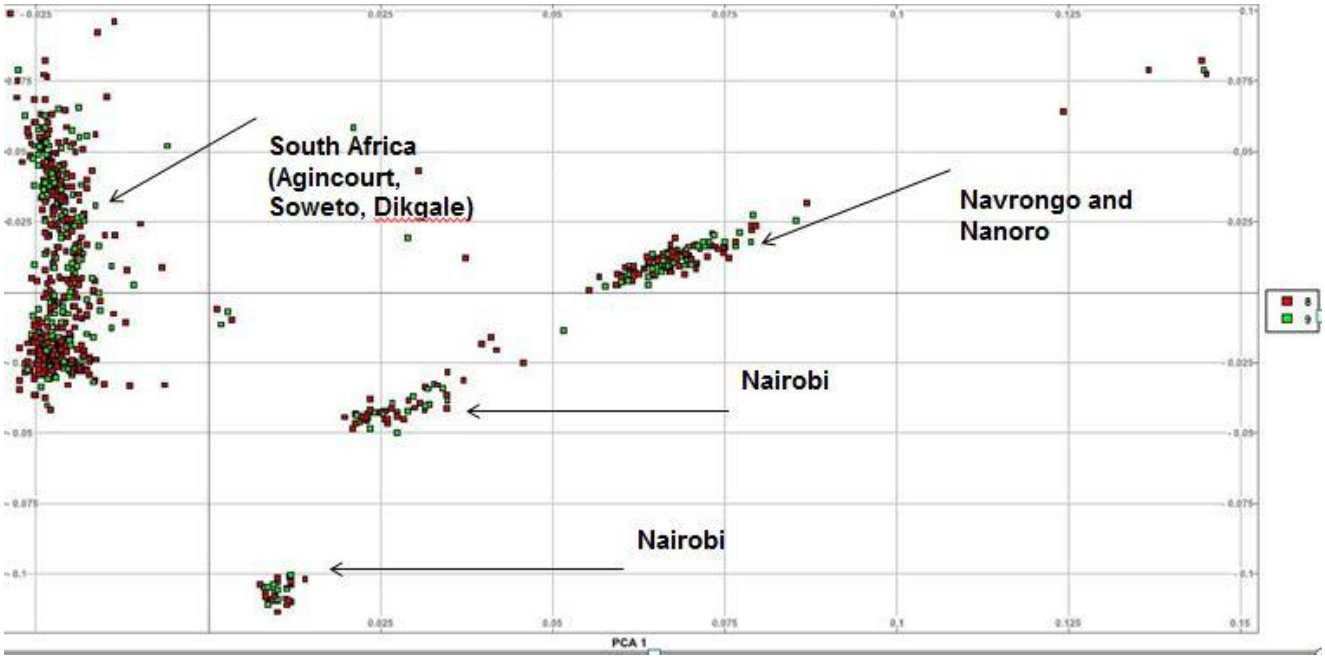


Figure 3.1: Principal component analysis plot, with the cases (red squares) and controls (green squares) from this study. In general the cases and controls from specific regions clustered together and there were a few (not many) apparent outliers that clustered between groups.

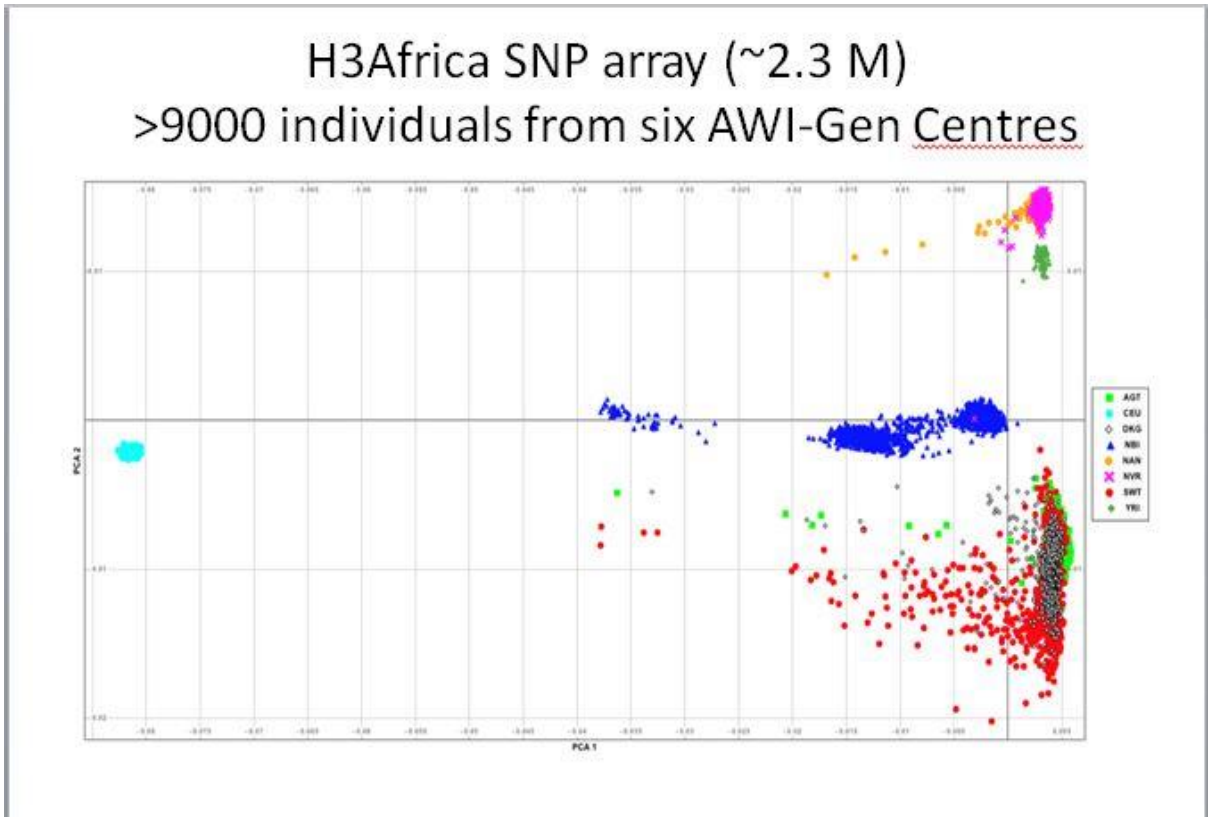


Figure 3.2: Principal analysis plot of the AWI-Gen participants, with Dikgale (White diamonds), Navrongo (Pink crosses), Soweto (Red dots), Namibia (Dark Blue triangles), Nanoro (Orange circles), Agincourt (Green squares), Yoruba (Green diamonds) and lastly Europeans (Blue circles).

The phenotype data shown in Table 3.2 used for the AWI-Gen participants were sex, age, and BMI. They were stratified into two groups representing cases and controls. The mean age of the T2D cases and matched controls was 53.23 and 53.87 years, respectively. No phenotype showed any significant differences, as expected, because participants were matched according to these parameters.

Table 3.2: Phenotype characterisation of the 836 cases and controls; according to sex, age and BMI.

Phenotype Characterisation					
Phenotype	Diabetic cases (n = 418)		Matched controls (n = 418)		P value
Sex (%F)	65.07%		65.07%		0.756
	Mean	Standard deviation	Mean	Standard deviation	
Age (years)	53.23	±8.04	53.87	±8.97	0.195
BMI (kg/m^2)	29.47	±4.21	27.92	±3.66	0.345

3.2 TL measurement assay

The TL measurement assay was divided into two separate sub-assays, hence the name two-plate method. It consists of two reactions: the SCG reaction and the telomere reaction. In this study the SCG gene of choice was the acidic ribosomal phosphoprotein PO gene (*36B4* or *RPLPO*). Throughout this study the *36B4* shorthand is used. The two reactions in the measurement assay are used to calculate the telomere/SCG (T/S) ratio; hence both crucial to the performance of the method and have to work before the ration can be calculated. The results from the *36B4* gene will be presented first then the telomere results after.

3.2.1 Standard curves

The standard curve contained six dilution points and non-template controls (NTCs), each plated in triplicate. The concentration points were 20ng/µl, 10ng/µl, 5ng/µl, 2.5ng/µl, 1.25ng/µl, and 0.625ng/µl. The reference or golden sample was used for all standard curves as it would also serve a purpose in QC, to calculate inter- and intra-assay variation. Moreover, the standard curve sample is also used to eventually calculate the T/S ratio. From the initiation of the assay the reference gene reaction performed better than the telomere reaction. Figures 3.1 and 3.2 illustrate the

standard curves for two *36B4* reactions performed. The slopes are -3.168 and -3.067 respectively; this means that the efficiency was close to 100%, which is ideal. A slope of -3.33 gives an efficiency of 100%. The efficiency percentages were 106.836% and 103.589% respectively. Lastly the parameter R^2 , which is also a measure of accuracy was 0.961 and 0.981 for plots A and B. Ideally the R^2 value should be 1, or as close to one as possible.

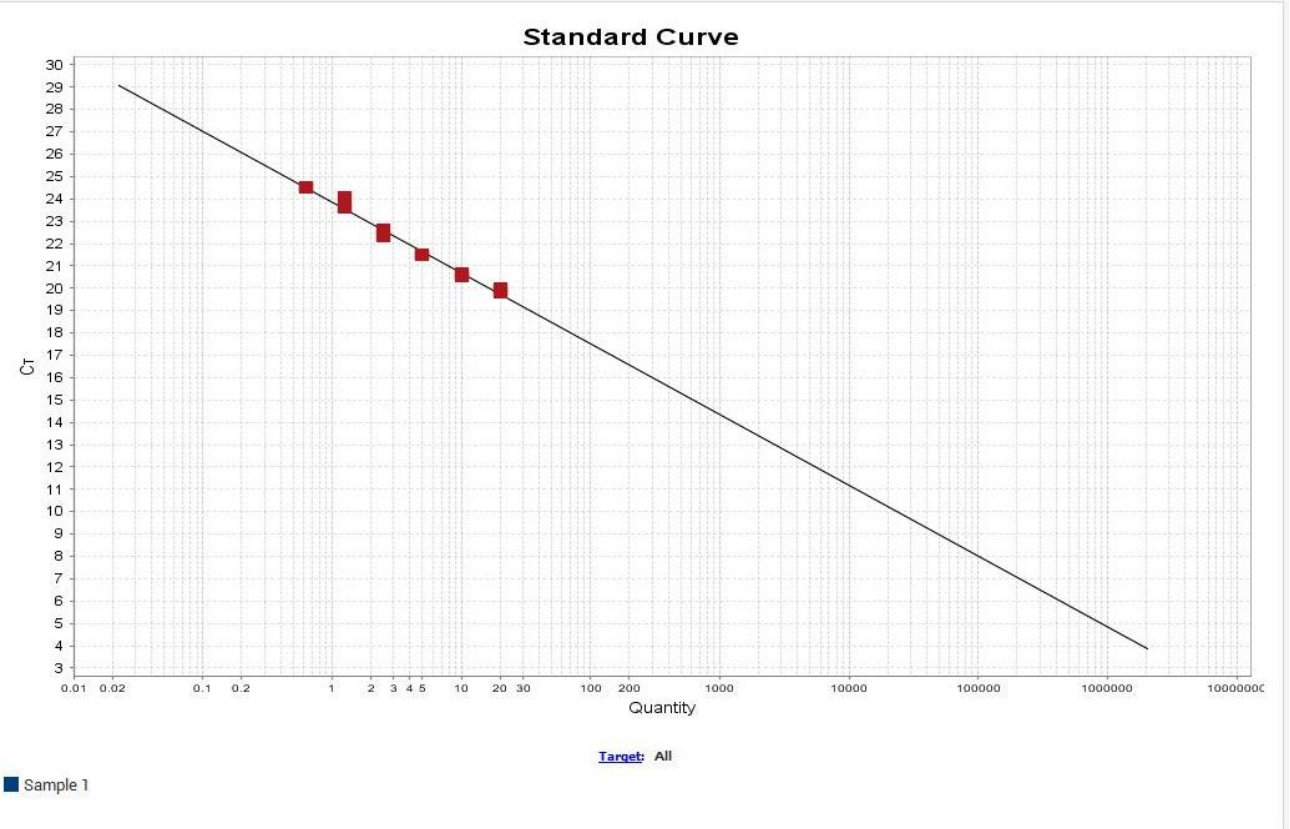


Figure 3.3: Standard curve analysis for one of the *36B4* gene reactions prior to amplification of the study samples. (R^2 = 0.961, slope = -3.168 and efficiency = 106.836 %).

This is the QC step incorporated into the assay, to show that the amplification curves seen are accurate. The standard curve is then used for the calculation of the T/S ratio, using the Pfaffl method, explained in the methods chapter (Pfaffl, 2001). The slope was -3.168 this means that the efficiency was close to 100%, which is ideal. The efficiency percentage was 106.836 %. R^2, which is also a measure of accuracy was 0.961 and for this plot.

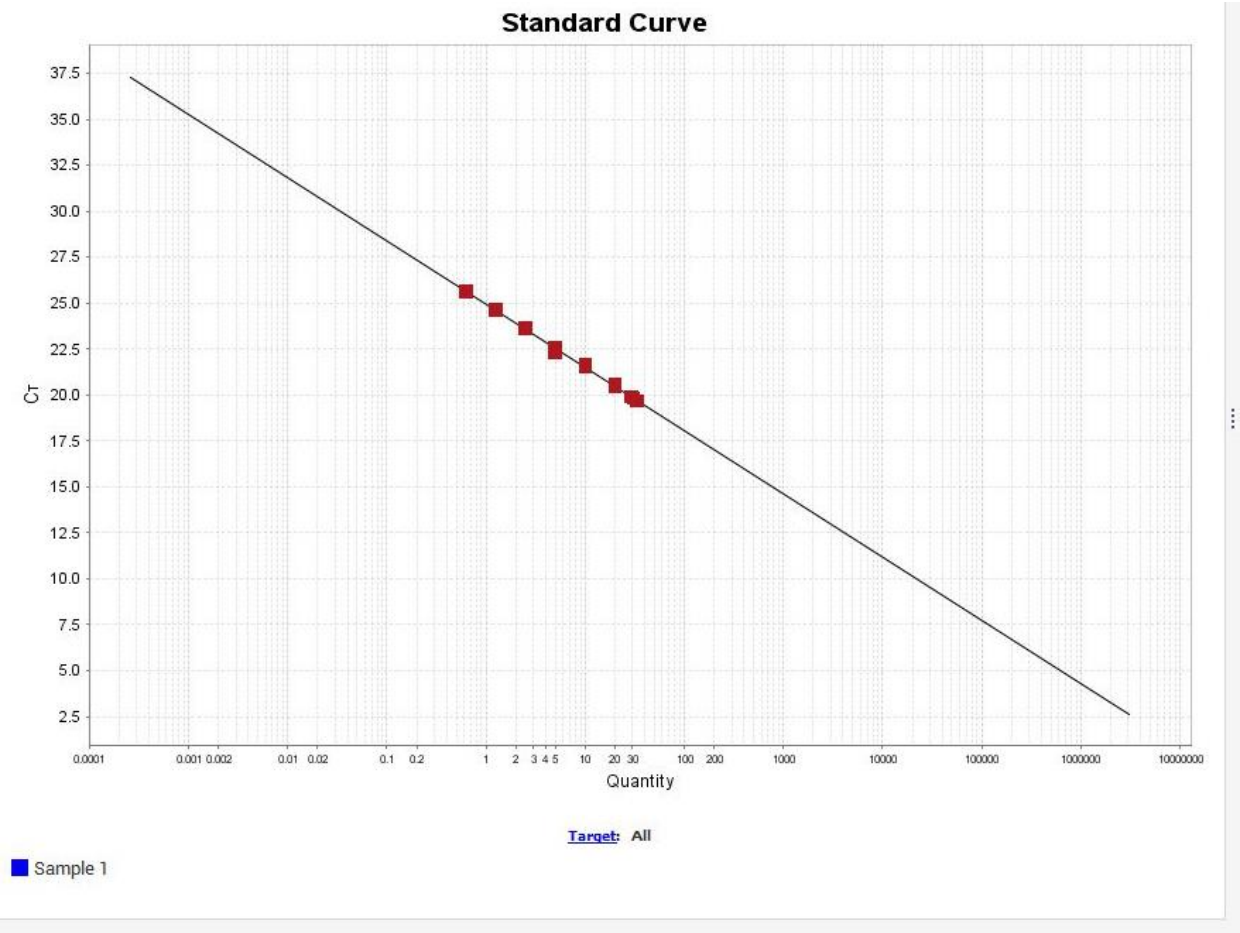


Figure 3.4: **The *36B4* gene reaction amplification curve was performed prior to amplification of study samples (R2 = 0.981, slope= -3.067 and efficiency = 103.589).**

The standard curve amplification plots above show the analysis for the *36B4* gene conducted prior to amplification of the study samples. This is the QC step incorporated into the assay, to show that the amplification curves seen are accurate. The standard curve is then used for the calculation of the T/S ratio, using the Pfaffl method, explained in the methods chapter (Pfaffl, 2001).

The standard curves for the telomere reaction were performed under the telomere cycling conditions but with the same parameters as the *36B4* gene. Figures 3.5 and 3.6 illustrate the telomere reaction standard curves established prior to study sample amplification.

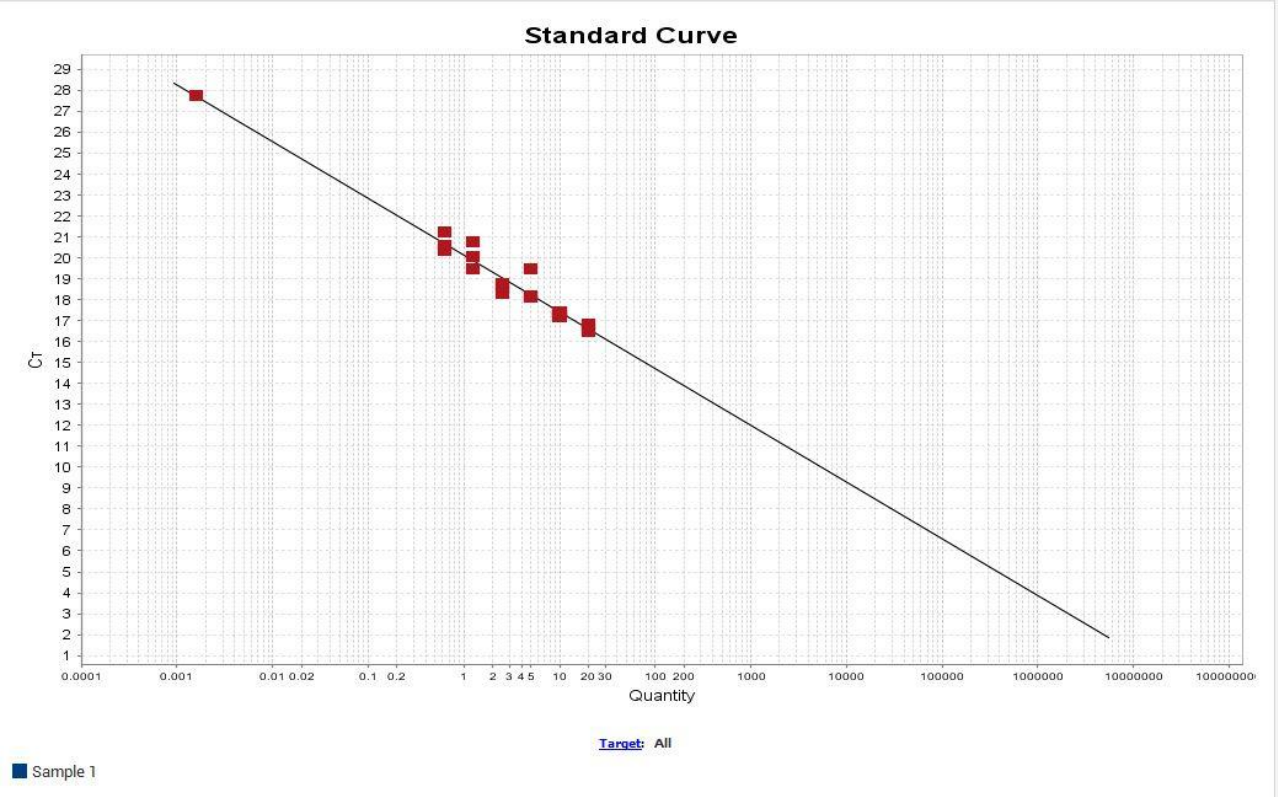


Figure 3.5: Telomere reaction as a standard curve. In this Figure the reaction seems to have worked relatively well. (R^2 = 0.900, slope -2.713, efficiency=133.673 %).

This slope is quite far from the ideal -3.33, and some of the samples do not fall on the regression line. This suggests that the reaction worked, but not optimally.

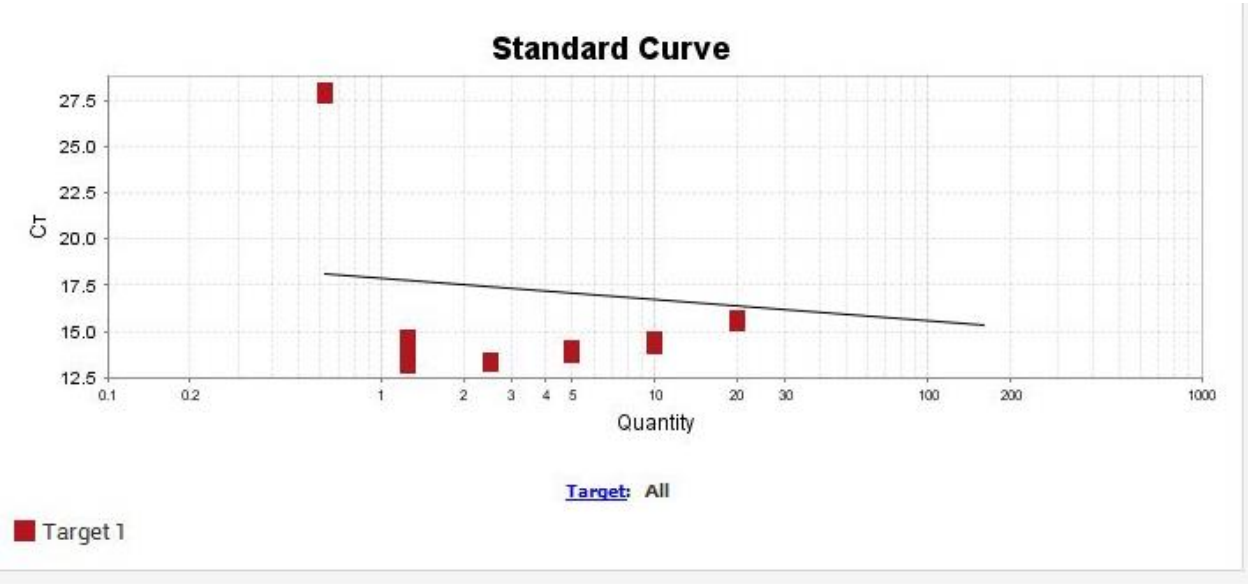


Figure 3.6:Standard curve of a *36B4* reaction that did not work well (R^2 = 0.467, slope= -0.713, efficiency= 54.835 %).

These parameters were far from ideal as the efficiency was substantially less than 100, and the role was less than 0.5, when a slope of more than 0.9 is required. Moreover, none of the samples lay within the regression line. This suggests that the serial dilution and amplification thereafter were flawed.

3.2.2 Amplification of the *36B4* gene

Following successful standard curve analysis, the test samples were assayed according to the cycling conditions mentioned in chapter 2. The Figures 3.7, 3.8, 3.9 and 3.10 show the amplification of 4 different SCG reactions. The amplification curves show the desired S-shaped curve in all the plots, despite the late amplification, and failure of amplification of a sample seen as the jagged line in Figure 3.7. During the initial assay of the study samples, the standard curve was conducted only twice before the samples were assayed. Only upon the second attempt of optimisation was a standard curve run on each and every plate. This was done to be able to effectively and accurately evaluate the efficacy of our sample amplification for each individual plate. Firstly, all the samples were normalised to a concentration of 5 ng/ul and run according to respective reaction cycling conditions. As seen with the standard curve, the *36B4* reaction performed better than the telomere reaction. Eighty-eight samples were chosen for the trial and were each replicated three times. A NTC was plated in each plate in triplicate with the reference (golden) sample also plated thrice.

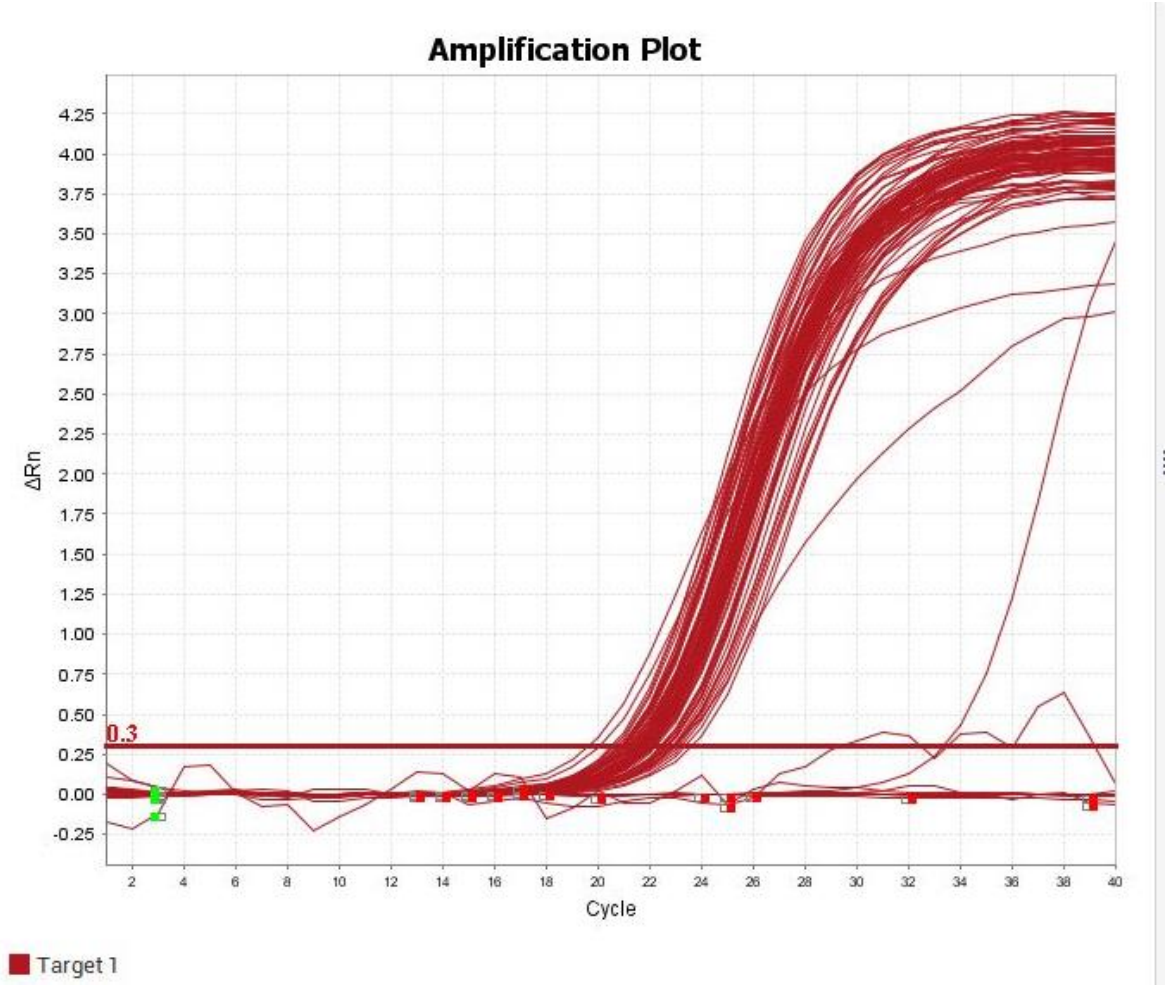


Figure 3.7: A representative amplification plot of the optimised *36B4* PCR.

The threshold was assigned at 0.3, as prescribed by the protocol by Cawthon, (2002). Which is ideal, as the threshold should lie between 0.2 and 0.4. The lines that remain flat below the threshold show the NTCs, as there is no amplification, there was no contamination with the controls. However one sample failed depicted by the jagged line and a few did not amplify optimally to present the S-shaped curve.

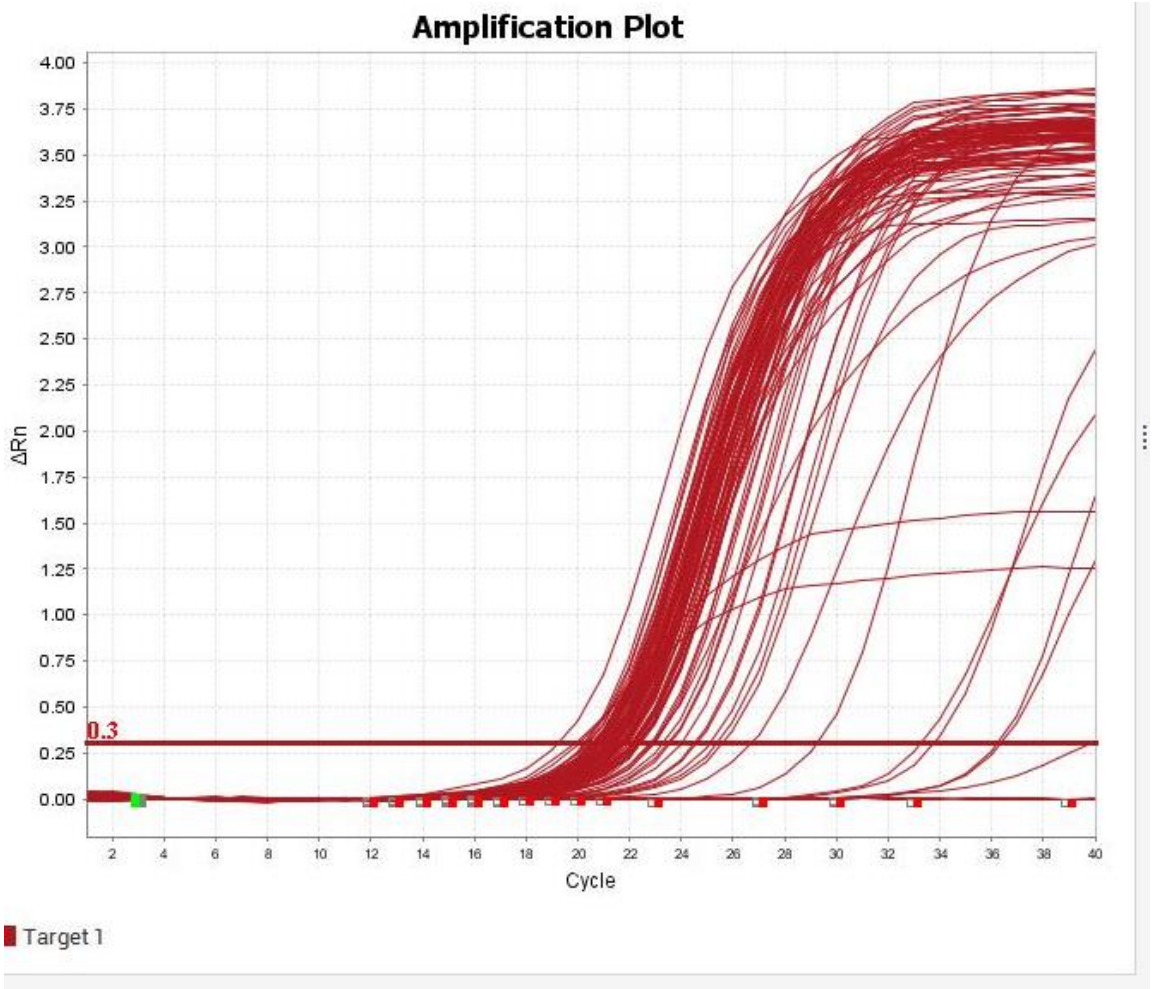


Figure 3.8: A representative amplification plot of the optimised *36B4* PCR.

The threshold was assigned at 0.3, as prescribed by the protocol by Cawthon, (2002). Which is ideal, as the threshold should lie between 0.2 and 0.4. The NTCs showed no amplification as expected, a few samples, as with the previous plate did not amplify optimally to depict the S-curved shape.

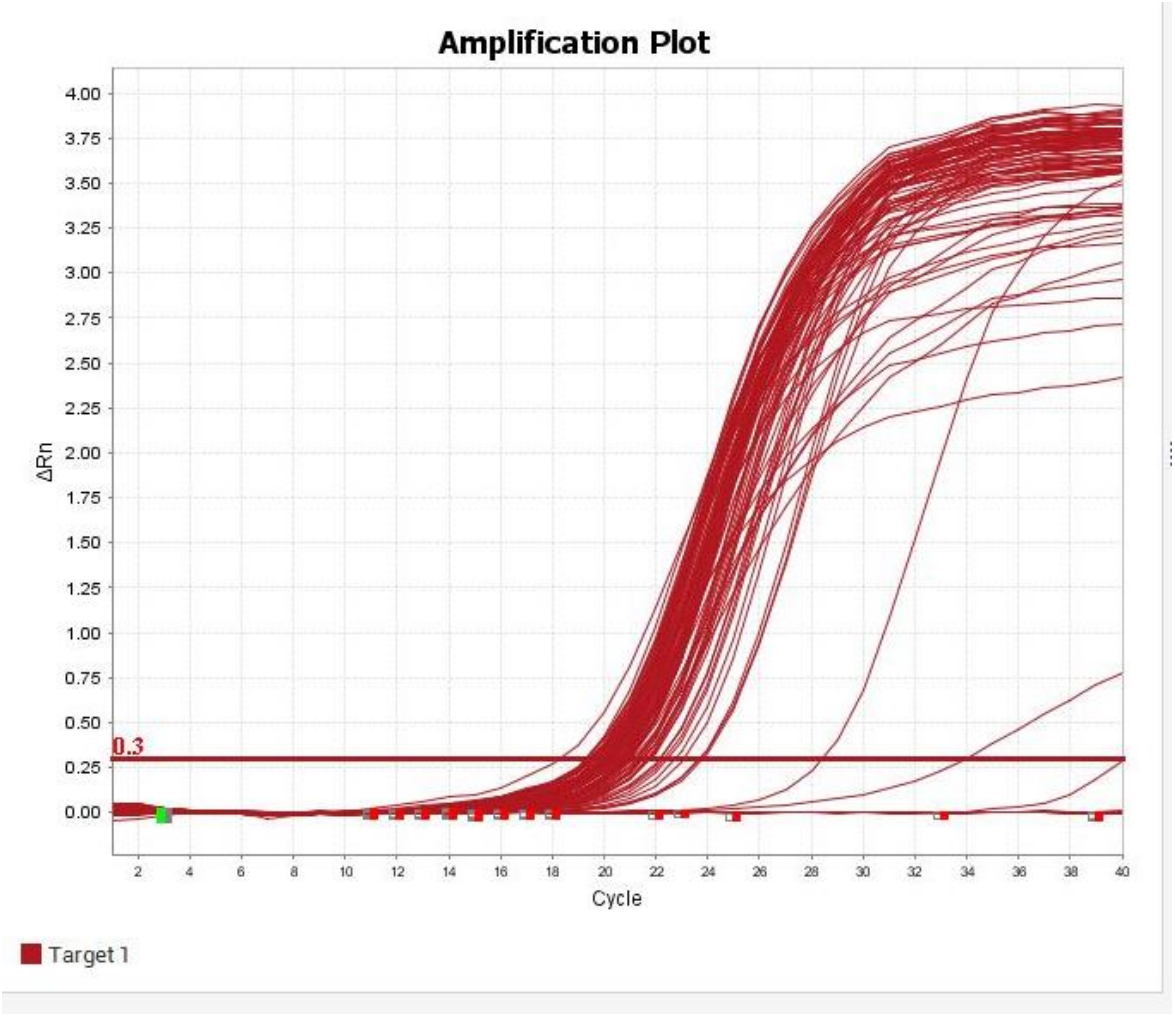


Figure 3.9: A representative amplification plot of the optimised *36B4* PCR.

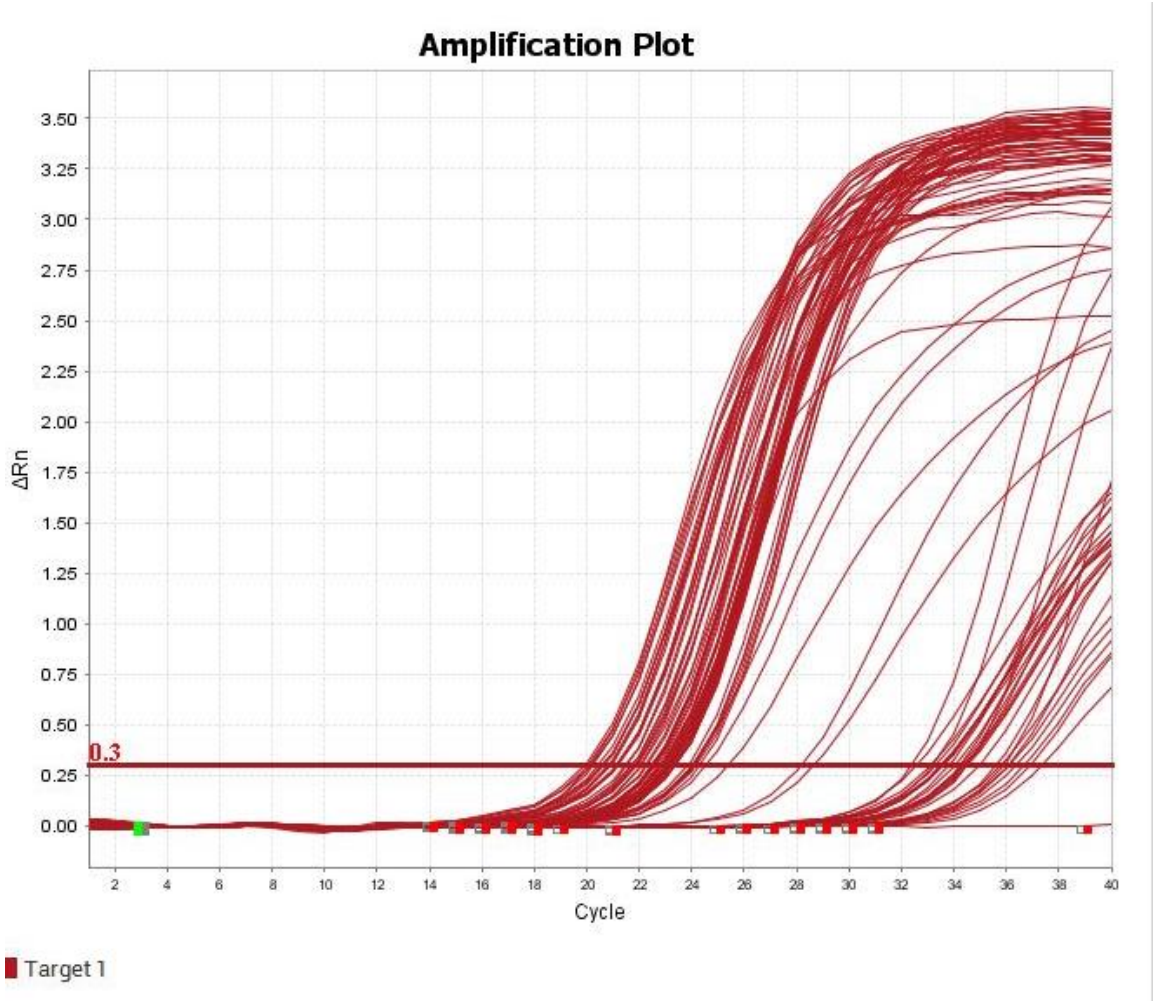


Figure 3.10: A representative amplification plot of the optimised *36B4* PCR.

As seen in the Figures above, not all samples amplified optimally. Some start amplifying at later stages as seen in this plot. This may be due to pipetting error resulting in a sample that is not of the correct concentration.

After amplification, for quality assurance, the amplification curves of the reference (golden sample), and the NTCs are viewed separately. When amplification is present in the NTCs that means that the sample plate was contaminated. Figures 3.11, 3.12, 3.13 and 3.14, illustrate the amplification curves of the golden sample and the NTCs.

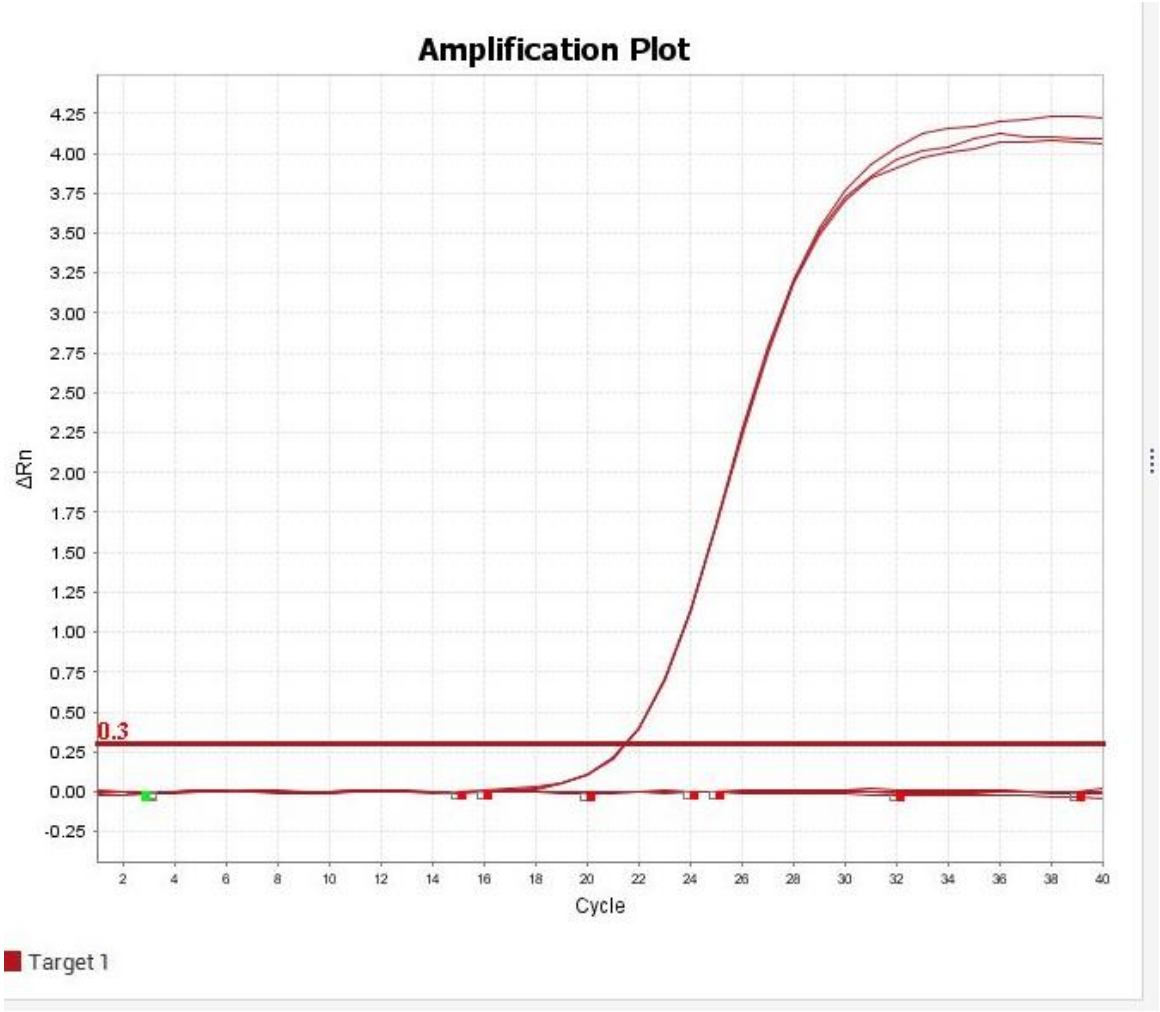


Figure 3.11: A representative amplification plot of the optimised *36B4* gene PCR, showing only the amplification of the golden sample and the NTCs.

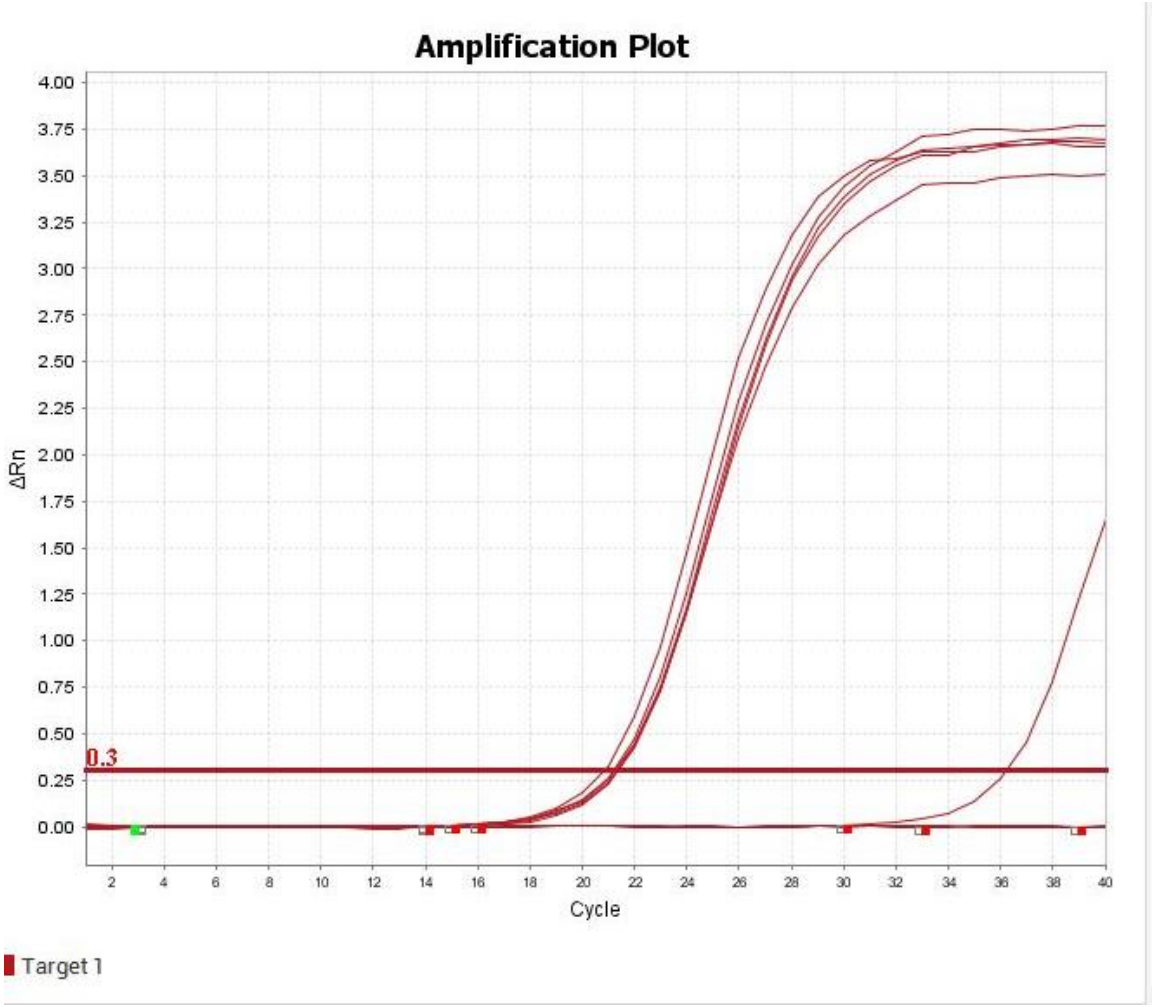


Figure 3.12: A representative amplification plot of the optimised *36B4* gene PCR, showing only the amplification of the golden samples and the NTCs. One NTC sample started to amplify towards the end of cycling (round 31) in this plate.

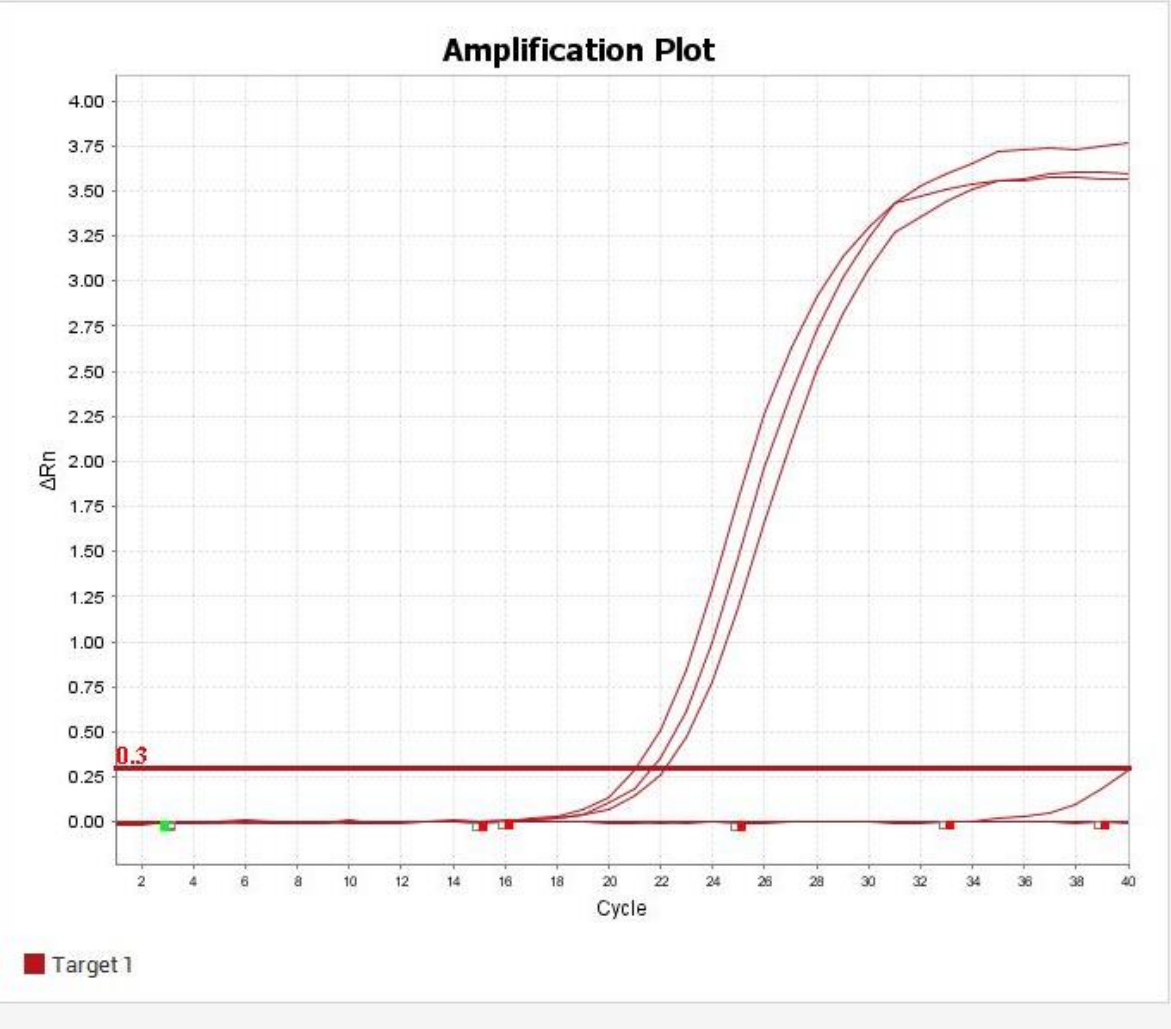


Figure 3.13: A representative amplification plot of the optimised *36B4* gene PCR, showing only the amplification of the golden sample and the NTCs.

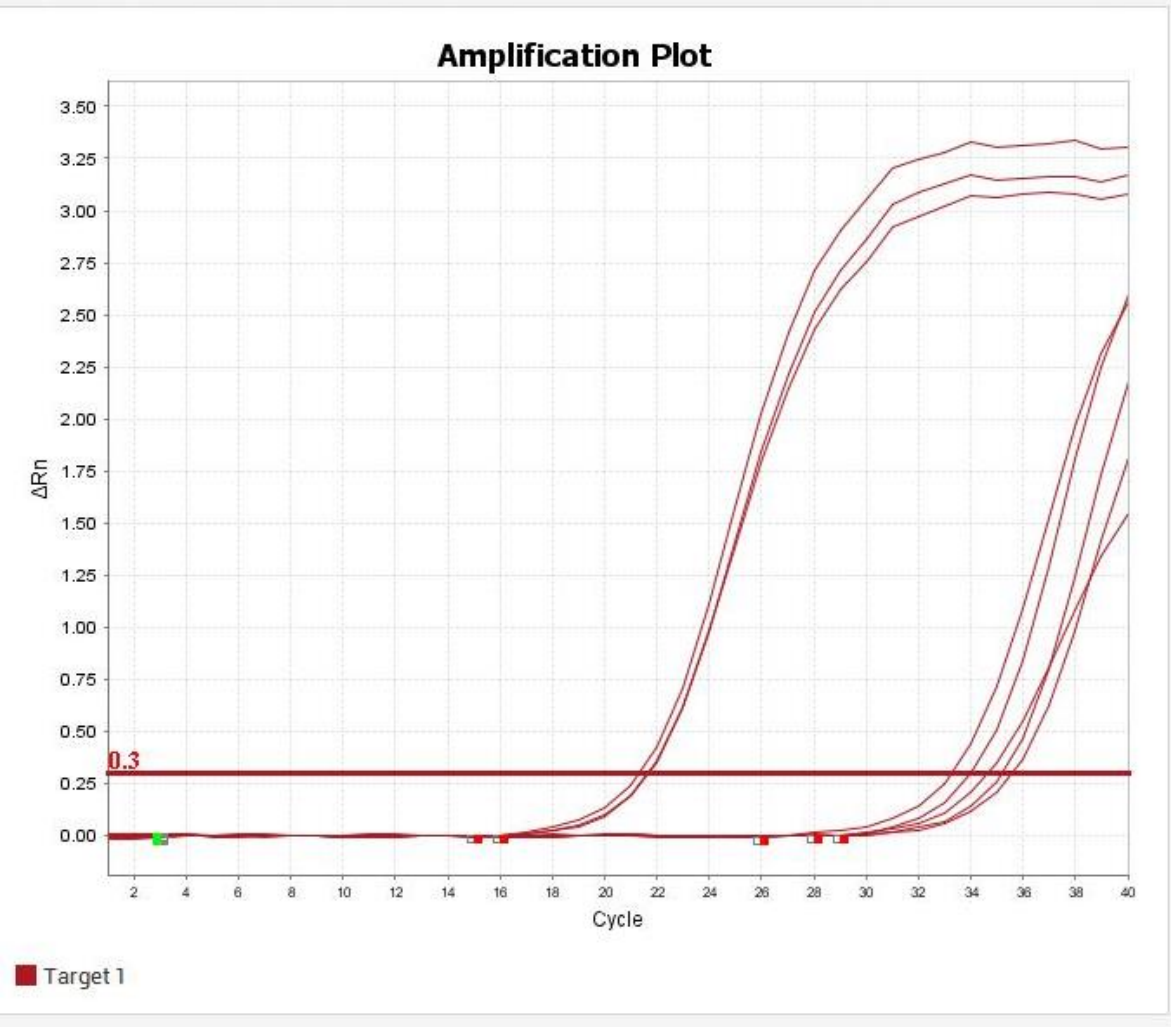


Figure 3.14: A representative amplification plot of the optimised *36B4* gene PCR, showing only the amplification of the golden sample and the NTCs. All the NTCs started to amplify in the late stages of cycling depicting some contamination.

The last of the amplification data analysis was the (high resolution melt-curve) HRM analysis. This analysis was done to confirm that the amplified PCR product was correct and of the correct size, which was deduced from the melting temperature of the PCR product. Figures 3.15 through to 3.18 show the melt-curve plots from the four plates above. Most of the samples are peaking between 79°C and 80°C for the SCG. The melting temperature that was recorded by Cawthon, 2008 was 80°C; hence this proves that the primers were amplifying the correct sequence from the genome (Cawthon, 2008).

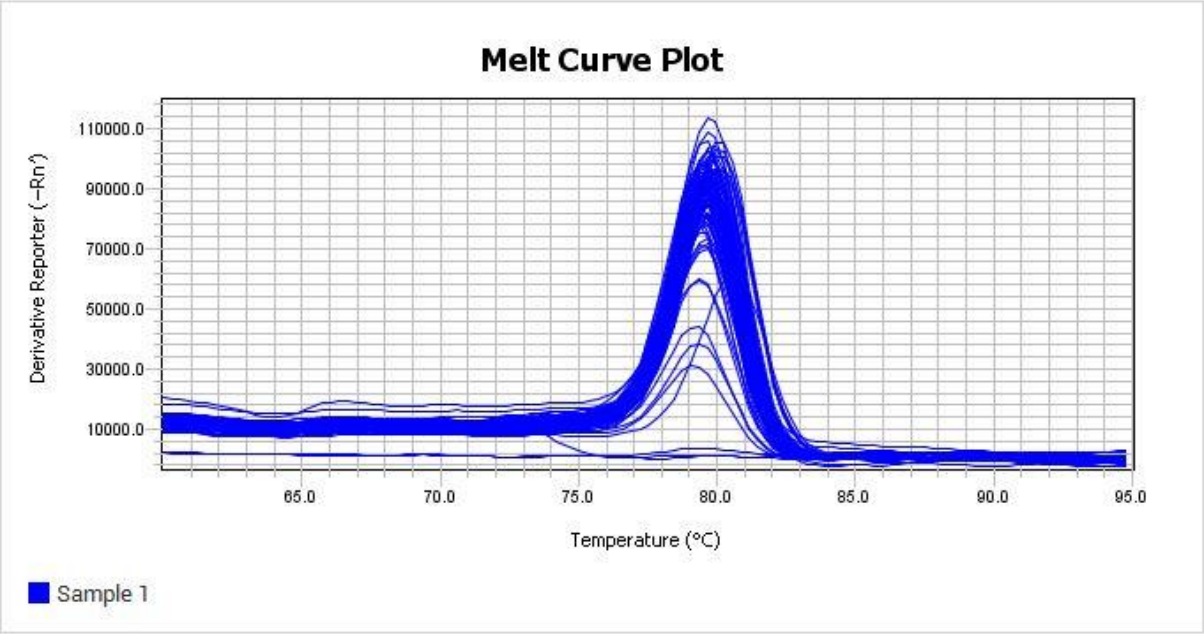


Figure 3.15: Melting curve of *36B4* PCR product, peaking between 79 and 80 °C as expected. The NTCs did not show any peaks as they had no PCR do melt.

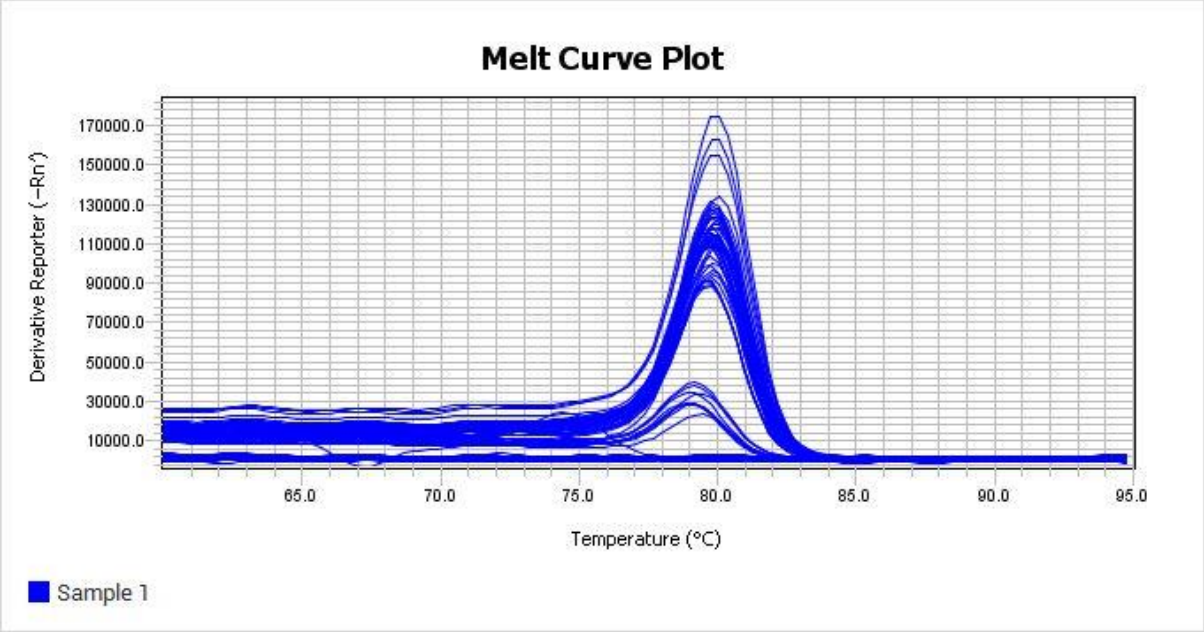


Figure 3.16: Melting curve of *36B4* PCR product, from a consecutive reaction plate with a melting temperature of between 79 and 80 °C for each sample run.

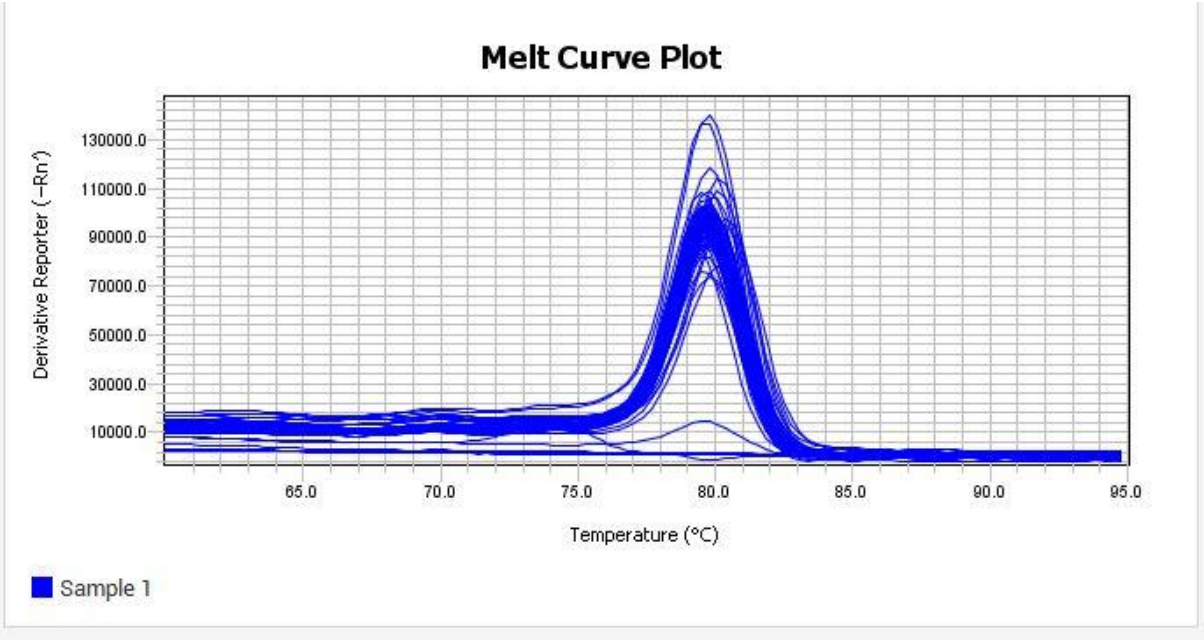


Figure 3.17: Melting curve of *36B4* PCR product, where samples peaked at the expected temperature of 80 °C. As expected, there was no peak seen in the NTCs.

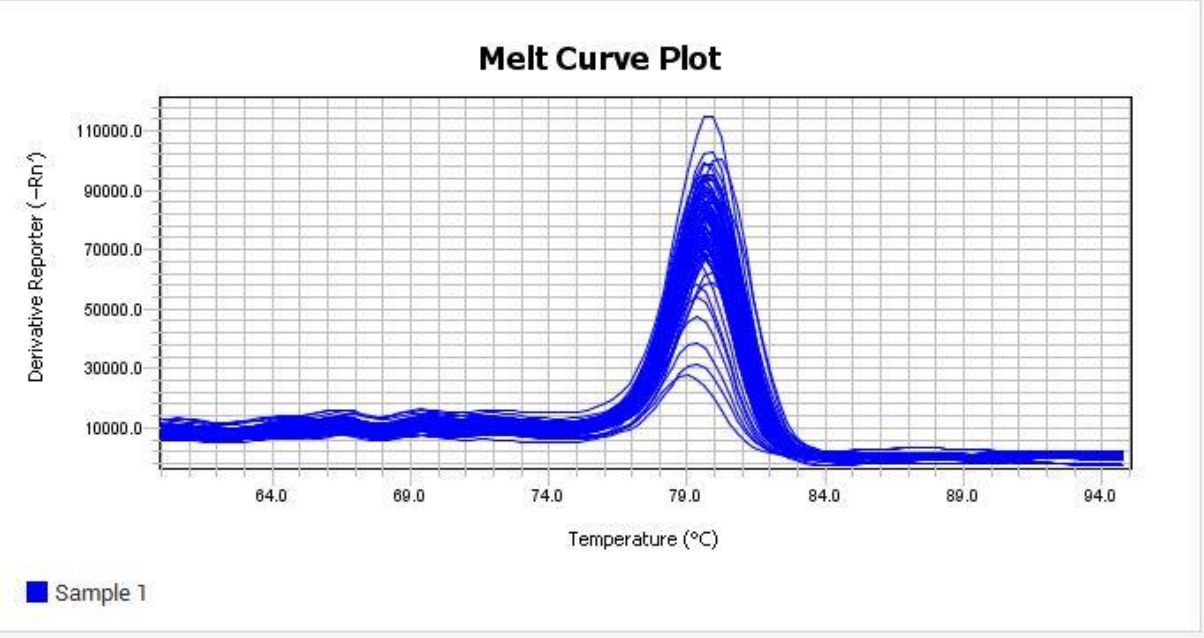


Figure 3.18: Melting curve of *36B4* PCR product, where the NTCs peaked at 79 °C.

The melt-curve plots are the graphic representation of the high-melt resolution, where the PCR products are dissociated. The melting temperature of each sample is then used to infer accuracy of amplification which should be different for different base-pairs. Only a single peak should be seen with all samples that have accurately

and successfully amplified. A peak that deviates far from the melting temperature of choice is deemed inaccurate. The NTCs do not peak; they remained flat-lined as there is no amplification in these wells. However in Figure 3.18 there were peaks visible in the NTCs as well, this would be expected from this reaction plate as it showed amplification of the NTCs in Figure 3.14, confirming that this was contamination of the NTC wells.

3.2.3 Amplification of the telomere copy repeats

Upon completion of the standard curve analysis for the telomere reaction, the study samples were then assayed. Figures 3.19 to 3.30 illustrate the amplification curves of the telomere reaction, the amplification of the golden sample and the NTCs and lastly the HRM analysis for the telomere reaction.

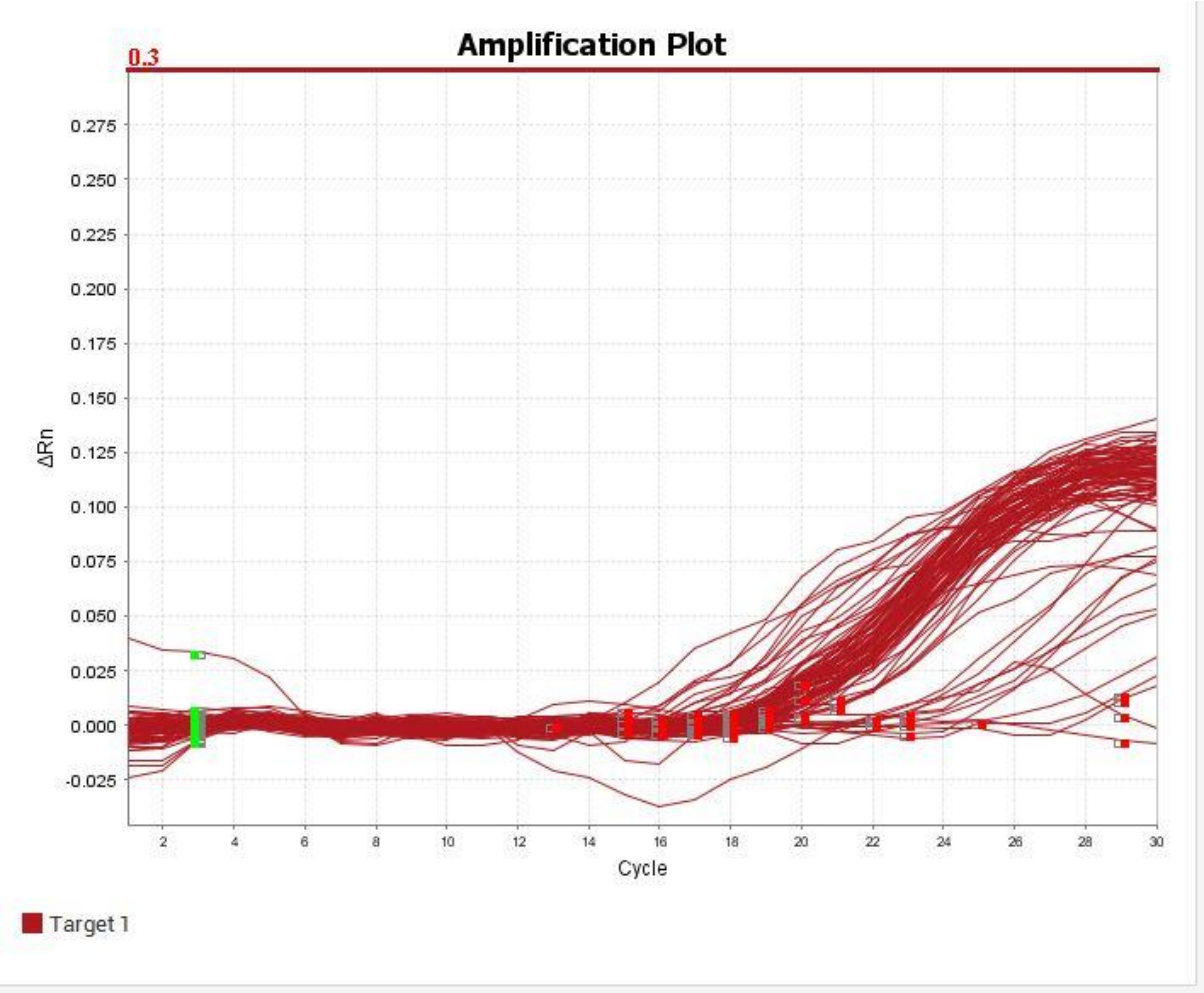


Figure 3.19: The amplification plot generated from the telomere PCR.

The expected S-curved shape was not achieved, meaning that optimal amplification of the desired PCR product was not achieved. It is also crucial to mention that all the amplifications were below the threshold of 0.3.

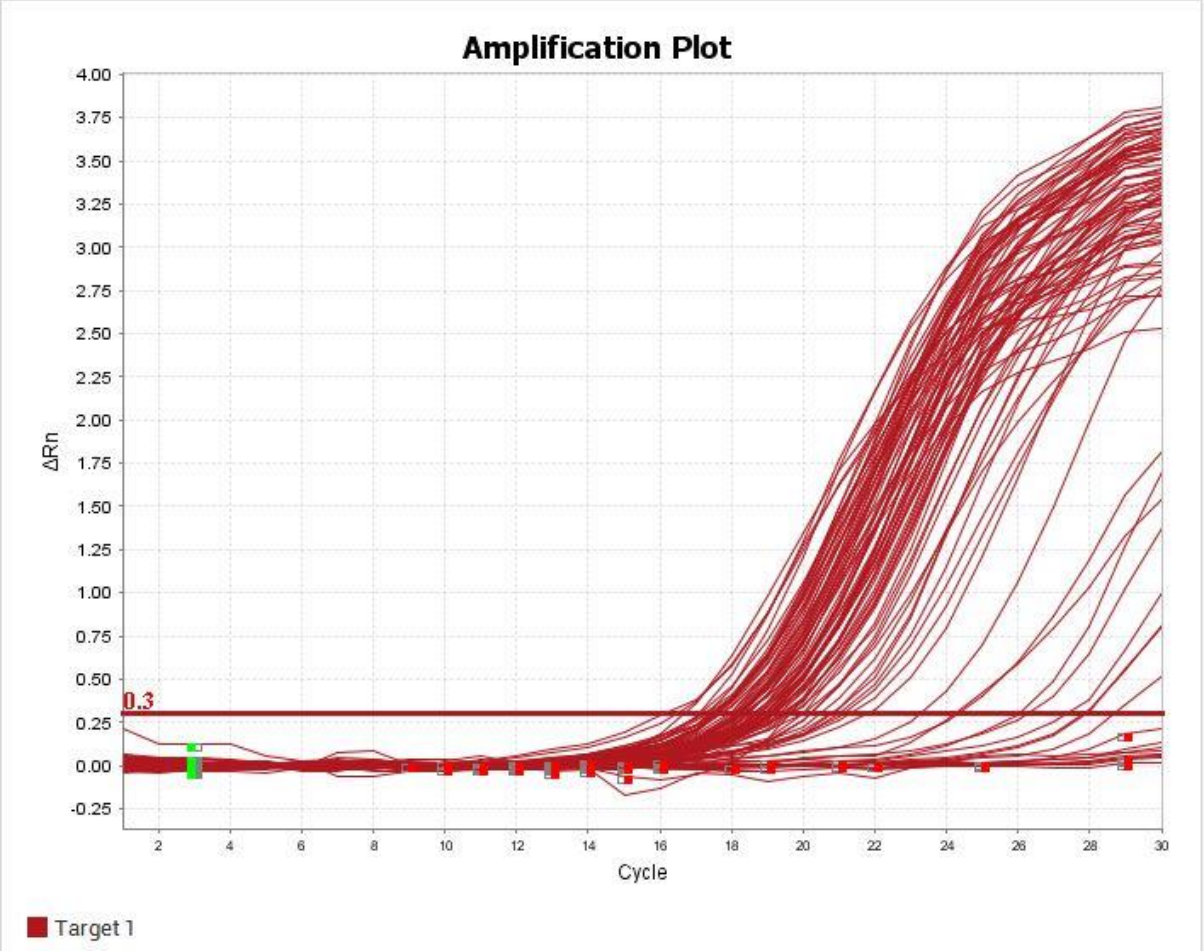


Figure 3.20: The amplification plot generated from the telomere PCR which resulted in the S-shaped curve that is required.

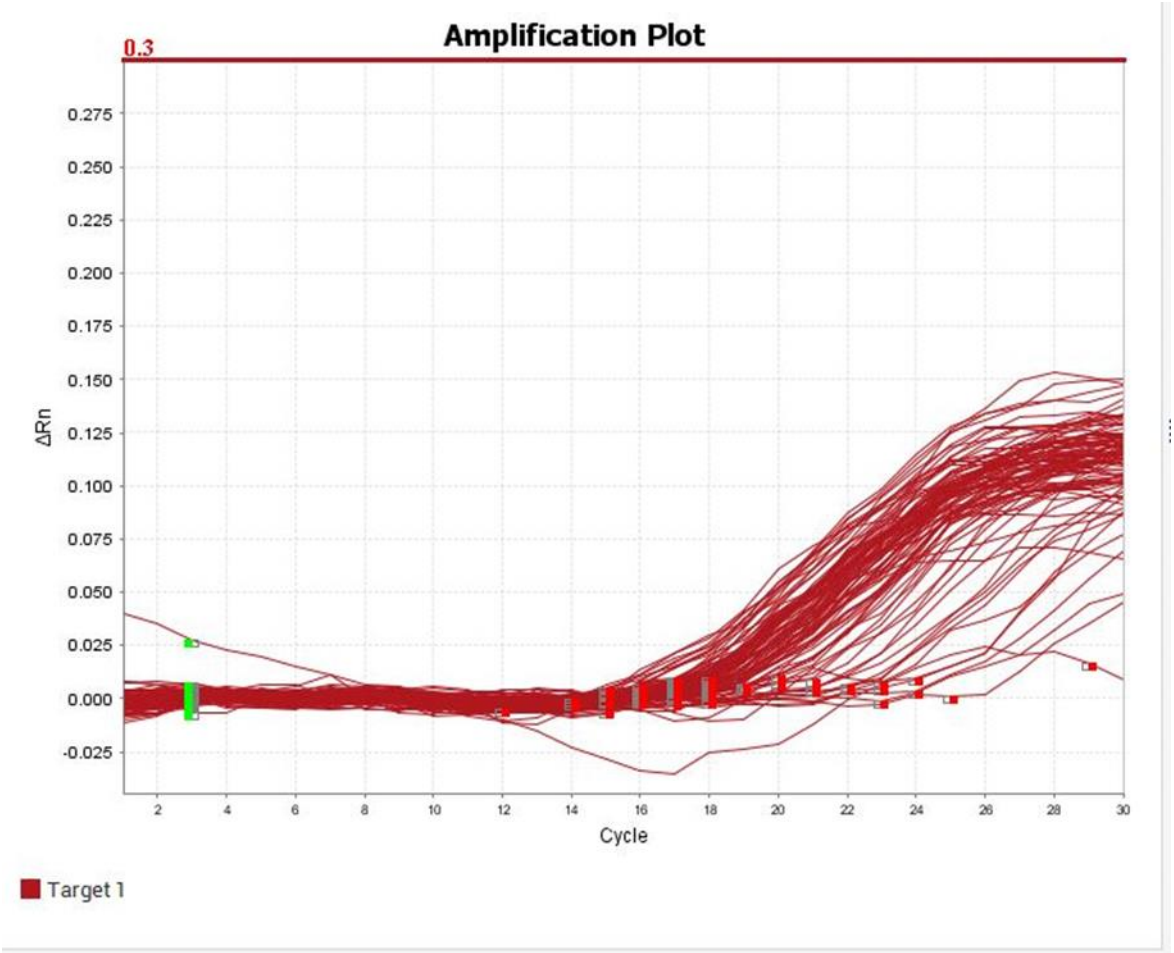


Figure 3.21: The amplification plot generated from the telomere PCR where amplification was yet again below the 0.3 threshold.

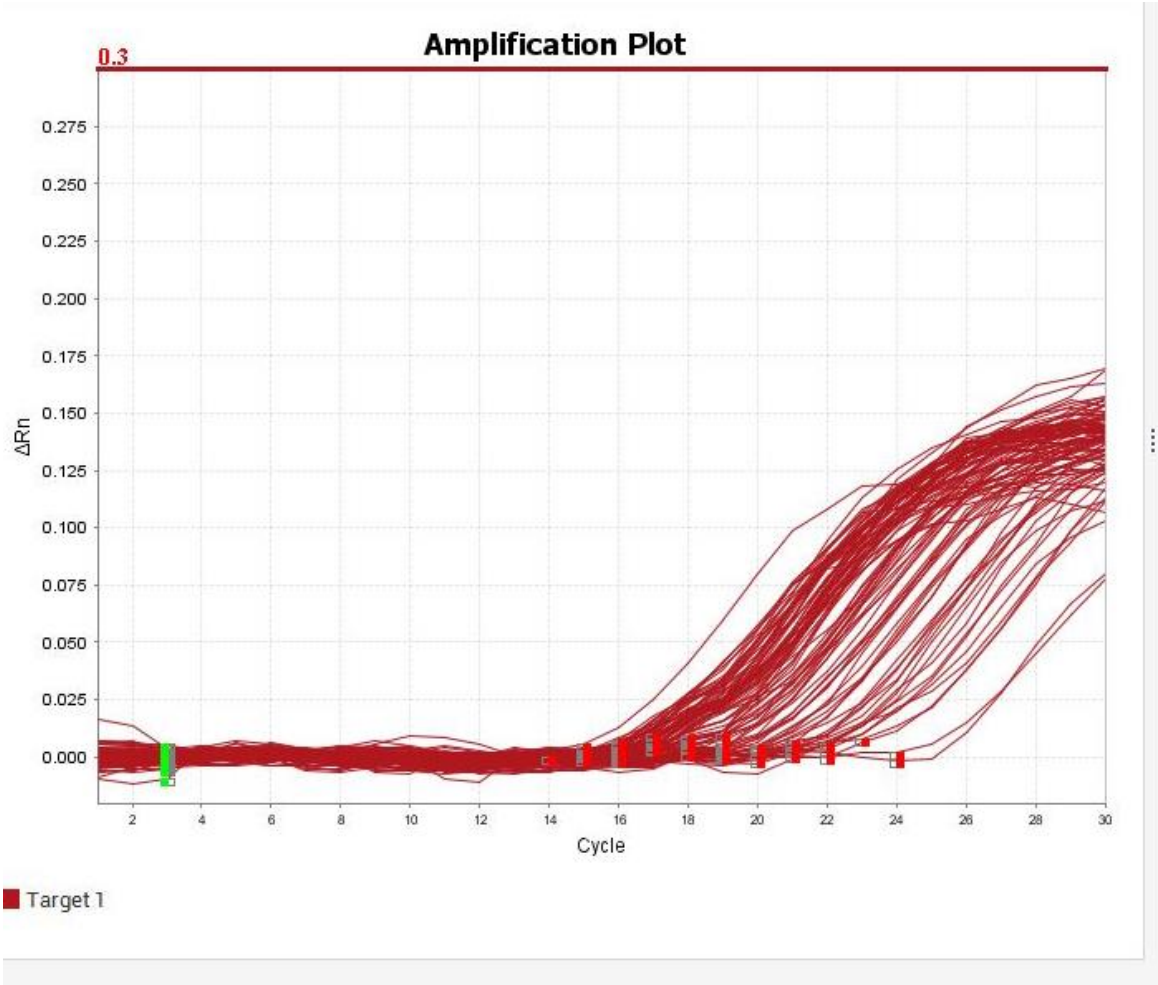


Figure 3.22: The amplification plot generated from the telomere PCR, where the amplification was also below the 0.3 threshold and there was amplification in NTCs, suggesting contamination.

Figures 3.19 to 3.22 show the initial amplification of the telomere aspect of the assay. The threshold was also set as 0.3 as noted by the original protocol. The TL amplification performed rather badly as there was no characteristic S-shaped curve in Figures 3.19, 3.21 and 3.22. The amplification also does not rise above the required threshold except for in Figure 3.20, although most samples in this reaction plate started to amplify late and at varying times, which could reflect inaccuracy of pipetting resulting in varying quantities of DNA in the reaction plate wells. However this reaction plate was the only plate that managed to depict the S-curved shape. This was the only plate which showed amplification beyond the specified threshold.

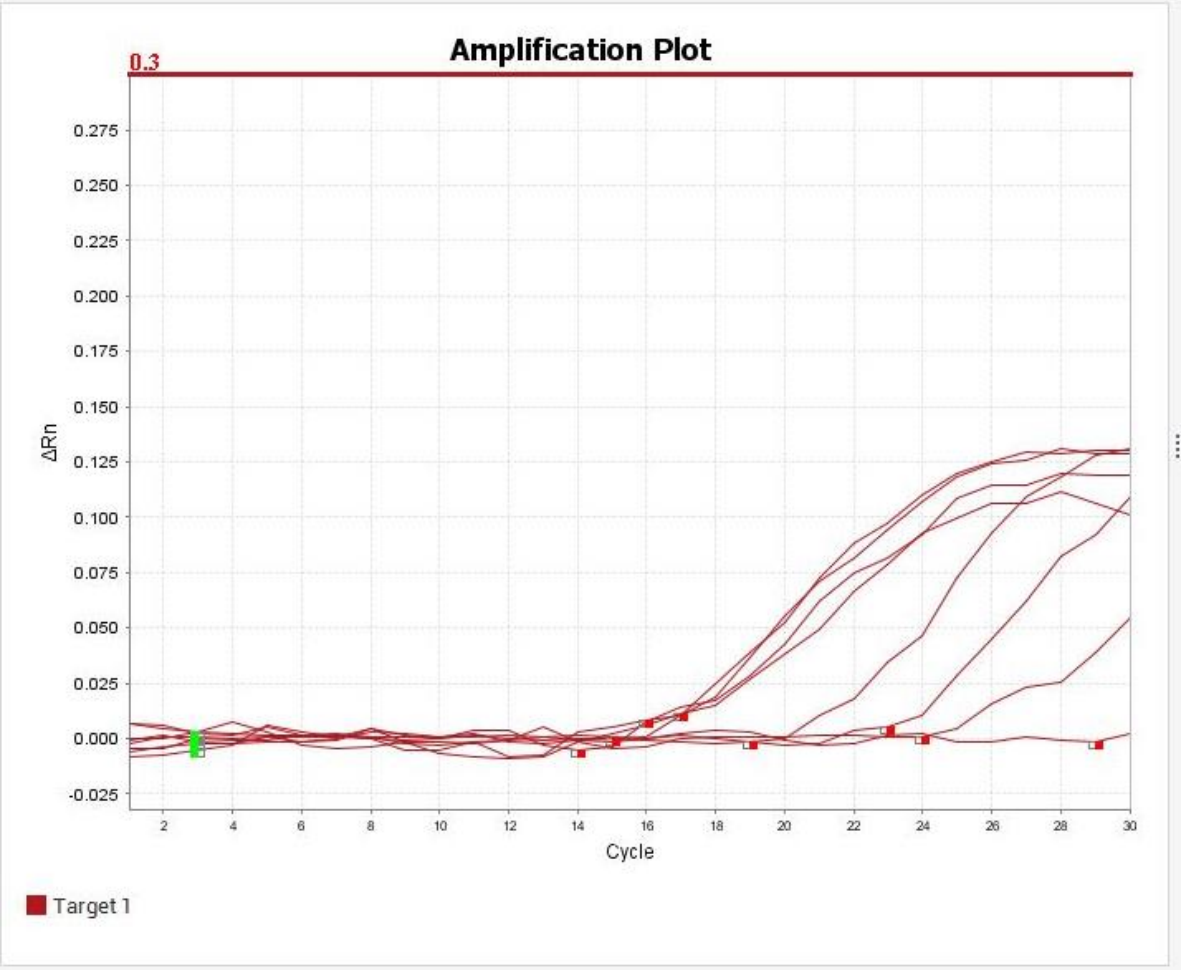


Figure 3.23: The amplification plot generated from the telomere reaction, showing only the golden sample and the NTCs.

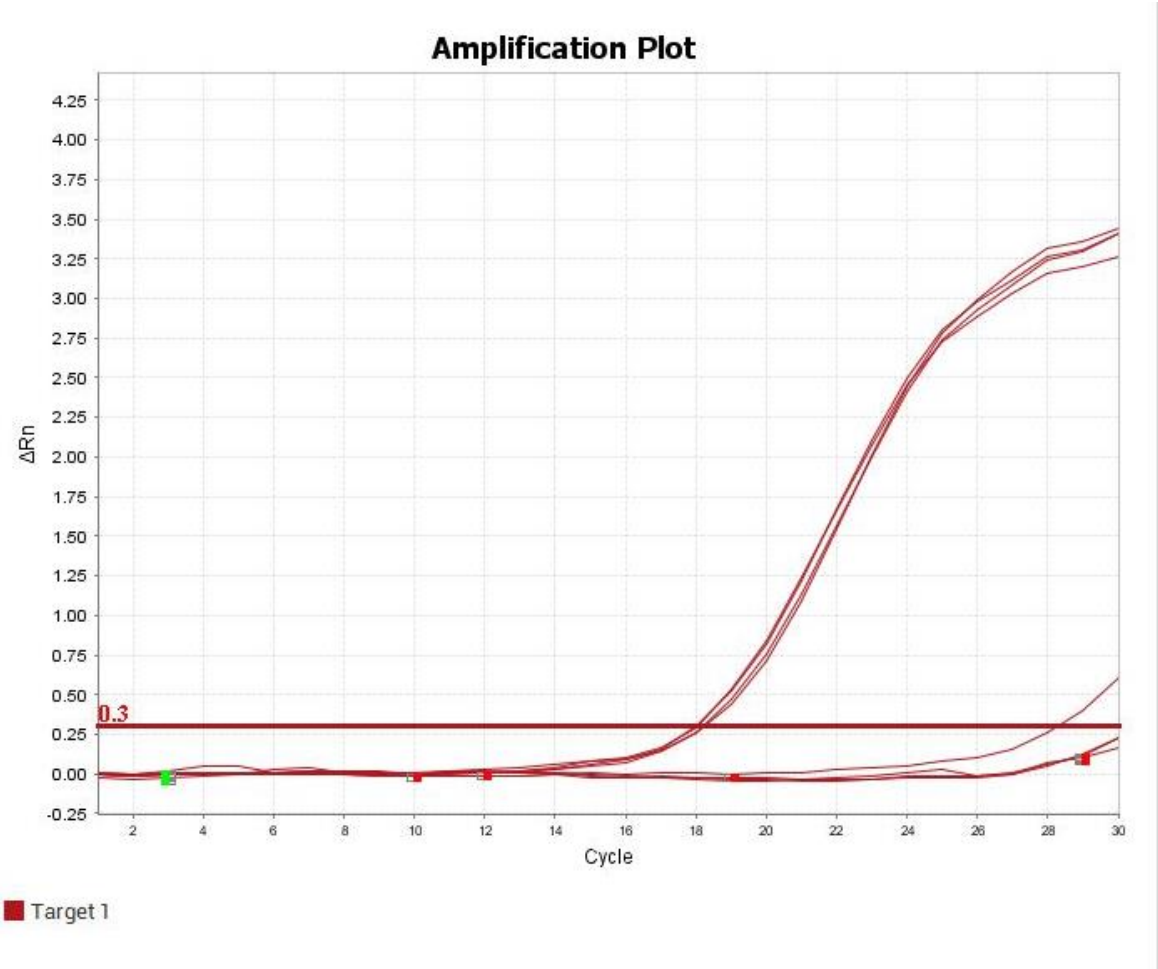


Figure 3.24: The amplification plot generated from the telomere reaction, showing only the golden sample and the NTCs.

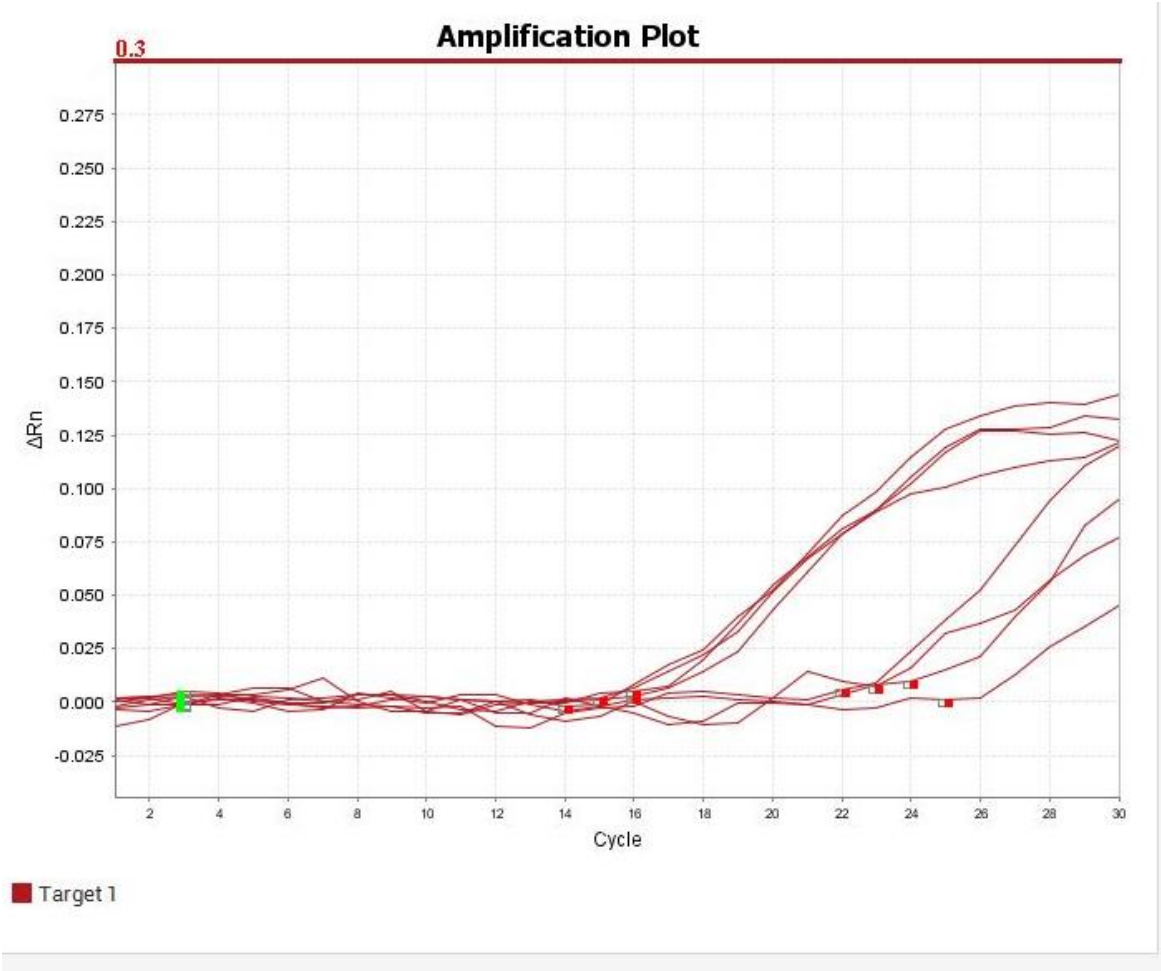


Figure 3.25: The amplification plot generated from the telomere reaction, showing only the golden sample and the NTCs.

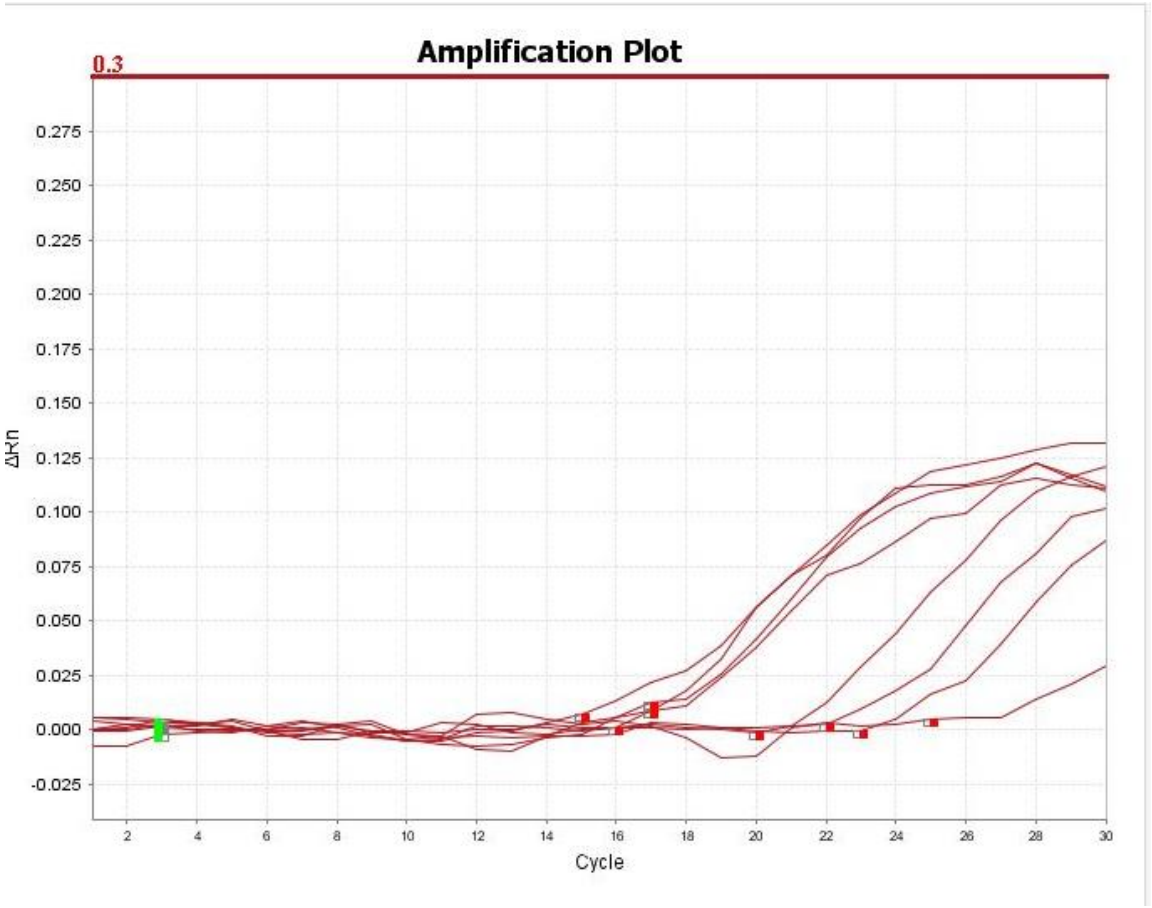


Figure 3.26: The amplification plot generated from the telomere reaction, showing only the golden sample and the NTCs.

Figures 3.23 through to 3.26 show the amplification plots of both the reference sample (golden sample) and the NTCs of the plates from the previous Figures 3.18 to 3.22, respectively. Figure 3.24 shows that there is likely some contamination in the plate of interest as there is a rise from the flat seen towards the end of the PCR cycles, but as they do not cross the threshold line they are regarded as insignificant. The rest of the plots show that amplification was not successful as in the Figures above.

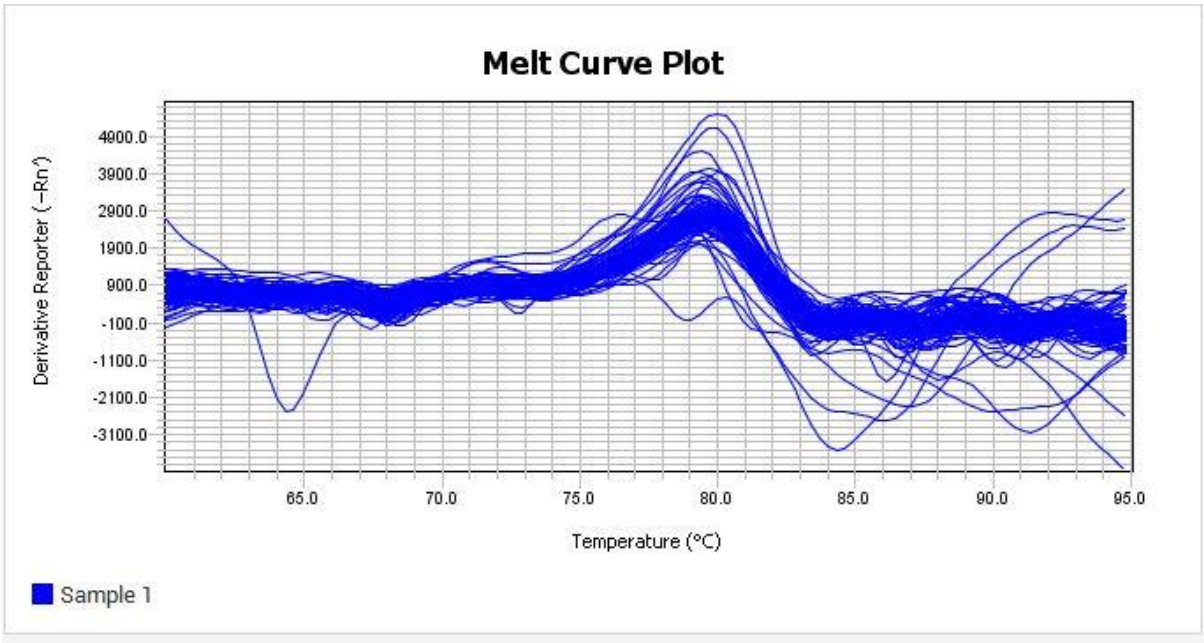


Figure 3.27: A melting curve plot of the telomere product.

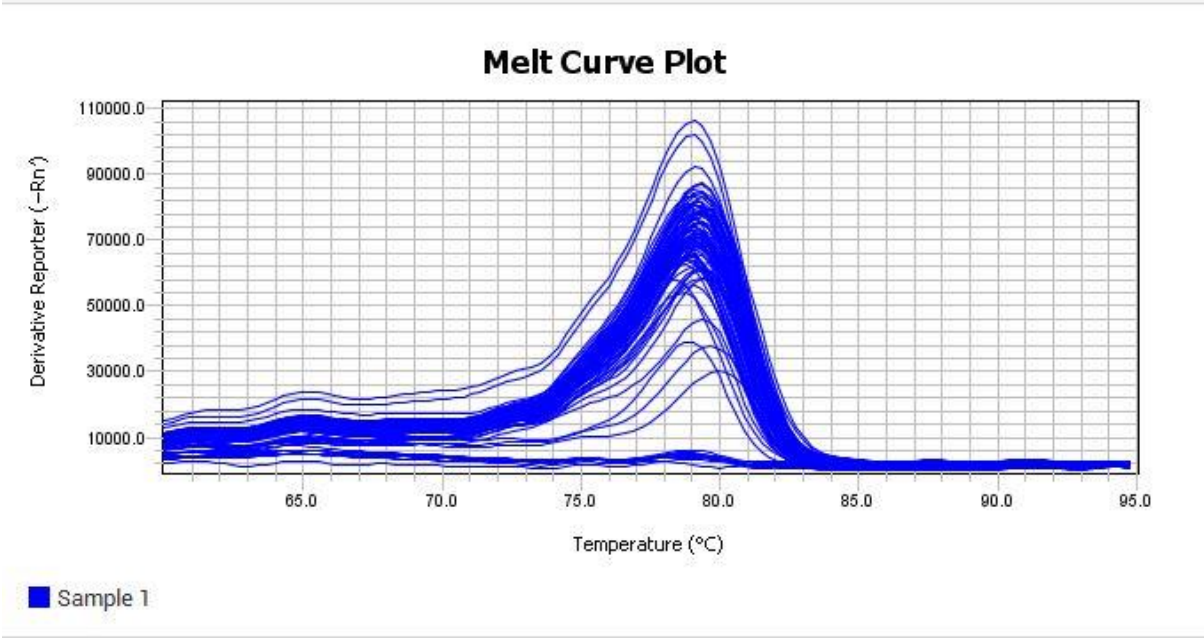


Figure 3.28: A melting curve plot of the telomere product.

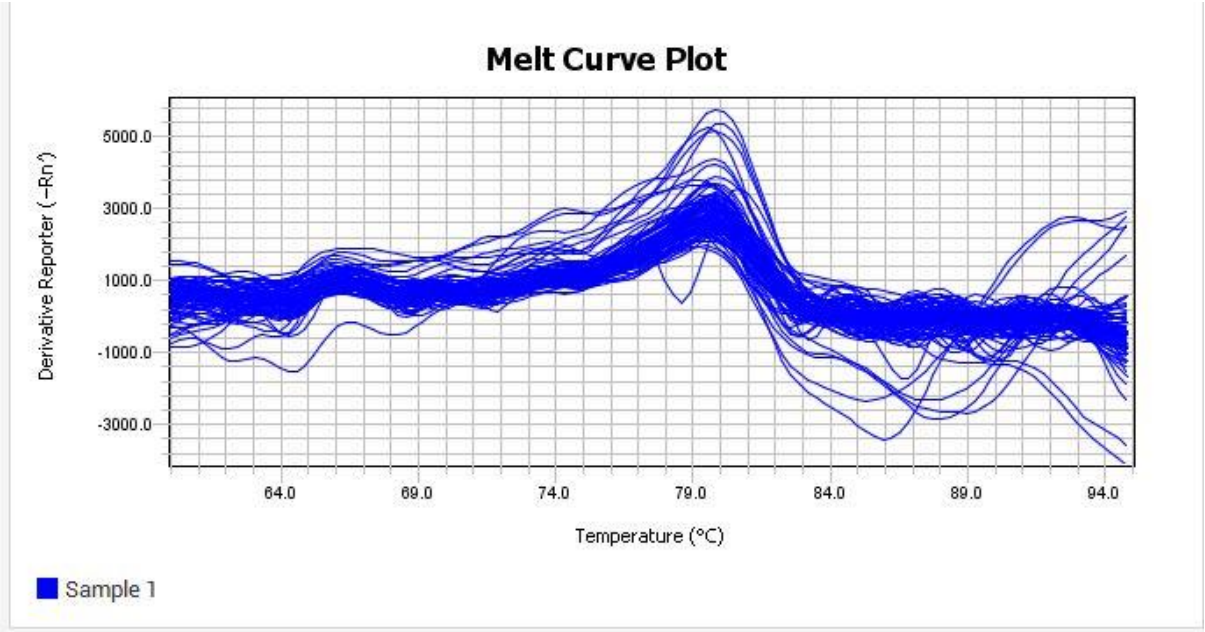


Figure 3.29: A melting curve plot of the telomere product.

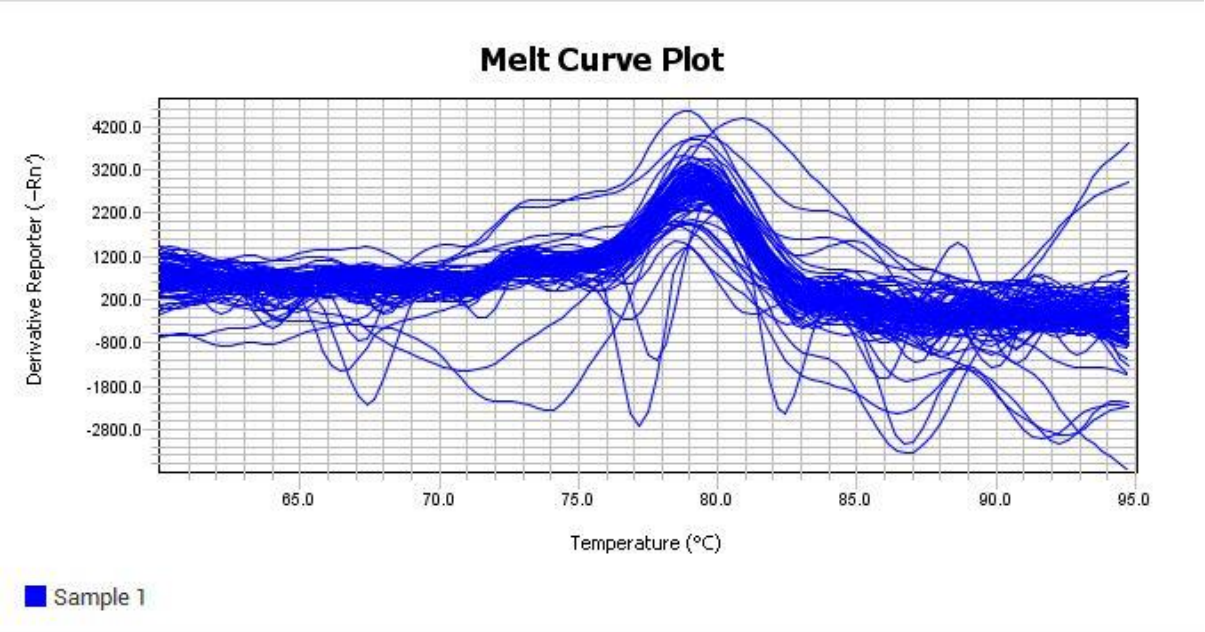


Figure 3.30: A melting curve plot of the telomere product.

Figures 3.25 to 3.28 illustrate the melt-curve plots from the high melt resolution of the telomere reaction. Only Figure 3.26 showed melt-curve analysis that is expected in this type of analysis. Multiple peaks are however seen in the melt-curve plots of Figures 3.25, 3.27 and 3.28, meaning that some samples' amplification was not from the desired telomeric base pairs. Moreover, the peaks are jagged and not uniform as expected, which confirms that the telomeric region of interest was not amplified with these sample plates. The NTCs should not peak at all but little peaks are seen and these would be from the primer dimers.

As shown in the images above, the *36B4* gene and telomere reaction assays did not perform optimally. During troubleshooting, cycling conditions were changed from 30 cycles to 40 cycles, as it was seen by chance or a mistake in one of the telomere plates, that when the sample is cycled longer a better prominent S-shaped curve is seen. From the number of cycles, the annealing temperature was modified in order to find the optimal temperature at which the primers would anneal to the desired region and not to each other and forming primer dimers in the process. Upon failure to clarify these issues through the change in number of cycles and annealing temperature, the decision was made to purchase a new set of primers.

3.3 Troubleshooting for PCR assay

From the first batch of primers the SCG (*36B4*) reaction performed optimally. The telomere reaction resulted in non-specific amplification. As shown in the flow diagram (Figure 3.31), firstly annealing temperature was decreased, and then increased by a temperature range of 2°C (from 50°C to 60°C). That resulted in no change from the previous results. Decreasing the concentration of the primers according to a serial dilution of 1/3 at 3 different concentrations of the telomere primers did not change the results. Upon purchasing a new batch of primers, a new problem was encountered, there was no amplification suggesting no product. After the decision to decrease sample concentration, preparing new solutions and reagents, all reaction tubes were autoclaved prior to use and the reaction was performed with 40 cycles instead of 30, but that made no difference. Upon purchase of the third batch of primers, there was persistent contamination of the NTCs, so fresh filtered tips were used and a designated area for the addition of DNA samples into the wells was set up with the help of a Thermo Fisher sales representative that

had been requested to assist. Furthermore, a fresh pair of gloves for each step was used to reduce possible contamination along with fresh solutions and reagents. In all the final attempts to get the assay to work, I consulted with Dr Daniel Nussey (who held a TL diversity workshop that I attended) who suggested that I try all the iterations, but this was to no avail. Lastly adjusting the components that made up the master mix (e.g. magnesium concentration) were not feasible as the fluorescent dye (SYBR Green) came mixed with all the PCR master mix components in a single tube.

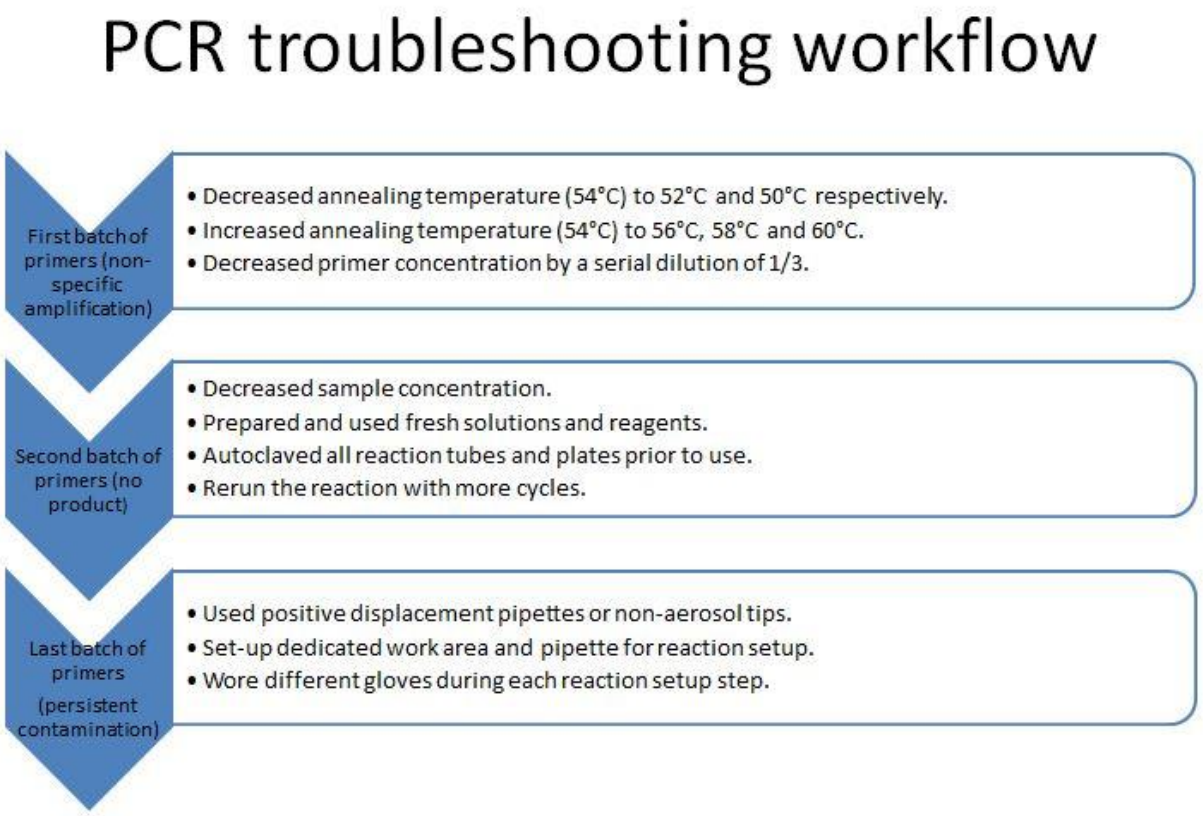


Figure 3.31: PCR troubleshooting workflow.

3.4 Optimisation of assay

The aim of the optimisation was to decipher the problem specifically with the telomere reaction primers, as they gave the most inaccurate results. The SCG amplification for the quantitative PCR assay was performing well; but the telomere

assay required further optimisation to get it to work at a satisfactory level. A limited number of images are shown and were chosen because they worked well, while showing some of the failures. Prior to optimisation an agarose gel was run to verify that the PCR product for the telomere assay was not the desired product. Figure 3.32 shows the gel electrophoresis of both the *36B4* and telomere reactions flanked by 1kb ladder and 50bp ladders. It was not possible to accurately judge the size of the PCR products because the bands were not clear on the gel. The expected sizes of the telomere and *36B4* amplicons were 79 base pairs, and 106 base pairs respectively. Unfortunately there was a PCR product in each of the negative controls (water only), indicating contamination. This was not evident in the amplification curves, because different reactions were used for the amplification curves and gel electrophoresis. The bands seen in the negative control could have also resulted from non-specific amplification of PCR products and primer dimer formation.

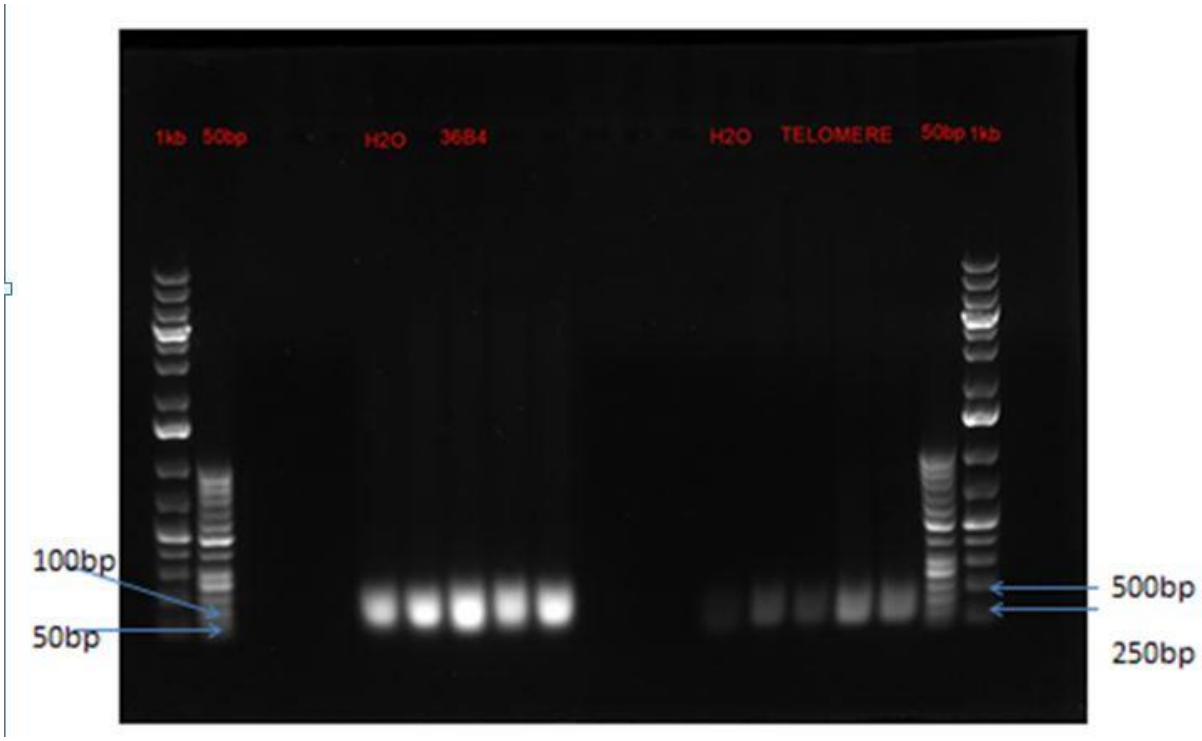


Figure 3.32: Agarose gel electrophoresis of *36B4* and telomere reaction PCR products. The horizontal lanes from the left read: 1kb ladder, 50bp ladder, H2O (water), *36B4* gene, H2O (water), TELOMERE, 50bp ladder and 1kb ladder.

A new set of primers was purchased for the second attempt at optimising the telomere assay. The single gene assay was working and did not require further optimisation. Test DNA samples from the SBIMB Biobank were used in this

optimisation process to prevent depletion of the participant samples during the optimisation process. The telomere reaction was run using the same conditions as before with only new primers and with this attempt, the assay still did not perform at a satisfactory level. Different annealing temperatures, different sample DNA concentrations, change in primer concentrations, and removal of DTT from the master mix did not improve the results. There was no PCR amplification in any of the different scenarios. The conclusion was that the problem arose through faulty primer manufacture, and a third and last batch of primers were purchased for the telomere reaction.

With the third batch of telomere primers, amplification was observed, and is illustrated in Figure 3.33. At first glance, one might say the telomere reaction finally worked but upon evaluation of the standard curve and HRM analysis, the assay was still not correct.

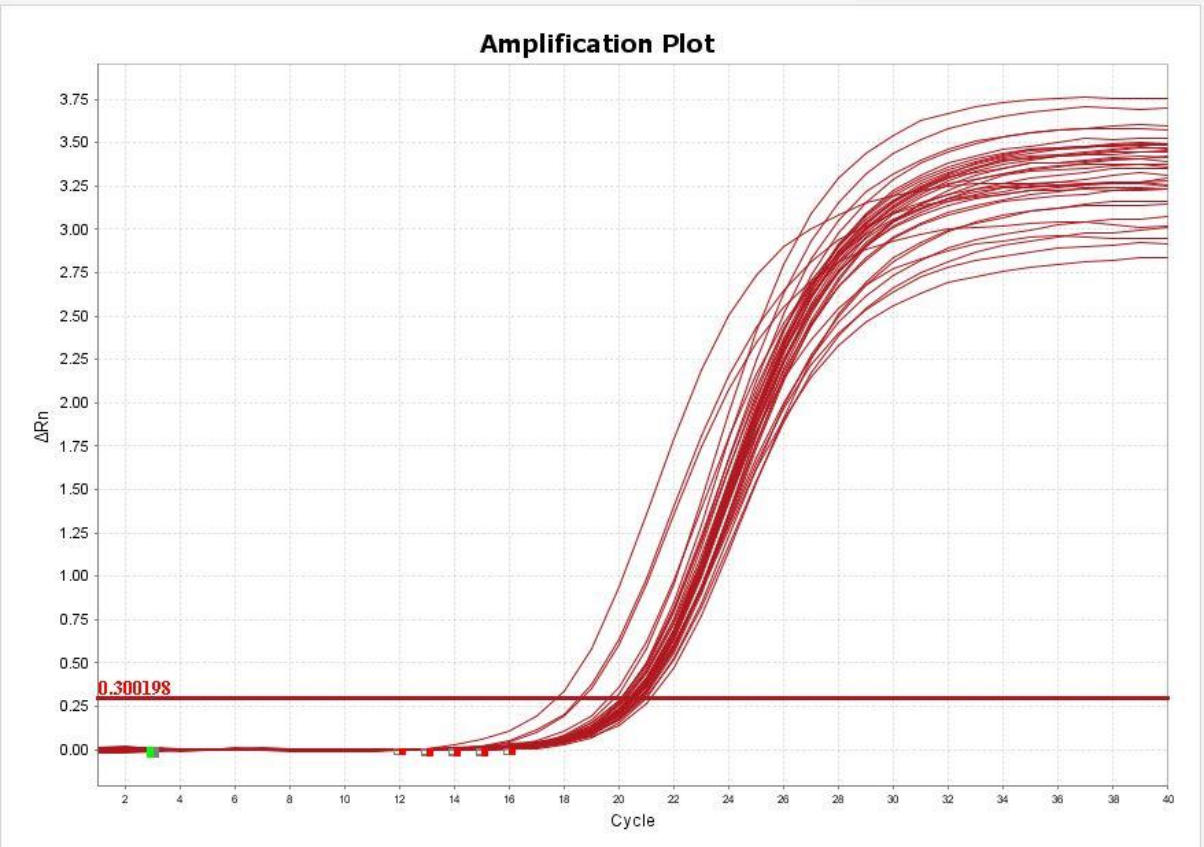


Figure 3.33: Linear amplification plot of the telomere reaction, from the new set of primers was purchased.

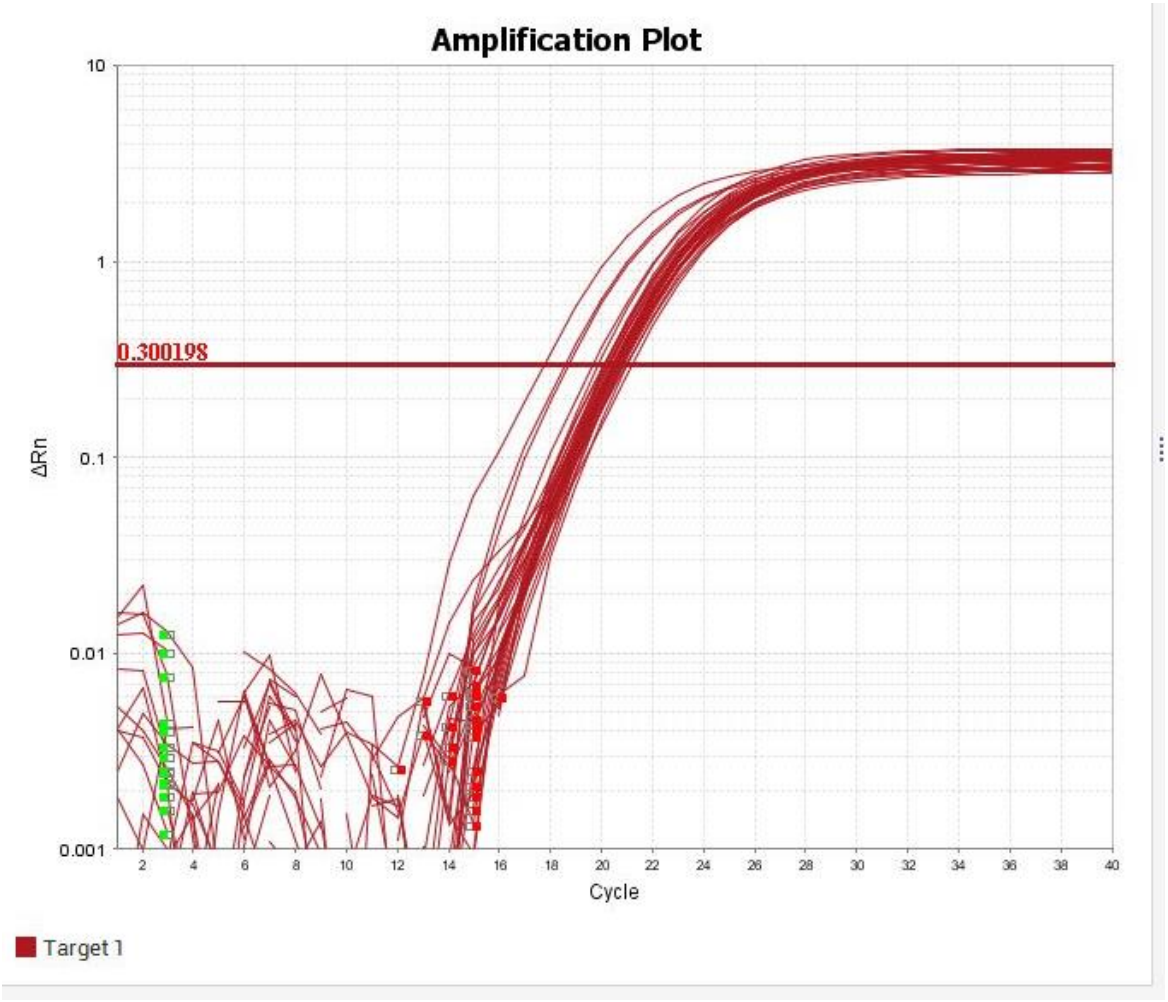


Figure 3.34: Log-transformed amplification of the telomere PCR from the new set of primers purchased.

Figures 3.34 and 3.35 show the linear amplification and the log-transformed amplification, respectively, of the last batch of telomere primers. The samples appear to amplify at the same range, despite the fact that different concentrations of DNA were used as was with the standard curve dilutions. This shows that the amplification seen is not of the desired telomere product. The amplification curves were supposed to group with each other in triplicates to show the amplification of each respective concentration from the serial dilution.

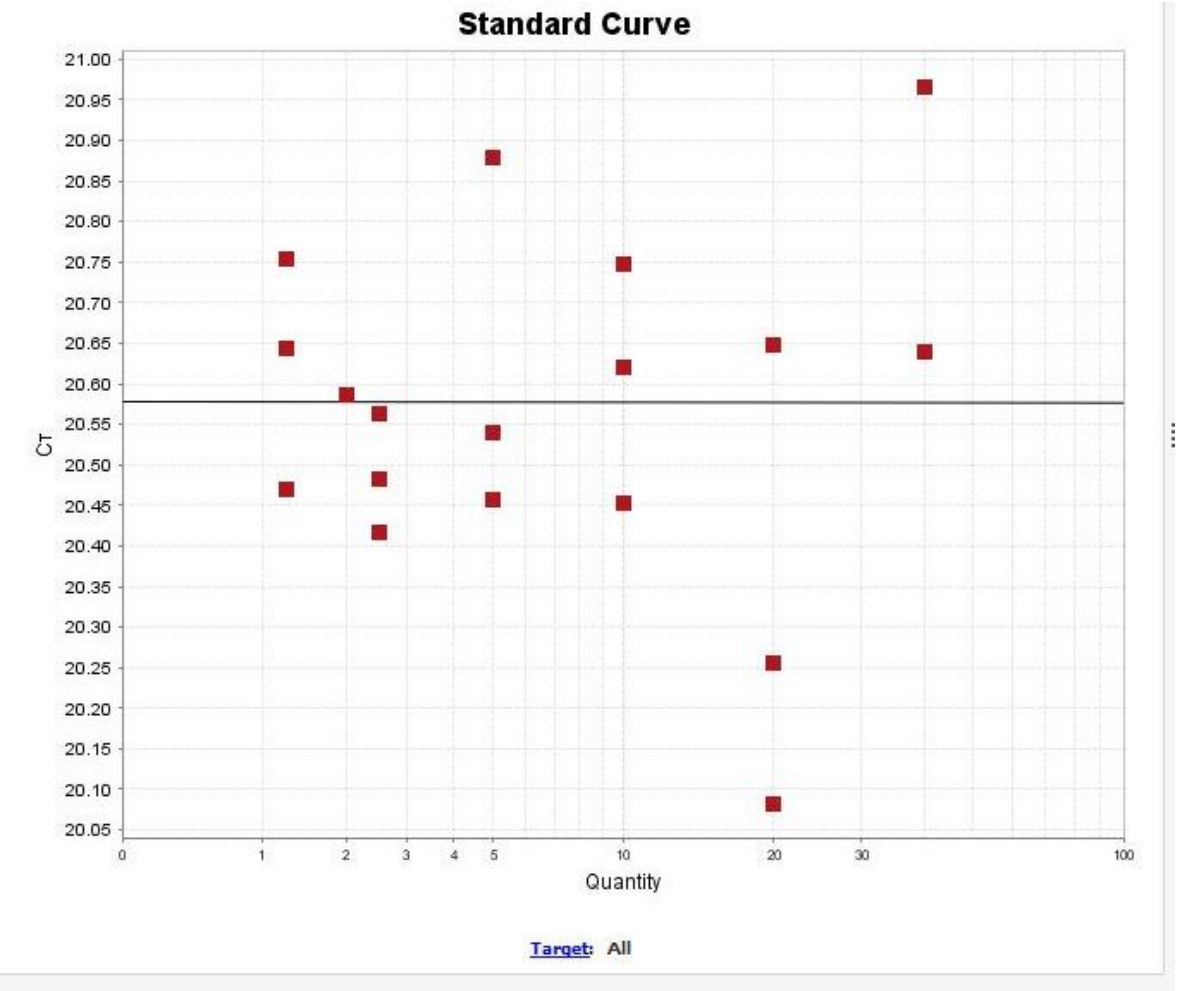


Figure 3.35: The standard curve from the last batch of primers purchased.

The regression line in Figure 3.35 has no slope, suggesting that the standard curve is not accurate and hence the amplification of the telomere product is not actually the telomere repeats but primer dimer formation.

Upon many attempts to get the telomere reaction to work accurately, it failed. Hence the measurement of TL was not possible as the telomere reaction is a crucial aspect for the calculation of the T/S ratio. An appropriate human reference gene was chosen as the single copy control against which to measure the telomere amplification. Relative quantification of TL could not be achieved if the reference gene behaves as a copy number variant, in the way that telomeres do.

3.5 Genetic association study using AWI-Gen Genotype data

There were 53 variants identified from Caucasian, Asian, and African American populations in the literature that were associated with TL. All these had reached

genome-wide significance in their respective GWASs. Twenty eight of these SNPs were present in the H3Africa array data and of the SNPs present, 16 were genotyped and 12 were imputed. Table 3.3 lists all the SNPs that were used in the association. Moreover, the table describes the study where the SNP was discovered, including the method of telomere measurement, ethnicity, the age range of the participants, and the outcome of the study.

3.6 Sample and Genotyping QC

H3A Bionet developed a nextflow pipeline for the H3A Consortium to use for all of the GWAS. The AWI-Gen GWAS analysis team at the SBIMB optimised the pipeline for analysis of the AWI-Gen dataset. This meant that original cut-off values were modified for each of the QC steps both in terms of sample, and SNP QC. In terms of samples we started with 836 participants. No participants were removed due to sex discordance and twenty-six had more than 2% missingness. Twenty samples were removed during the heterozygosity check and 10 individuals were removed due to having a pi hat value above 0.8. For SNP QC none of the SNPs were duplicates. Six SNPs were removed because they were non-autosomal. A further 6 SNPs had a MAF of less than 0.01, or were monomorphic. Seven SNPs were removed due to having missingness above 0.01, and another 6 SNPs had a p-value below the Hardy Weinberg cut-off of $p=5 \times 10^{5}$. After QC we ended up with 770 individuals and 28 SNPs. All the remaining 28 SNPs showed no differential missingness, which checks whether controls and cases differ in missingness. Data from a total of 28 SNPs and 770 individuals were used for further analysis.

3.7 Summary of genes and SNPs analysed using the H3Africa chip

Table 3.3 below lists the 28 SNPs analysed using the H3Africa array, with the gene involved, population in which the study was first conducted, aim and outcome of the study. Among the genes represented, only the *TERC* and *TERT* genes which have direct functions in telomere biology. *TERC* had 5 SNPs (rs10936599, rs12638862, rs1093660, rs1317082 and rs1920116) representative of the gene in this study. *TERC*, which stands for telomerase RNA component on RNA gene, encodes a protein that forms part of the enzyme telomerase. The second gene of importance is the telomerase reverse transcriptase (*TERT*) gene that also encodes a protein that

forms part of the telomerase enzyme. This gene had 1 SNP representative in this study (rs2736100). The rest of the genes in which the SNPs reside have no specific known role in telomere biology, they are however implicated with the cell cycle and apoptosis. Of the 13 SNPs that were found in African American populations, only 3 SNPs (rs11703393, rs594119, rs2300383) were analysed on the H3Africa array, and none of them were of any statistical significance.

Table 3.3 List of SNPs associated with TL with gene implicated, ethnicity and outcome

Reference	Genes with associated SNPs	Aim of the study	Ethnicity	Age range (years)	Outcome
Codd *et al.*,2013	*TERC*; rs10936599	Mean TL and NCDs	Caucasian	17 - 103	Seven SNPs associated with TL and other diseases
Codd *et al.*,2013	*TERT*; rs2736100	Mean TL and NCDs	Caucasian	17 - 103	Seven SNPs associated with TL and other diseases
Codd *et al.*,2013	ACYP2; rs11125529	Mean TL and NCDs	Caucasian	17 - 103	Seven SNPs associated with TL and other diseases
Codd *et al.*,2013	OBFC1; rs9420907	Mean TL and NCDs	Caucasian	17 - 103	Seven SNPs associated with TL and other diseases
Delgado *et al* 2018	PRR20; rs9537514	To enhance our understanding of genetic determinants of	South Asian	17 - 75	Novel second association signal at the RTEL1 locus
Delgado *et al* 2018	*TERC*; rs12638862	Enhance our understanding of genetic determinants of TL across populations	South Asian	17 - 75	Novel second association signal at the RTEL1 locus
Gu *et al.*,2011	WDR65; rs621559	Association of TL and cancer	Caucasian	NA	Found SNP rs398652 associated with longer TL and decreased bladder cancer risk

Telomere length = TL, African American = AA

Table 3.3: List of SNPs associated with TL with gene implicated, ethnicity and outcome…continued

Reference	Genes with associated SNPs	Aim of the study	Ethnicity	Age range (years)	Outcome
Gu *et al.*,2011	PELI2; rs398652	Association of TL and cancer	Caucasian	NA	Found SNP rs398652 associated with longer TL and decreased bladder cancer risk
Gu *et al.*,2011	DHX35; rs6028466	Association of TL and cancer	Caucasian	NA	Found SNP rs398652 associated with longer TL and decreased bladder cancer risk
Lee *et al.*,2014	SYT16; rs4902100	To identify TL variation SNPs in long lived families	Caucasian	29 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2 on 4q30
Lee *et al.*,2014	DKK2; rs7680468	To identify TL variation SNPs in long lived families	Caucasian	24 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2
Lee *et al.*,2014	LOC100128993; rs11787341	To identify TL variation SNPs in long lived families	Caucasian	24 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2
Lee *et al.*,2014	FXYD4; rs10466239	To identify TL variation SNPs in long lived families	Caucasian	24 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2
Lee *et al.*,2014	TMPRSS7; rs16859140	To identify TL variation SNPs in long lived families	Caucasian	29 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2

Telomere length = TL, African American = AA

Table 3.3: List of SNPs associated with TL with gene implicated, ethnicity and outcome…continued

Reference	Genes with associated SNPs	Aim of the study	Ethnicity	Age range (years)	Outcome
Levy *et al.*,2010	ZNF676; rs1975174	GWAS of TL	European and AA	20 - 95	Association identifies OBFC1 as a locus involved in human TL biology
Liu et., 2014	KRT80; rs17653722	GWAS of mean TL, 2632	Chinese	53 - 71	Identified two loci associated with TL
Mangino *et al* 2012	*TERC*; rs1317082	Genome-wide meta-analysis	Caucasian	18 - 101	CTC1 and ZNF676 as genes regulating telomere homeostasis in humans
Mangino *et al* 2012	BRUNOL4; rs2162440	Genome-wide meta-analysis	Caucasian	18 - 101	CTC1 and ZNF676 as genes regulating telomere homeostasis in humans
Mangino *et al.*,2015	DCAF4; rs2535913	To establish whether DCAF4 (DDB1 and CUL4-associated factor 4) is associated with TL	European	18 - 95	DCAF4, a novel gene associated with leucocyte TL
Prescott *et al.*,2011	*TERC*; rs10936601	To identify common genetic variants associated with TL	European	43 - 82	Telomere biology associated SNPs, previously found
Saxena et 2014	CSNK2A2; rs74019828	Association with relative LTL and cardio metabolic traits	South Asian	> 60	Shorter TL with T2D and CVD

Telomere length = TL, African American = AA

Table 3.3: List of SNPs associated with TL with gene implicated, ethnicity and outcome…continued

Reference	Genes with associated SNPs	Aim of the study	Ethnicity	Age range (years)	Outcome
Walsh *et al* 2014	*TERC*; rs1920116	GWAS to explore TL variants associated with glioma risk	Caucasian, Hispanics, AA	>18	Variants near *TERC* and *TERT* genes are associated with glioma risk
Saxena et 2014	C5orf42; rs2098713	Association with relative LTL and cardio metabolic traits	South Asian	> 60	Shorter TL with T2D and CVD
Walsh *et al* 2014	RTEL1; rs6010620	GWAS to explore TL variants associated with glioma risk	Caucasians, Hispanics, AA	>18	Variants near *TERC* and *TERT* genes are associated with glioma risk
Zeiger *et al.,*2018	PARVB; rs11703393	Explore telomere variants in paediatric populations	AA	8 - 20	Novel association between rs143898 and TL
Zeiger *et al.,*2018	NKAIN2; rs594119	Explore telomere variants in paediatric populations	AA	8 - 20	Novel association between rs143898 and TL
Zeiger *et al.,*2018	ITSN1; rs2300383	Explore telomere variants in paediatric populations	AA	8 - 20	Novel association between rs143898 and TL
Lee *et al.,*2014	ASCC2; rs73394838	To identify TL variation SNPs in long lived families	Caucasian	24 - 110	Novel intergenic SNP rs7680468 located near PAPSS1 and DKK2

Telomere length = TL, African American = AA

3.8 Case-Control association analysis

Logistic regression was carried out to determine whether there was a SNP association that significantly differentiated T2D cases from matched controls. Allelic association was carried out using PLINK version 1.9. None of the SNPs were significantly associated with T2D, tested under the Fisher model. Table 3.4 shows the allelic association of the 28 TL associated variants that were tested in this cohort and the genes in which they were found. As the *p* values and the OR show, there was no association between these SNPs and T2D. Table 3.5 illustrates the logistic regression analysis results.

Table 3.4: Allelic association of 28 variants adjusted for covariates sex, age, BMI and geographic region in 770 individuals with allele and genotype frequencies

SNP	Cases					Controls					P-value	L95	U95
	Genotypes			Alleles		Genotypes			Alleles				
	A/A	A/B	B/B	A	B	A/A	A/B	B/B	A	B			
rs621559	54 (0.139)	172 (0.442)	163 (0.419)	280 (0.360)	498 (0.640)	50 (0.131)	177 (0.465)	154 (0.404)	277 (0.364)	485 (0.636)	0.816	0.4309	0.7314
rs11125529	13 (0.033)	111 (0.285)	265 (0.681)	137 (0.176)	641 (0.824)	10 (0.026)	109 (0.286)	262 (0.688)	129 (0.169)	633 (0.831)	0.915	0.5503	0.8180
rs16859140	2 (0.005)	35 (0.090)	352 (0.905)	39 (0.050)	739 (0.950)	1 (0.003)	37 (0.097)	343 (0.900)	39 (0.051)	723 (0.949)	1	0.4046	0.7632
rs12638862	15 (0.085)	138 (0.784)	23 (0.131)	168 (0.216)	610 (0.784)	20 (0.124)	117 (0.727)	24 (0.149)	157 (0.206)	605 (0.794)	0.2918	1.1860	2.0430
rs10936599	1 (0.003)	16 (0.041)	372 (0.956)	18 (0.023)	760 (0.977)	1 (0.003)	19 (0.050)	361 (0.948)	21 (0.028)	741 (0.972)	0.803	0.2782	0.8223
rs1317082	1 (0.003)	16 (0.041)	370 (0.956)	18 (0.023)	756 (0.977)	1 (0.003)	18 (0.047)	362 (0.950)	20 (0.026)	742 (0.974)	0.864	0.6680	0.9573
rs10936601	58 (0.216)	197 (0.735)	13 (0.049)	313 (0.402)	465 (0.598)	72 (0.282)	169 (0.663)	14 (0.055)	313 (0.411)	449 (0.589)	0.7557	0.5445	0.9699
rs1920116	4 (0.010)	81 (0.208)	304 (0.781)	89 (0.114)	689 (0.886)	7 (0.018)	68 (0.178)	306 (0.803)	82 (0.108)	680 (0.892)	0.4164	0.5702	0.9909

A=Minor allele, B=Major allele, L95= Lower confidence interval, U95= Upper confidence interval, Decimals in brackets denote the frequency of each genotype and allele in the association results.

Table 3.4: Allelic association of 28 variants adjusted for covariates sex, age, BMI and geographic region in 770 individuals with allele and genotype frequencies …continued

SNP	Cases					Controls					P-value	L95	U95
	Genotypes			Alleles		Genotypes			Alleles				
	A/A	A/B	B/B	A	B	A/A	A/B	B/B	A	B			
rs10466239	10 (0.026)	75 (0.193)	303 (0.781)	95 (0.122)	681 (0.878)	10 (0.026)	86 (0.226)	285 (0.748)	106 (0.139)	656 (0.861)	0.5293	0.7544	1.0730
rs9420907	61 (0.249)	169 (0.690)	15 (0.061)	291 (0.374)	487 (0.626)	66 (0.264)	170 (0.680)	14 (0.056)	302 (0.396)	460 (0.604)	0.6806	0.6237	1.2100
rs17653722	0 (0.000)	12 (0.031)	377 (0.969)	12 (0.015)	766 (0.985)	0 (0.000)	15 (0.039)	366 (0.961)	15 (0.020)	747 (0.980)	0.5612	0.7877	1.4750
rs9537514	31 (0.157)	146 (0.737)	21 (0.106)	208 (0.267)	570 (0.733)	27 (0.145)	138 (0.742)	21 (0.113)	192 (0.252)	570 (0.748)	0.8038	0.8306	1.2580
rs398652	58 (0.241)	167 (0.693)	16 (0.066)	283 (0.364)	495 (0.636)	55 (0.232)	167 (0.705)	15 (0.063)	277 (0.364)	485 (0.636)	0.9563	0.8226	1.1740
rs4902100	46 (0.204)	162 (0.717)	18 (0.080)	254 (0.326)	524 (0.674)	33 (0.138)	192 (0.800)	15 (0.063)	258 (0.339)	504 (0.661)	0.6266	0.5503	0.8180
rs2535913	5 (0.013)	114 (0.293)	270 (0.694)	124 (0.159)	654 (0.841)	12 (0.031)	94 (0.247)	275 (0.722)	118 (0.155)	644 (0.845)	0.099	0.4046	0.7632
rs74019828	1 (0.003)	31 (0.080)	357 (0.918)	33 (0.042)	745 (0.958)	0 (0.000)	24 (0.063)	357 (0.937)	24 (0.031)	738 (0.969)	0.4028	1.1860	2.0430

A=Minor allele, B=Major allele, L95= Lower confidence interval, U95= Upper confidence interval, Decimals in brackets denote the frequency of each genotype and allele in the association results

Table 3.4: Allelic association of 28 variants adjusted for covariates sex, age, BMI and geographic region in 770 individuals with allele and genotype frequencies …continued

SNP	Cases					Controls					P-value	L95	U95
	Genotypes			Alleles		Genotypes			Alleles				
	A/A	A/B	B/B	A	B	A/A	A/B	B/B	A	B			
rs6028466	36 (0.182)	141 (0.712)	21 (0.106)	213 (0.274)	565 (0.726)	35 (0.173)	148 (0.733)	19 (0.094)	218 (0.286)	544 (0.714)	0.7522	0.5445	0.9699
rs6010620	0 (0.000)	15 (0.039)	374 (0.961)	15 (0.019)	763 (0.981)	0 (0.000)	13 (0.034)	368 (0.966)	13 (0.017)	749 (0.983)	0.848	0.5702	0.9909
rs2300383	9 (0.023)	132 (0.339)	248 (0.638)	150 (0.193)	628 (0.807)	17 (0.106)	119 (0.744)	24 (0.150)	153 (0.201)	607 (0.799)	0.2139	0.9388	1.5350
rs73394838	37 (0.175)	155 (0.735)	19 (0.090)	229 (0.294)	549 (0.706)	25 (0.121)	162 (0.786)	19 (0.092)	212 (0.278)	550 (0.722)	0.2952	0.8306	1.2580
rs2736100	85 (0.310)	177 (0.646)	12 (0.044)	347 (0.446)	431 (0.554)	91 (0.239)	192 (0.504)	98 (0.257)	374 (0.491)	388 (0.509)	0.1076	0.8226	1.1740
rs2098713	8 (0.021)	85 (0.219)	296 (0.761)	101 (0.130)	677 (0.870)	6 (0.016)	89 (0.234)	286 (0.751)	101 (0.133)	661 (0.867)	0.7924	0.5503	0.8180
rs594119	8 (0.021)	127 (0.326)	254 (0.653)	143 (0.184)	635 (0.816)	6 (0.016)	104 (0.273)	271 (0.711)	116 (0.152)	646 (0.848)	0.207	0.4046	0.7632
rs11787341	6 (0.015)	74 (0.190)	309 (0.794)	86 (0.111)	692 (0.889)	3 (0.008)	59 (0.155)	319 (0.837)	65 (0.085)	697 (0.915)	0.2671	1.1860	2.0430

A=Minor allele, B=Major allele, L95= Lower confidence interval, U95= Upper confidence interval, Decimals in brackets denote the frequency of each genotype and allele in the association results.

Table 3.4: Allelic association of 28 variants adjusted for covariates sex, age, BMI and geographic region in 770 individuals with allele and genotype frequencies …continued

SNP	Cases					Controls					P-value	L95	U95
	Genotypes			Alleles		Genotypes			Alleles				
	A/A	A/B	B/B	A	B	A/A	A/B	B/B	A	B			
rs7680468	**0 (0.000)**	**2 (0.005)**	**387 (0.995)**	**2 (0.003)**	**776 (0.997)**	**0 (0.000)**	**4 (0.010)**	**377 (0.990)**	**4 (0.005)**	**758 (0.995)**	**0.4468**	**0.9388**	**1.5350**
rs1975174	47 (0.204)	166 (0.722)	17 (0.074)	260 (0.335)	516 (0.665)	41 (0.185)	164 (0.739)	17 (0.077)	246 (0.325)	512 (0.675)	0.0664	0.6680	0.9573
rs2162440	101 (0.366)	174 (0.630)	1 (0.004)	376 (0.483)	402 (0.517)	88 (0.322)	174 (0.637)	11 (0.040)	350 (0.459)	412 (0.541)	0.629	0.2782	0.8223
Rs11703393	86 (0.310)	178 (0.646)	10 (0.044)	349 (0.446)	431 (0.554)	91 (0.239)	192 (0.504)	98 (0.257)	374 (0.491)	388 (0.509)	0.1076	0.8226	1.1740

A=Minor allele, B=Major allele, L95= Lower confidence interval, U95= Upper confidence interval, Decimals in brackets denote the frequency of each genotype and allele in the association results.

Table 3.5: Logistic regression with 28 SNPs, adjusted for covariates sex, BMI and region in 770 individuals

SNP ID	GENE(S)	CHR	A	B	MAF	P	OR
rs621559	WDR65	1	A	G	0.361	0.891	0.985
rs11125529	ACYP2	2	A	C	0.172	0.729	1.048
rs10936599	*TERC*	3	T	T	0.050	0.599	0.847
rs10936601	*TERC*	3	C	A	0.024	0.717	0.963
rs12638862	*TERC*	3	G	A	0.211	0.642	1.060
rs1317082	*TERC*	3	G	T	0.406	0.727	0.895
rs16859140	TMPRSS7	3	C	G	0.111	0.935	0.981
rs1920116	*TERC*	3	A	C	0.025	0.667	1.072
rs7680468	DKK2, PAPSS1	4	T	G	0.003	0.416	0.492
rs2098713	C5orf42	5	C	T	0.131	0.870	0.975
rs2736100	*TERT*	5	C	A	0.468	0.083	0.839
rs594119	NKAIN2	6	A	G	0.168	0.085	1.278
rs11787341	LOC100128993	8	A	G	0.098	0.102	1.324
rs10466239	FXYD4, RASGEF1A	10	T	C	0.130	0.347	0.871
rs9420907	OBFC1	10	A	C	0.385	0.388	0.915
rs17653722	KRT80	12	T	G	0.017	0.515	0.772
rs9537514	PRR20	13	T	G	0.259	0.508	1.079
rs2535913	DCAF4	14	A	A	0.332	0.793	1.038
rs398652	PELI2	14	A	G	0.157	0.971	1.004
rs4902100	SYT16	14	G	G	0.363	0.633	0.948

OR = odds ratio, A= minor allele, B= major allele, MAF= minor allele frequency, CHR=chromosome, P = adjusted P-value.

Table 3.5: Logistic regression with 28 SNPs, adjusted for covariates sex, BMI and region in 770 individuals…continued

SNP ID	GENE(S)	CHR	A	B	MAF	P	OR
rs74019828	**CSNK2A2**	**16**	**A**	**G**	**0.037**	**0.258**	**1.363**
rs2162440	BRUNOL4	18	G	A	0.471	0.363	1.093
rs1975174	ZNF676	19	G	T	0.329	0.667	1.047
rs6010620	RTEL1	20	A	G	0.018	0.737	1.139
rs6028466	DHX35	20	A	G	0.279	0.597	0.943
rs2300383	ITSN1	21	A	G	0.197	0.666	0.945
rs11703393	PARVB	22	G	A	0.424	0.653	1.048
rs73394838	ASCC2	22	G	A	0.286	0.480	1.084

OR = odds ratio, A= minor allele, B= major allele, MAF= minor allele frequency, CHR=chromosome, P = adjusted P-value

Chapter 4: Discussion

4.1 Chapter outline

TL is hypothesised to be a biomarker for ageing as it decreases at each successive cell division (Mather *et al.,* 2011). A substantial number of scientific studies have associated TL with various age-related diseases, such as T2D, cancer, CVD, and atherosclerosis (Khan *et al.,* 2012; Wang *et al.,* 2016; Zhou *et al.,* 2016; Zhu *et al.,* 2016). The aim of this dissertation was to determine the association of TL with T2D in SSA populations. Two approaches were used to achieve this aim. The first was to establish and validate the qPCR TL measurement assay, developed by Cawthon (Cawthon, 2002). The second approach was to examine the relationship of TL-associated variants with T2D in SSA populations. For the second approach, a MR approach was used to assess whether TL was causal in its relationship with T2D. Although both approaches have their limitations and challenges, this study is the first to attempt to assess TL, telomere-related genetic variants, and T2D in a non-admixed African cohort. Studies conducted in populations of European descent were used as a point of reference for the selection of the telomere measurement assay and the TL-associated variants. This was done as there is a significant lack of pre-existing genetic information pertaining to TL in African populations. The TL measurement assay was not effectively established and validated for this study, due to technical challenges that could not be resolved in the time frame for this project. Therefore, TL was not measured and the association of TL with T2D in this cohort could not be examined.

Genetic variants that had been previously found to be associated with TL biology in Caucasian and Asian populations were used to indirectly assess the relationship of TL with T2D in this cohort. Twenty-eight of the 53 variants found to be associated with TL at a genome-wide significance level in other studies were present in the dataset available for this study. These were used to test their association with T2D in 770 of the 836 participants that were selected for the study. The number of participants was lower in the genetic association study as we did not have genotyping data for all the cases, 33 of the 418 individuals with diabetes and their 33 matched controls were removed from further analysis. Logistic regression analysis

was subsequently carried out using a case-control study design to determine whether the selected SNPs were associated with T2D.

4.2 Characterisation of participant phenotype

This study was nested in a larger study on the genetic and environmental factors that contribute to cardiometabolic diseases (including T2D) in Africans, referred to as AWI-Gen. The participants were from 4 different countries, in 6 geographically different study sites; South Africa (Agincourt, Dikgale and Soweto), Kenya (Nairobi), Burkina-Faso (Nanoro) and Ghana (Navrongo) (Ramsay *et al.,* 2016; Ali *et al.,* 2018). The samples represent groups from South, West, and East of Africa, encompassing both rural and urban communities. A group matched case-control association approach for T2D was used for this study, an approach that is said to increase the power of certain epidemiological investigations (Strunner and Brenner, 2001). With the condition that the matching is accounted for in the analysis, if it was not, the matching would be a confounding factor (Pearce, 2016). The participants were matched according to geographical site, sex, age, and BMI. This matching was done to obtain groups of case and control pairs that would be phenotypically similar, with the exception of the disease state (T2D) (Pearce, 2016). This was done with the intention of obtaining an association that would be as accurate as possible. From the approximately 10 000 AWI-Gen participants aged 40 to 60 years, only 600 had health and phenotype data consistent with a likely diagnosis of T2D (told by a doctor or healthcare worker that they had T2D or with a fasting glucose level >7 mmol/l). These cases were then matched to identify individuals suitable for their respective controls according to the above-mentioned criteria. Initially, a total of 1200 participants were included in this study. Due to QC measure requirements of the qPCR assay, explained in the methods chapter, 364 samples failed to meet QC requirements. The final sample size was then reduced to 418 cases and 418 controls. There was no significant difference between cases and controls regarding age, sex and BMI, which was expected because these parameters were used in the matching process. Nonetheless, the phenotype data (sex, age, and BMI) were used as covariates in the subsequent analyses because it is known that they all contribute to either TL attrition, or diabetes progression.

4.3 Summary of findings

The summary of findings will be discussed separately pertaining to the 2 objectives. Firstly the establishment and validation of the qPCR assay for TL and secondly, the association of telomere-associated genetic variants with T2D, using an approach that mimics MR.

4.3.1 Assay establishment

The qPCR-based TL measurement assay used in this project, is the one that is most often used globally for epidemiological studies. It is cost effective, can be used in a high throughput setting, and is less laborious when compared to the other TL measurement methods. Large sets of samples can be measured in a short period of time, and only small amounts of DNA are required for the qPCR assay (Montpetit *et al.,* 2014). From the conception of the original assay by Cawthon (Cawthon, 2002), many variations have been formulated, however, the original protocol was used even though it has considerable disadvantages. Some of the disadvantages of the assay include large variations in TL measurements across different laboratories, and the fact that the measure of TL is an average across all chromosomes from all leucocytes, and moreover, it is calculated as a relative ratio, so it may not be an accurate representation of an individual's true TL. Lastly, standardised protocols for the use of qPCR for TL assessment are lacking, strict guidelines are deficient, and thus this leads to variation across different laboratories (Lai *et al.,* 2018).

In this study, the qPCR assay was unsuccessful despite countless attempts to perform troubleshooting to eliminate inconsistencies and to get the assay to work. The first problem encountered with the telomere assay was non-reproducibility of the telomere PCR assay reaction. Whereas the single-copy gene PCR assay reaction performed adequately, the telomere PCR assay reaction repeatedly failed to depict amplification of the desired telomeric regions. One reason for this could be primer degradation. In optimal conditions, DNA primers should last a long time, i.e. 5 years (Pinto *et al.,* 2018). Fast degradation of primers might be due to insufficient purification of primers during manufacturing (Pinto *et al.,* 2018). However, new primers were purchased, but the problem persisted. Other factors that may affect primer degradation are; exposure to UV radiation, temperature changes, pH, and salt

contamination (Dean *et al.,* 2001). It is unlikely that the above conditions were a contributing factor in our laboratory, as the primers were resuspended and aliquoted and kept at the recommended temperature of -20°C. Moreover, the variables such as salt contamination and pH would be difficult to change in commercially bought primers optimised for PCR conditions.

With the second batch of telomere PCR primers, there was limited success and amplification was specific to a certain type of sample and reagent combination. To decipher this problem, the assay was performed with the previously purchased primers, and then the current primers, using either the first set, or the new set of PCR reagents. Amplification was selective to the combination of the previously purchased primers and latest reagents (SYBR Green). I concluded that the new set of primers was defective. A last and final set of primers and reagents (SYBR Green, master mix) were purchased, and finally, amplification was achieved within both the SCG and telomere aspects of the assay as depicted in Figures 3.33 and 3.34 within the results section. Nonetheless, amplification was also evident in the no-template controls (NTCs), thus suggesting contamination in the wells (Nybo, 2011), making the results unreliable. Upon sterilisation of pipettes and reaction plates prior to use, and use of the fresh stock of primers and reagents, contamination in NTCs persisted. The reason for the amplification in the NTC wells could have been a result of two reasons, firstly the primer stock solution could be the source of contamination or the amplification could be from primer dimer formation. It is important to note that with the single copy reaction there was no contamination, but the two assays need to be used together to estimate TL, so no data was generated that could be used for the comparison between T2D cases and controls. The amplification in the NTC wells demonstrated that the signals seen in the sample wells were inaccurate or false (Bustin and Nolan, 2004). Due to the issues with primer specificity, intercalating dye specificity, and the ever-persistent contamination in the NTC wells, the establishment and validation of the qPCR assay were unsuccessful. Considered that there may be a SNP in the primer binding regions of the telomere primers, but that was unlikely, as the primers were tested on samples of different ethnicities in the laboratory and the same issue was still prevalent. Moreover, the SNPs are unlikely to be present on all the chromosomes of a single cell, as the primers would anneal on all telomere

binding sites of all 46 chromosomes, so a SNP in 1 or 2 would not affect the binding and amplification of the telomere primers to a significant level.

The tel1 and tel2 telomere primers used in this qPCR assay have a natural inclination to form primer dimers (Hsieh *et al.,* 2016). This occurs due to the 3' three base overlap found in the primers, regardless of the mismatch incorporated at the fifth base of the complementary primers (Cawthon, 2002). In theory, this mismatch should prevent primer dimer formation (Cawthon*,* 2008; Jodczyk, 2011). Telomeres being repetitive nucleotides in nature are thus very difficult to accurately measure (Massey *et al.,* 2018). The assay itself relies on many steps in the pre-DNA extraction processing, DNA extraction, and post-extraction processing (Lin *et al.,* 2019). Factors like the source of sample (blood or saliva), storage of sample, DNA extraction method, DNA storage, PCR conditions, and method for data analysis, all lead to variation in TL when using the qPCR method (Lin *et al.,* 1997). There is currently no reliable and robust method available that will ensure uniformity in results considering the above aspects. Thus, guidelines need to be developed and validated for accurate average TL assessment and implemented globally.

In this project the primer dimer formation in the NTC wells could occur with the same conditions as in the sample containing wells, which is expected in the sample wells but not in the NTC wells. This implies that amplification signals observed in the sample wells were also due to primer dimer formation and thus were not true telomere amplification signals. However, the use of hot-start Taq polymerase is designed to prevent the formation of primer dimers (Lebeclev *et al.,* 2008). Hot-start Taq polymerases are specialised DNA polymerase enzymes that have an aptamer-based inhibitor incorporated into them (Lin and Jayasena, 1997). This inhibitor prevents the activation of the enzyme below a temperature of 45°C, above this temperature, the inhibitor is released, and the enzyme performs optimally (Lebeclev *et al.,* 2008). This ensures optimal amplification of the desired PCR product. Nonetheless, even with the hot-start, PCR primer dimer formation was evident in this experiment. This might possibly suggest that the hot-start Taq polymerase enzyme was active at lower temperatures, thus self-annealing of primers occurred, preventing amplification of the true telomeric product.

Though many studies have measured TL successfully using qPCR, only one study has reported failure to measure it accurately (Jodczyk, 2011). However, discussions with researchers in several laboratories suggest that there is frequent failure to get the assay to work and the literature is biased toward the groups who managed to get it to work. Since telomere measurement using qPCR has been reported to be irreproducible and inconsistent across laboratories (Jodcyzk *et al.,* 2011; Lin *et al.,* 2019), it is perhaps not unexpected that it proved to be variable, poorly reproducible, and poorly validated within my project.

For all the available TL measurement methods, as discussed in chapter one, there is considerable methodological variation. In my study, I chose to use the qPCR method as it was the only method suitable for the number samples used in this study. It required low amounts of DNA and most importantly was cost effective and high throughput. The lack of standardisation and transparency in publications makes it difficult to put control measures in place. Furthermore, the way in which samples are handled in different laboratories may contribute to variation seen with the results (Aubert *et al*., 2012). At the present time and to my knowledge, there is no method that accurately, easily, and rapidly measures TL.

4.3.2 Genetic associations with TL

The third objective of the study was to ascertain whether there is an association between genetic variants known to be associated with TL variation and T2D in an African population. This sub-study started with a literature review to identify genetic variants associated with TL, but data were only available from non-African populations. A MR study approach was used to attempt to determine causality of TL for T2D by using a set of SNPs known to be associated with TL. Where T2D was the outcome of interest, TL attrition was the risk factor or exposure variable, the telomere genetic variants were the instrumental variables and lastly; age, sex, and BMI were potential confounders. MR studies are genetic epidemiological studies that scientists have recently developed to disentangle the cause and effect of diseases (Roberts, 2018). The instrumental variable method used in MR studies was adapted from methodologies used in sociology and economics studies (Burgess *et al.,* 2017).

The genetic variants used for the association were sourced from previous studies that had conducted GWAS and reported their genome-wide significant SNPs with TL (Table 3.3). Most of these studies were conducted in European and Asian populations, with little representation from African-American populations and, to date, there is no GWAS for TL that has been conducted in non-admixed African populations (Zeiger *et al.,* 2018). Fifty-three genome-wide significant SNPs were found in the literature and only 28 of these were genotyped. These 28 variants were used as genetic proxies for TL. Previously, 2 of the 28 variants were specifically associated with T2D in an Asian population, and they are found in the *CSNK2A2* and *C5orf42* genes (Saxena *et al.,* 2014). This may partly explain why these two associations were not observed in the SSA populations. In addition to being related to TL, the remaining 26 variants were implicated in various cancers and NCDs as shown in Table 3.3.

As previously mentioned no GWAS for TL has been performed in a non-admixed African population, so the use of variants found in other populations would limit discovery in African populations due to the differences in population genome architecture. To elaborate on the differences between populations, a study conducted in East Africa found that individuals of European descent tended to have shorter telomeres when compared to non-admixed East Africans, suggesting that there are genetic differences with regards to telomere variation between populations (Hansen *et al.,* 2016). Moreover, a GWAS conducted in African-Americans found 6 novel variants associated with TL in adolescents, which were not replicated in any of the association studies in European and Asian populations (Zeiger *et al.,* 2018).

4.3.3 Association of TL variants

Following QC, 385 pairs of T2D cases and controls were included in the genetic association analysis (total n=770). Logistic regression analysis was performed with the 28 SNPs that passed QC. The association was performed with age, sex, and BMI as covariates. None of the 28 variants was significantly associated with T2D before and after adjusting for the above mentioned covariates. This would suggest that TL, as assessed through the telomere-related SNPs, is unlikely to be associated with T2D in these SSA populations. The flaw in this reasoning is, however, that we have not done an independent association between TL and genetic variants in the

African populations we are studying and the European proxies may not apply in Africans.

Bearing in mind that our sample size was relatively small and the variants tested were also small in number and not validated for African populations, it is not surprising that no associations were found between the telomere-associated variants and T2D. Moreover, for the genetic association analysis, a post hoc power calculation was performed and it proved that this study was indeed under powered to detect associations with low to moderate effects. Thus indicating that we cannot strongly conclude with certainty that all the variants were not associated with T2D. Since statistical power was sub-optimal and the sample size relatively small one cannot confidently accept the lack of association found in our study. Considering that the association study was performed in African populations, using SNPs reported in Caucasian, Asian, and African American populations (Levy *et al*., 2010; Mangino *et al*., 2012; Saxena *et al*., 2014; Zeiger *et al*., 2018), it is likely to result in no associations being found, in the case where the SNPs are specific to the specified populations.

Population heterogeneity has to be considered, as it has been shown that non-admixed African populations have shorter linkage disequilibrium (LD) blocks (Wall and Pritchard, 2003) with higher genetic diversity (Campbell and Tishkoff, 2008). With this being acknowledged, the participants from this study are from 6 different geographical regions, meaning that the genetic differences in our cases might hinder the ability to detect valuable associations that could be present in one African population, but not in others. We were underpowered to look at the differences between regions. However, the genotyping was performed on the specifically designed H3Africa SNP array, which has better representation of common African variation and should be able to detect associations, as it incorporates elements of higher genetic diversity and more SNPs, to detect the shorter LD blocks seen in African populations. Further adequately powered studies are required to examine associations between TL and T2D in African populations.

Inconsistent associations have been reported with regards to TL and T2D by some epidemiological studies (Zhang *et al.,* 2015). In this study, in order to minimise

biases, we performed a MR approach where TL-associated genetic variants served as a proxy for the actual TL. No association was detected.

4.3.4 Telomere association with T2D

Previous investigations of the association of TL and T2D have provided conflicting results. Studies performed in European and Asian populations could not be replicated in African American populations (Zeiger *et al.,* 2018). However, novel variants were discovered in such studies. A study conducted in an Indian population discovered a novel variant in the *CSNK2A2* gene associated with shorter TL ($P=4.52\times10^{-8}$), but this association was not replicated in Asian and European populations, suggesting that the variant could be population specific (Saxena *et al.,* 2014). Considering that the cohort of interest in this study is African, it might be possible that none of the discovered variants are specific to the African population. When investigating the relationship between TL and T2D in patients with ischaemic heat disease (IHD), in an Indian population, it was found that TL significantly ($p = 0.00019$) shortened in a period of 2 years in IHD participants who had T2D compared to those who did not have T2D (Piplani *et al.,* 2018). This association showed that there is a role played by TL attrition in patients with T2D. However a longitudinal study of 6 years in an Asian population, which observed the change in TL in diabetic patients, showed conflicting results between diabetic patients on insulin treatment and those on other forms of treatment. It was found that 30% of the participants had increased TLs and 70% showed an attrition in TL. The majority of the patients who were on insulin treatment had decreased TL when compared to patients on other glucose lowering drugs (Zeng *et al.,* 2018). This implies that other glucose lowering drugs have the potential to initiate mechanisms that increase TL that insulin does not, hence depending on the type of treatment in T2D patients, TL is likely to increase or decrease. Thus drug interactions should be an avenue of interest when investigating TL and T2D, as it may hinder the detection of true associations. In this particular study, the type of medication and duration thereof, was not captured, and based on the study by Zeng *et al*. (2018) medication does have an effect on TL and should be considered as a confounder in future studies. Moreover many other co-existing diseases like HIV infection could also impact

telomere attrition or elongation, even more so, when individuals are on anti-retro viral medication.

A longitudinal study that investigated TL in T2D patients and its associated biochemical and clinical variables, found that there was no correlation between TL and T2D in the cases and controls, but, there was a significant correlation between TL and fasting plasma glucose levels (Rosa *et al.,* 2018). Glucose control was inversely associated with TL, meaning that in participants with overall higher plasma glucose levels, TL was significantly shorter. This supports the assertion that telomeres are affected by hyperglycaemia, which increases inflammation and oxidative stress in T2D patients.

4.4 Strengths and limitations

This study is the first to look at the association between TL and T2D in a non-admixed African population. It has added to the technological and methodological insights that need to be considered when investigating TL. The method with which one measures TL is crucial in performing a successful study of TL. This should be extensively probed before initiating a study on TL. Furthermore, the phenotypic data that was available for the study was limited, and when one is investigating TL it would be useful to have additional biomarkers that are relevant to TL. For example, indicators of inflammation and oxidative stress (e.g. CRP) were not available and therefore their effects on TL in this population could not be investigated.

Regarding the genetic study, a GWAS or genetic association for TL had not yet been conducted in a non-admixed African population and genetic variants as proxies for TL that are specific to African populations were not available. Therefore, generalisation of the previously discovered SNPs for investigation in an African population is a potential pitfall of this study, when trying to find the association between TL and T2D. Moreover the cohort that was used in this study was relatively small in size for a genetic association analysis, and if the effect of TL on T2D is small, this could be the reason why no association was found with any of the TL-related variants.

4.5 Summary and conclusion

The association of TL with T2D could not be fully explored in this study due to the problems in getting the TL assay to work reliably. Given more time and a substantial budget, it would be worthwhile exploring the association of TL with T2D is an African cohort. A good starting point would be exploring other methods to measure TL and visiting laboratories where the assays are performing reliably. The genetic variants tested did not show any association with T2D and therefore no inference could be made regarding this negative result. It is possible that the SNPs were poor predictors of TL in African populations and this then violates one of the pillars of a MR approach. For future studies it would be important to first perform an association study for TL in African populations to validate the MR instrument. Only then could an appropriate set of SNPs be used to examine causality of TL in the development of T2D. The approach used in this study was convenient, given the available data. The small sample size, and small number of SNPs used in the investigation affected the power to detect association signals. Moreover, the SNPs were chosen based on data from genetically different populations. Further exploratory approaches need to be used in African populations, in order to discover novel African genetic contributions to TL variation and to enable studies that relate TL to T2D.

5 References

- Adaikalakoteswari, A., Balasubramanyam, M., Ravikumar R. *et al*., 2007. Association of telomere shortening with impaired glucose tolerance and diabetic macroangiopathy. *Atherosclerosis*, 195(1), pp. 83-89.

- Ali, S. A., Soo C., Agongo G. *et al*., 2018. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. *Global Health Action*, 16 11, 11(sup2), pp. 1507133.

- Allman, P. H., Aban, I. B., Tiwari, H. K. *et al*., 2018. An introduction to Mendelian randomization with applications in neurology. *Multiple Sclerosis and Related Disorders,* 8, Volume 24, pp. 72-78.

- Allsopp, R. C., Chang, E., Kashefi-Aazam, M. *et al*., 1995. Telomere Shortening Is Associated with Cell Division in Vitro and in Vivo. *Experimental Cell Research*, 1 9, 220(1), pp. 194-200.

- Alter, B.P., Baerlocher, G.M., Savage, S.A. *et al*., 2007. Very short telomere length by flow fluorescence *in situ* hybridization identifies patients with dyskeratosis congenita. *Blood*, 111(5), pp. 1439-1447.

- Andrew, T., Aviv, A., Falchi, M. *et al*., 2006. Mapping genetic loci that determine leukocyte telomere length in a large sample of unselected female sibling pairs*. American journal of human genetics*, 78(3), pp.480–486.

- Atkinson, M. A. 2012. The Pathogenesis and Natural History of Type 1 Diabetes*. Cold Spring Harbor Perspectives in Medicine*. 2(11), pp. a007641.

- Aubert, G., Hills, M. and Lansdorp, P. M., 2011. Telomere length measurement-Caveats and a critical assessment of the available technologies and tools. *Mutation Research,* Volume 730, pp. 59-67.

- Bahendeka, S. K., 2017. Diabetes in sub-Saharan Africa: let us not forget type 1. *The Lancet, Diabetes & Endocrinology*. 5(8) pp. 575-577.

- Barrett, L.B., and Richardson, D.S., 2011. Sex differences in telomeres and lifespan. *Aging Cell*, 10(6), pp. 913-921.

- Benetos, A., Okuda, K., Lajemi, M. *et al*., 2001. Telomere length as an indicator of biological aging: the gender effect and relation with pulse pressure and pulse wave velocity*. Hypertension*. 2, 37(2 Pt 2), pp. 381-385.

- Blackburn, E. H., Epel, E. S. and Lin, J., 2015. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science,* 350(6265), pp. 1193-1198.

- Bonnefond, A., Froguel, P. and Vaxillaire, M., 2010. The emerging genetics of type 2 diabetes. Trends in Molecular Medicine. 16(9), pp. 407-416.

- Brouilette, S., Singh, R.K., Thompson, J.R. *et al*., 2003. White Cell Telomere Length and Risk of Premature Myocardial Infarction*. Arteriosclerosis, Thrombosis, and Vascular Biology*, 5, 23(5), pp. 842-846.

- Burgess, S., Small, D. S. and Thompson, S. G., 2017. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research,* 17 10, 26(5), pp. 2333-2355.

- Bustin, S. A., and Nolan, T. 2004. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *Journal of biomolecular techniques: JBT*. 15(3), pp. 155–166.

- Butt, H. Z., Attura, G., London, N..J *et al*., 2010. Telomere length dynamics in vascular disease: A review. *European Journal of Vascular and Endovascular Surgery*, 40(1), pp. 17-26.

- Calado, R. T., and Young, N. S. 2009. Telomere diseases*. The New England journal of medicine*, 361(24),pp. 2353–2365.

- Calado, R.T., Regal, J.A., Hills, M. *et al*., 2009. Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia*. Proceedings of the National Academy of Sciences*. 106 (4) pp.1187-1192.

- Campbell, M. C. and Tishkoff, S. A., 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.. *Annual review of genomics and human genetics,* Volume 9, pp. 403-33.

- Cawthon, R. M., 2002. Telomere measurement by quantitative PCR.. *Nucleic acids research,* 30(10), pp. e47.

- Cawthon, R. M., 2008. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Research,* 37(3), pp. 1-7.

- Chatterjee, S., Khunti, K. and Davies, M. J., 2017. Type 2 diabetes. *The Lancet,* 6, 389(10085), pp. 2239-2251.

- Codd, V., Mangino, M., van der Harst, P. *et al*., 2010. Variants near *TERC* are associated with mean telomere length*. Nature genetics*, 42(3), pp. 197-199.

- Codd, V., Nelson, C. P., *ALB*recht, E. *et al*., 2013. Identification of seven loci affecting mean telomere length and their association with disease*. Nature genetics*, 45(4), pp.422–427.

- Cole, L., Kramer, P. R., Cole, L. *et al*., 2016. Diabetes. *Human Physiology, Biochemistry and Basic Medicine,* 1 1.pp. 211-213.

- de Pauw, E.S., Roelofs, H., Zwinderman, A. *et al*., 2005. Studying the biological and technical sources of variation in telomere length of individual chromosomes. *Cytometry*, 65(1), pp. 35-39.

- Dean, M. D., William, J. and Ballard, O., 2001. Factors affecting mitochondrial DNA quality from museum preserved Drosophila simulans*. Entomologia Experimentalis et Applicata*, 98, pp. 279–283.

- Deelen, J., Beekman, M., Codd, V. *et al*., 2014. Leukocyte telomere length associates with prospective mortality independent of immune-related parameters and known genetic markers*. International journal of epidemiology*, 43(3), pp. 878–886.

- DeFronzo, R. A., and Tripathy, D. 2009. Skeletal muscle insulin resistance is the primary defect in type 2 diabetes. *Diabetes care*, 32 Suppl 2(Suppl 2), pp. S157–S163.

- Delgado, D. A., Zhang, C., Chen, L. S. *et al*. 2018. Genome-wide association study of telomere length among South Asians identifies a second RTEL1 association signal*. Journal of medical genetics*, 55(1), pp. 64–71.

- Diez Roux, A. V., Ranjit, N., Jenny, N. S. *et al*., 2009. Race/ethnicity and telomere length in the Multi-Ethnic Study of Atherosclerosis. *Aging cell*, 8(3), pp. 251–257.

- Elmarakby, A. A., Abdelsayed, R., Yao Liu, J. *et al*., 2010. Inflammatory cytokines as predictive markers for early detection and progression of diabetic nephropathy. *The EPMA journal*, 1(1), pp. 117–129.

- Erusalimsky, J. D. and Skene, C., 2009. Mechanisms of endothelial senescence. *Experimental Physiology,* 1 3, 94(3), pp. 299-304.

- Factor-Litvak, P., Susser, E., Kezios, K. *et al*., 2016. Leukocyte Telomere Length in Newborns: Implications for the Role of Telomeres in Human Disease. *Pediatrics*, 137(4), pp. e20153927.

- Garcı´a-Calzo´n, S., Gea, A., Razquin, C. *et al*., 2014. Longitudinal association of telomere length and obesity indices in an intervention study with a Mediterranean diet: the PREDIMED-NAVARRA trial*, International Journal of Obesity*, 38, pp. 177–182.

- Gorenjak, V., Akbar, S., Stathopoulou, M. G. *et al*., 2018. The future of telomere length in personalized medicine.. *Frontiers in bioscience (Landmark edition),* 1 3, Volume 23, pp. 1628-1654.

- Greider, C.W. and Blackburn, E.H. 1985. Identification of a Specific Telomere Terminal Transferase Activity in *Tetrahymena* Extracts. *Cell*, 43, 405-413.

- Greider, C.W. and Blackburn, E.H. 1989. 'A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis' *Nature*, 337(6205), pp. 331-337.
- Gu, J., Chen, M., Shete, S. *et al*., 2011. A Genome-wide Association Study Identifies a Locus on Chromosome 14q21 as a Predictor of Leukocyte Telomere Length and as a Marker of Susceptibility for Bladder Cancer. *Cancer Prev Res (Phila)*, 4(4), pp. 514-521.
- Hansen, M. E., Hunt, S. C., Stone, R. C. *et al*., 2016. Shorter telomere length in Europeans than in Africans due to polygenetic adaptation*. Human molecular genetics*, 25(11), pp. 2324–2330.
- Hara, K., Shojima, N., Hosoe, J. *et al*., 2014. Genetic architecture of type 2 diabetes. *Biochemical and Biophysical Research Communications,* 452(2), pp. 213-220.
- Harris, S. E., Martin-Ruiz, C., von Zglinicki, T. *et al*., 2012. Telomere length and aging biomarkers in 70-year-olds: The Lothian Birth Cohort 1936. *Neurobiology of Aging*, 33(7), pp. 1486.e3-1486.e8.
- Hayflick, L., 1998. How and why we age.. *Experimental gerontology,* 33(7-8), pp. 639-653.
- Honig, L. S. *et al*., 2006. Shorter telomeres are associated with mortality in those with APOE ??4 and dementia. *Annals of Neurology,* 60(2), pp. 181-187.
- Honig, L. S., Kang, M. S., Cheng, R. *et al*., 2015. Heritability of telomere length in a study of long-lived families. *Neurobiology of aging*, 36(10), pp. 2785–2790.
- Honig, L. S., Schurpf, N., Lee, J.H. *et al*., 2006. Shorter telomeres are associated with mortality in those with APOE epsilon4 and dementia. *Annals of Neurology*, 60(2), pp. 181-187.
- Hsieh, A., Saberi, S., Ajaykumar, A. *et al*., 2016. Optimization of a Relative Telomere Length Assay by Monochromatic Multiplex Real-Time Quantitative

PCR on the LightCycler 480: Sources of Variability and Quality Control Considerations. *The Journal of molecular diagnostics* : JMD, 18(3), pp. 425–437.

- Hunt, S. C., Chen, W., Gardner, J. P. *et al*., 2008. Leukocyte telomeres are longer in African Americans than in whites: the National Heart, Lung, and Blood Institute Family Heart Study and the Bogalusa Heart Study. *Aging cell,* 7(4), pp. 451–458.
- IDF Diabetes Atlas, Fourth Edition 2015, www.diabetesatlas.org.
- Jaacks, L. M., Siegel, K. R., Gujral, U. P. *et al*., 2016. Type 2 diabetes: A 21st century epidemic. *Best Practice and Research Clinical Endocrinology and Metabolism,* 30(3), pp. 331-343.
- Jodczyk, S., 2011. Establishment and validation of a telomere length assay. Issue November, pp. 1-100.
- Joubert, J. and Bradshaw, D., 2005. Population Ageing and Health Challenges in South Africa. *Chronic Diseases of Lifestyle in South Africa since 1995-2005,* pp. 204-219.
- Kasuga M. 2006. Insulin resistance and pancreatic beta cell failure. *The Journal of clinical investigation*, 116(7), pp. 1756–1760.
- Khan, S., Chuturgoon, A. A. and Naidoo, D. P., 2012. Telomeres and atherosclerosis.. *Cardiovascular journal of Africa,* 11, 23(10), pp. 563-71.
- Khodabandehloo, H., Gorgani-Firuzjaee, S., Panahi, G. *et al*., 2016. Molecular and cellular mechanisms linking inflammation to insulin resistance and β-cell dysfunction. *Translational Research,* Volume 167, pp. 228-256.
- Kimura, M., Stone, R.C, Hunt, S.C. *et al*.,2010. Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. Nature Protocols, ;5(9), pp. 1596-607.

- Krasnienkov, D.S., Khalangot, M.D., Kravchenko, V.I.*et al*., 2018. Hyperglycemia attenuates the association between telomere length and age in Ukrainian population. *Experimental Gerontology*, 9(110), pp. 247-252.

- Kushitor, M. K. and Boatemaa, S., 2018. The double burden of disease and the challenge of health access: Evidence from Access, Bottlenecks, Cost and Equity facility survey in Ghana.. *PloS one,* 13(3), pp. e0194677.

- Lai, T. P., Wright, W. E., and Shay, J. W. 2018. Comparison of telomere length measurement methods. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 373(1741), pp. 20160451.

- Laimer, M., Melmer, A., Lamina, C. *et al*., 2015. Telomere length increase after weight loss induced by bariatric surgery : results from a 10 years prospective study. *International Journal of Obesity*, 40(5), pp. 773-778.

- Lebedev, A. V., Paul, N., Yee, J. *et al*., 2008. Hot start PCR with heat-activatable primers: a novel approach for improved PCR performance*. Nucleic acids research*, 36(20), pp.e131.

- Lee, J. H., Cheng, R., Honig, L. S. *et al*., 2014. Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study. *Frontiers in genetics*, 4(310).

- Levy, D., Neuhausen, S. L., Hunt, S. C. *et al*., 2010. Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20), pp. 9293–9298.

- Lin, J., Smith, D. L., Esteves, K. *et al*., 2019. Telomere length measurement by qPCR – Summary of critical factors and recommendations for assay design. *Psychoneuroendocrinology,* 1 1, Volume 99, pp. 271-278.

- Lin, Y. and Jayasena, S. D., 1997. Inhibition of multiple thermostable DNA polymerases by a heterodimeric aptamer. *Journal of Molecular Biology,* 8 8, 271(1), pp. 100-111.

- Liu, Y., Cao, L., Li, Z *et al*., 2014. A genome-wide association study identifies a locus on *TERT* for mean telomere length in Han Chinese. *PloS one*, 9(1), pp. e85043..

- Lunenfeld, B. and Stratton, P., 2013. The clinical consequences of an ageing world and preventive strategies. *Best Practice and Research Clinical Obstetrics and Gynaecology,* Volume 27, pp. 643-659.

- Mangino, M., Christiansen, L., Stone, R. *et al*., 2015. DCAF4, a novel gene associated with leucocyte telomere length. *Journal of medical genetics*, 52(3), pp. 157–162.

- Massey, D. S., Wagner, B., Donnelly, L. *et al*., 2018. Neighborhood Disadvantage and Telomere Length: Results from the Fragile Families Study. *The Russell Sage Foundation journal of the social sciences* : *RSF*, 4(4), pp. 28–42.

- Mather, K. A., Jorm, A. F., Parslow, R. A. *et al*., 2011. Is telomere length a biomarker of aging? A review. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences,* 66 A(2), pp. 202-213.

- McCarthy, M. I., and Zeggini, E. 2009. Genome-wide association studies in type 2 diabetes. *Current diabetes reports*, 9(2), pp. 164–171.

- Miller, S. A., Dykes, D. D. and Polesky, H. F., 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research,* 16(3), pp. 1215.

- Minamino, T., Miyauchi H., Yoshida, T. *et al*., 2002. Endothelial cell senescence in human atherosclerosis: role of telomere in endothelial dysfunction. *Circulation*, 105(13), pp. 1541-4.

- Montpetit, A. J., Alhareeri, A. A., Montpetit, M. *et al*., 2014. Telomere length: a review of methods for measurement. *Nursing research*, 63(4), pp. 289–299.

- Mulder, N., Abimiku, A., Adebamowo, S. N. *et al*., 2018. H3Africa: current perspectives. *Pharmacogenomics and personalized medicine*, 11, pp. 59–66.

- Mundstock, E., Zatti, H., Louzada, F.M. *et al*., 2015. Effects of physical activity in telomere length: Systematic review and meta-analysis*. Aging Research Reviews*, 22, pp. 72-80.

- Muoio, D. M. and Newgard, C. B., 2008. Molecular and metabolic mechanisms of insulin resistance and β-cell failure in type 2 diabetes. *Nature Reviews Molecular Cell Biology,* 1 3, 9(3), pp. 193-205.

- Needham, B. L., Adler, N., Gregorich, S. *et al*., 2013. Socioeconomic status, health behavior, and leukocyte telomere length in the National Health and Nutrition Examination Survey, 1999-2002. *Social science & medicine (1982),* 85,pp. 1–8.

- Njajou, O. T., Hsueh, W.C., Blackburn, E.H. *et al*., 2009. Association between telomere length, specific causes of death, and years of healthy life in health, aging, and body composition, a population-based cohort study.. *The journals of gerontology. Series A, Biological sciences and medical sciences,* 64(8), pp. 860-4.

- Nussey, D. H., Baird, D., Barrett, E. *et al*., 2014. Measuring telomere length and telomere dynamics in evolutionary biology and ecology. *Methods in ecology and evolution*, 5(4), pp. 299–310.

- Nybo, K., 2009. Real-Time qPCR/qRT-PCR Methods: Amplification in Template-Free Controls. *BioTechniques,* 2, 46(2), pp. 91-93.

- O'Callaghan, N. J. and Fenech, M., 2011. A quantitative PCR method for measuring absolute telomere length.. *Biological procedures online,* 13(1), pp. 3.

- Olokoba, A. B., Obateru, O. A. and Olokoba, L. B., 2012. Type 2 Diabetes Mellitus: A Review of Current Trends, *Oman medical journal*, 2012 vol: 27 (4) pp. 269-73.

- Paschou, S. A., Papadopoulou-Marketou, N., Chrousos, G. P *et al*., 2018. On type 1 diabetes mellitus pathogenesis. *Endocrine Connections* 7(1), pp. R38–R46.

- Pearce, N., 2016. Analysis of matched case-control studies.. *BMJ (Clinical research ed.),* 25(352), pp. i969.

- Pfaffl M. W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic acids research*, 29(9), pp. e45.

- Pillay-van Wyk, V. Msemburi, W.2., Laubscher, R. *et al*., 2016. Mortality trends and differentials in South Africa from 1997 to 2012: second National Burden of Disease Study. *The Lancet Global Health,* 4(9), pp. e642-e653.

- Pinto, A., Chen, S. X. and Zhang, D. Y., 2018. Simultaneous and stoichiometric purification of hundreds of oligonucleotides*. Nature Communications*, 9(1), pp. 2467.

- Piplani, S., Alemao, N.N., Prabhu, M. *et al*., 2018. Correlation of the telomere length with type 2 diabetes mellitus in patients with ischemic heart disease. *Indian Heart Journal*, 12(70), pp. S173-S176.

- Poitout, V., Amyot, J., Semache, M. *et al*., 2010. Glucolipotoxicity of the pancreatic beta cell*. Biochimica et biophysica acta (BBA) - Molecular and Cell Biology of Lipids*, 1801(3), pp. 289–298.

- Pooley, K. A., Bojesen, S. E., Weischer, M. *et al* 2013. A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. *Human Molecular Genetics*, 22(24), pp. 5056-5064.

- Prescott, J., Kraft, P., Chasman, D. I. *et al*., 2011. Genome-Wide Association Study of Relative Telomere Length. *PLoS ONE*, 6(5), pp. 19635.

- Ramsay, M. and Sankoh, O., 2016. African partnerships through the H3Africa Consortium bring a genomic dimension to longitudinal population studies on the continent. *International Journal of Epidemiology,* pp. 305-308.

- Ramsay, M., Crowther, N., Tambo, E. *et al*., 2016. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and

environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Global health, epidemiology and genomics*, 1, pp. e20.

- Roberts, R., 2018. Mendelian Randomization Studies Promise to Shorten the Journey to FDA Approval.. *JACC. Basic to translational science,* 10, 3(5), pp. 690-703.

- Rosa, E. C. C. C., Dos Santos, R.C.C., Fernandes, L.F.A. *et al*., 2018. Leukocyte telomere length correlates with glucose control in adults with recently diagnosed type 2 diabetes*. Diabetes Research and Clinical Practice*, 1(135), pp. 30-36.

- Salpea, K. D. and Humphries, S. E., 2010. Telomere length in atherosclerosis and diabetes. *Atherosclerosis,* 209(1), pp. 35-38.

- Salpea, K. D., Nicaud, V., Tiret, L. *et al*., 2008. The association of telomere length with paternal history of premature myocardial infarction in the European Atherosclerosis Research Study II. *Journal of Molecular Medicine*, 86(7), pp. 815-824.

- Salpea, K. D., Talmud, P. J., Cooper, J. A. *et al*., 2010. Association of telomere length with type 2 diabetes, oxidative stress and UCP2 gene variation. *Atherosclerosis*, 209(1), pp. 42-50.

- Samani, N. J., Boulty, R., Butler, R. *et al*., 2001. Telomere shortening in atherosclerosis. *The Lancet*, 11 8, 358(9280), pp. 472-473.

- Sampson, M. J. and Hughes, D. A., 2006. Chromosomal telomere attrition as a mechanism for the increased risk of epithelial cancers and senescent phenotypes in type 2 diabetes. *Diabetologia,* 49(8), pp. 1726-1731.

- Sanders, J. L., and Newman, A. B. 2013. Telomere length in epidemiology: a biomarker of aging, age-related disease, both, or neither?. Epidemiologic reviews, 35(1), pp. 112–131.

- Saxena, R., Bjonnes, A., Prescott, J. *et al*., 2014. Genome-Wide Association Study Identifies Variants in Casein Kinase II (CSNK2A2) to be Associated

with Leukocyte Telomere Length in a Punjabi Sikh Diabetic Cohort. *Circulation: Cardiovascular Genetics*, 7(3), pp. 287-295.

- Seeker, L. A., Holland, R., Underwood, S. *et al*., 2016. Method specific calibration corrects for DNA extraction method effects on relative telomere length measurements by quantitative PCR. *PLoS ONE*, 11(10), pp. e0164046.

- Shaffer, J. A., Epel, E., Kang, M. S. *et al*., 2012. Depressive symptoms are not associated with leukocyte telomere length: findings from the Nova Scotia Health Survey (NSHS95), a population-based study. *PloS one*, 7(10), pp. e48318.

- Silva, E.F.F., Ferreira, C.M.M., and Pinho, L. 2017. Risk factors and complications in type 2 diabetes outpatients*. Revista da Associação Médica Brasileira*, 63(7), pp. 621-627.

- Skyler, J. S. *et al*., 2017. Differentiation of Diabetes by Pathophysiology, Natural History, and Prognosis. *immunology, metabolism, endocrinology,* Volume 66, pp. 10-12.

- Skyler, J. S., Bakris, G. L., Bonifacio, E. *et al*., 2017. Differentiation of Diabetes by Pathophysiology, Natural History, and Prognosis*. Diabetes*, 66(2), pp. 241–255.

- Slagboom, P. E., Droog, S., and Boomsma, D. I. 1994. Genetic determination of telomere size in humans: a twin study of three age groups. American journal of human genetics, 55(5), pp. 876–882.

- Smith, L., Luchini, C., Demurtas, J. *et al*., 2006. Telomere length and health outcomes: An umbrella review of systematic reviews and meta-analyses of observational studies*. Ageing Research Reviews*, 1(51), pp. 1-10.

- Sørensen, K., Van den Broucke, S., Fullam, J. *et al*., 2012. Health literacy and public health: a systematic review and integration of definitions and models. *BMC public health*, 1(12), pp. 80.

- Tamura, Y., Izumiyama-Shimomura, N., Kimbara, Y. *et al*., 2016a. Telomere attrition in beta and alpha cells with age. *Age (Dordrecht, Netherlands),* 38(3), pp. 61.

- Tamura, Y., Takubo, K., Aida, J. *et al*., 2016. Telomere attrition and diabetes mellitus. *Geriatrics and Gerontology international*.,16(1), pp. 66-74.

- Tangvarasittichai, S., 2015. Oxidative stress, insulin resistance, dyslipidemia and type 2 diabetes mellitus. *World Journal of Diabetes,* 6(3), pp. 456.

- Tedone, E., Arosio, B., Gussago, C. *et al*., 2014. Leukocyte telomere length and prevalence of age-related diseases in semisupercentenarians, centenarians and centenarians' offspring. *Experimental Gerontology*, 58, pp. 90-105.

- Tekola-Ayele, F., Adeyemo, A. a. and Rotimi, C. N., 2014. Genetic epidemiology of type 2 diabetes and cardiovascular diseases in Africa.. *Progress in cardiovascular diseases,* 56(3), pp. 251-60.

- Tentolouris, N., Nzietchueng. R., Cattan, V. *et al*., 2007. White blood cells telomere length is shorter in males with type 2 diabetes and micro*ALB*uminuria. *Diabetes Care*, 30(11), pp. 2909-2915.

- Terry, K. L., Tworoger, S.S., Vitonis, A.F. *et al*., 2012. Telomere length and genetic variation in telomere maintenance genes in relation to ovarian cancer risk. *Cancer Epidemiology Biomarkers and Prevention*, 21(3), pp. 504-512.

- Thorogood, M., Goudge, J., Bertram, M. *et al*., 2014. The Nkateko health service trial to improve hypertension management in rural South Africa: study protocol for a randomised controlled trial.. *Trials,* 15(1), p. 435.

- Tzanetakou, I. P., Katsilambros, N.L., Benetos, A. *et al*., 2012. " Is obesity linked to aging?". Adipose tissue and the role of telomeres. *Ageing Research Reviews,* 11(2), pp. 220-229.

- Vaiserman, A. and Lushchak, O., 2017. Implementation of longevity-promoting supplements and medications in public health practice:

achievements, challenges and future perspectives. *J Transl Med,* Volume 15, p. 160.

- Van Tilburg, J., Van Haeften, T. W., Pearson, P.*et al*., 2001. Review articles Defining the genetic contribution of type 2 diabetes mellitus*. Journal of Medical Genetics*, 38(9), pp. 596-578.

- Vasa-Nicotera, M., Brouilette, S., Mangino, M. *et al*., 2005. Mapping of a major locus that determines telomere length in humans*. American journal of human genetics*, 76(1), pp. 147–151.

- Vera, E., Bernardes de Jesus, B., Foronda, M. *et al*., 2012. The rate of increase of short telomeres predicts longevity in mammals*. Cell Reports*, 2(4), pp. 732-737.

- Vergès, B. 2015. Pathophysiology of diabetic dyslipidaemia: where are we?. *Diabetologia*, 58(5), 886–899.

- Verhulst, S., Dalgård, C., Labat, C. *et al*., 2016. A short leucocyte telomere length is associated with development of insulin resistance. *Diabetologia*. 59(6), pp. 1258-1265.

- Verzola, D., Gandolfo, M.T., Gaetani, G. *et al*., 2008. Accelerated senescence in the kidneys of patients with type 2 diabetic nephropathy*. Am J Physiol Renal Physiol*, 295(5), pp. F1563

- Von Kanel, R., Malan, N.T., Hamer, M. *et al*., 2015. Comparison of telomere length in black and white teachers from South Africa: the sympathetic activity and ambulatory blood pressure in Africans study. *Psychomatic Medicine*, 77(1), pp. 26-32.

- Von Zglinicki, T. and Martin-Ruiz, C., 2005. Telomeres as Biomarkers for Ageing and Age-Related Diseases. *Current Molecular Medicine,* 5(2), pp. 197-203.

- Wall, J. D. and Pritchard, J. K., 2003. Assessing the Performance of the Haplotype Block Model of Linkage Disequilibrium. *American Journal of Human Genetics,*73(3), pp. 502-515.

- Walsh, K. M., Codd, V., Smirnov, I. V. *et al*., 2014. Variants near *TERT* and *TERC* influencing telomere length are associated with high-grade glioma risk HHS Public Access Author manuscript. *Nature Genetics*, 46(7), pp. 731-735.

- Wang, F., Pan, X., Kalmbach, K. *et al*., 2013. Robust measurement of telomere length in single cells.. *Proceedings of the National Academy of Sciences of the United States of America*, 110(21), pp. E1906-12.

- Wong, J. Y., De Vivo, I., Lin, X. *et al*., 2014. The relationship between inflammatory biomarkers and telomere length in an occupational prospective cohort study*. PLoS ONE*, 9(1), pp. e87348.

- World Health Organisation 2013, *World Health Statistics 2013,* World Health Organisation, viewed 25 July 2017, <http:// www.who.int/gho/publications/worldhealthstatistics/2013/en/>.

- Zee, R. Y., Castonguay, A.J., Barton, N.S. *et al*., 2010. Mean leukocyte telomere length shortening and type 2 diabetes mellitus: a case-control study. *Translational Research*, 155(4), pp. 166-169.

- Zee, R. Y., Ridker, P. M. and Chasman, D. I. 2011. Genetic variants of 11 telomere-pathway gene loci and the risk of incident type 2 diabetes mellitus: The Women's Genome Health Study. *Atherosclerosis,* 1 9, 218(1), pp. 144-146.

- Zeggini, E., Weedon, M. N., Lindgren, C. M. *et al*., 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science (New York, N.Y.),* 316(5829), pp. 1336–1341.

- Zeiger, A. M., White, M. J., Eng, C. *et al*., 2018. Genetic Determinants of Telomere Length in African American Youth. *Scientific reports*, 8(1), pp. 13265.

- Zeng, J.B., Liu, H.B., Ping, F. *et al*., 2019. Insulin treatment affects leukocyte telomere length in patients with type 2 diabetes: 6-year longitudinal study. Journal of Diabetes and its Complications. *Journal of Diabetes and Complications*, 33(5), pp. 363-367.

- Zhan, Y., Song, C., Karlsson, R. *et al*., 2015. Telomere Length Shortening and Alzheimer Disease—A Mendelian Randomization Study*. JAMA Neurology*, 1 10, 72(10), pp. 1202-1203

- Zhang, C., Doherty, J. A., Burgess, S. *et al*., 2015. Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Human molecular genetics*, 24(18), pp. 5356–5366.

- Zhao, Z., Pan, X., Liu, L. and Liu, N., 2014. Telomere length maintenance, shortening, and lengthening. *Journal of Cellular Physiology,* 229(10), pp. 1323-1329.

- Zhou, Y., Ning, Z., Lee, Y. *et al*., 2016. Shortened leukocyte telomere length in type 2 diabetes mellitus: genetic polymorphisms in mitochondrial uncoupling proteins and telomeric pathways. *Clinical and translational medicine*, 3, 5(1), pp. 8.

- Zhu, H., Wang, X., Gutin, B. *et al*., 2011. Leukocyte telomere length in healthy caucasian and african-american adolescents: Relationships with race, sex, adiposity, adipokines, and physical activity. *Journal of Pediatrics,* 158(2), pp. 215-220.

- Zhu, X., Han, W., Xue, W. *et al*., 2016. The association between telomere length and cancer risk in population studies. *Scientific Reports,* 26 4, 6(1), pp. 22243.

6 Appendices

6.1 Appendix A: Extraction of genomic DNA

DNA extraction

Blood samples (in EDTA as anti-coagulant) from participants of the AWI-Gen study were collected and DNA extracted in the SBIMB biobank. These DNA samples were already available and stored at -80 °C at the SBIMB Biobank, Parktown, South Africa. The DNA was extracted at the SBIMB using the salting out method (Miller *et al*. 1988), with minor modifications. DNA extraction using this method was divided into two parts and spans over two days. The first part involved completely thawing samples to room temperature. The samples were then transferred to an appropriately marked 50 ml NUNC™ tubes and filled with cold sucrose-Triton lysing buffer, which lyses the red blood cells. The tubes were inverted several times until completely mixed and placed in a centrifuge (Beckman Coulter, USA). Using this procedure, the blood was centrifuged at 2400 rpm at 4 °C for 10 min. After, the supernatant was poured off gently into a bleach containing container, ensuring that the pellet does not dislodge. The pellet was washed with 20 – 25 ml of cold sucrose-Triton X lysing buffer. After, the tubes were placed in a -20 °C freezer for 5 minutes and then spun down for 5 minutes. The supernatant was poured off as before. This pellet washing step was repeated at least twice to ensure that most of the red cell debris was completely removed. About 3 ml T20E5, 200 µl of 10% SDS and 500 µl of proteinase-K mixture were added to the pellet containing tube and mixed by inversion. The tubes were then incubated overnight at 42 °C. This step lyses the white blood cells to release the DNA to the solution and degrades proteins.

The following morning about 1 ml of saturated NaCl is added to the lysate containing tube and mixed by vigorous agitation for about 20 seconds. The tubes were then placed at -20 °C for 5 to 10 minutes. The tubes were then spun down for 30 minutes until a white pellet appeared at the bottom of the tube. After, the supernatant was transferred to an appropriately labelled 50 ml NUNC™ tube. The tube with the pellet was then discarded. To the supernatant, about two supernatant volumes of absolute ethanol was added and mixed by gentle agitation. This allowed the DNA to aggregate in the solution. The DNA was then “fished” out from the solution and

placed in an appropriately labelled 1.5 ml cryovial. In order to remove excess salt, the DNA was then washed with ice cold 70% ethanol. The DNA was then air dried by placing the tubes on the bench and then resuspended in 400 µl.

6.2 Appendix B: Quantification of genomic DNA

Genomic DNA QC

In order to reduce or prevent downstream process failures, the quality and concentration of the extracted DNA was assessed using a NanoDropTM Spectrophotometer (Thermo Fisher ScientificTM). The quality and concentration were assessed by measuring the ratio of absorbance of 1 µl DNA at 260/280nm wavelength and results recorded. DNA samples with a ratio of absorbance of approximately 1.8 to 2.0 were accepted as good quality according to genotyping service provider standards.

6.3 Appendix C: Table with SNPs and genes in AWI-Gen imputed data

Table 6.1 SNPs and genes in AWI-Gen imputed data

Gene and SNP	Role/function	Association and biological significance
ACYP2 **rs11125529**	Acylphosphatase 2, muscle type (*ACYP2*) has a specific role in muscle differentiation and stress induced apoptosis. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association between previously identified loci affecting TL and coronary heart disease (CHD) in Han Chinese population. Association analysis of TL related gene ACYP2 with the gastric cancer risk in the northwest Chinese Han population. Risk Allele C. OR 0.06 [0.036-0.076] kb decrease in TL.
AK097143 **rs34596385**	The record was discontinued because the defining sequence (AK097143.1) appears to be of retroviral origin and is not in scope for Entrez Gene. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele T
ASCC2 **rs73394838**	Encodes protein involved in the DNA damage repair pathway. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele G
BRUNOL4 **rs2162440**	Regulate pre-mRNA alternative splicing and may also be involved in mRNA editing, and translation. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association of *TERC* and OBFC1 Haplotypes with Mean Leukocyte TL and Risk for Coronary Heart Disease. Polymorphisms Associated with TL Affect Kidney Allograft Function After Transplantation. A genome-wide association study identifies a novel locus on chromosome 18q12.2 influencing white cell TL. Risk Allele G

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
PIKC3C **rs2162440**	**Protein Coding gene. Diseases associated with PIK3C3 include Amyotrophic Lateral Sclerosis. Involved in trafficking pathways. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Association of *TERC* and OBFC1 Haplotypes with Mean Leukocyte TL and Risk for Coronary Heart Disease. Polymorphisms Associated with TL Affect Kidney Allograft Function After Transplantation. A genome-wide association study identifies a novel locus on chromosome 18q12.2 influencing white cell TL. Risk Allele G**
C5orf42 **rs2098713**	The protein encoded by this gene has putative coiled-coil domains and may be a transmembrane protein. Defects in this gene are a cause of Joubert syndrome. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte TL in a Punjabi Sikh diabetic cohort. Risk Allele T OR 0.25 [0.15-0.35] kb decrease in TL.
CSNK2A2 **rs74019828**	It is involved in various cellular processes, including cell cycle control, apoptosis, and circadian rhythms. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	An intronic variant in this gene (alpha 2) may be associated with leukocyte TL in a South Asian population. Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte TL in a Punjabi Sikh diabetic cohort. Risk Allele A. OR 0.38 [0.26-0.50]kb decrease in TL.
CXCR4 **rs4452212**	This gene encodes a CXC chemokine receptor. It acts with the CD4 protein to support HIV entry into cells and is also highly expressed in breast cancer cells. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Risk Allele A. OR 0.05 [0.03-0.07] kb decrease in TL.
DCAF4 **rs2535913**	This gene encodes a WD repeat-containing protein that interacts with the Cul4-Ddb1 E3 ligase macromolecular complex. Diseases associated with DCAF4 include TL, and Mean Leukocyte. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	*DCAF4*, a novel gene associated with leucocyte TL.
DHX35 **rs6028466**	Implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly. Involved in embryogenesis, spermatogenesis, and cellular growth and division.	A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte TL and as a marker of susceptibility for bladder cancer. Leukocyte TL-related genetic variants in 1p34.2 and 14q21 loci contribute to the risk of oesophageal squamous cell carcinoma.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
DKK2 **rs7680468**	**Involved in embryonic development through its interactions with the Wnt signalling pathway. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele T**
PAPSS1 **rs7680468**	Three-prime-phosphoadenosine 5-prime-phosphosulfate (PAPS) is the sulphate donor co-substrate for all sulfotransferase (SULT) enzymes. SULTs catalyse the sulphate conjugation of many endogenous and exogenous compounds, including drugs and other xenobiotics. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele T
FXYD4 **rs10466239**	Encodes a family of transmembrane proteins. Among its related pathways are Cardiac conduction and Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele T
RASGEF1A **rs10466239**	Encodes protein involved in the integrated breast cancer pathway and RET signalling. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele T
ITSN1 **rs2300383**	Encodes Intersectin 1 which is a cytoplasmic membrane-associated protein that indirectly coordinates endocytic traffic. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American population (Stathoupouloua *et al*, 2015 and Zeiger *et al.,*2018).
KPNA5 **rs654128**	Encodes a gene that belongs to the importin alpha protein family and is involved in the transport of proteins into the cell. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Leukocyte TL-related genetic variants in 1p34.2 and 14q21 loci contribute to the risk of oesophageal squamous cell carcinoma. A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte TL and as a marker of susceptibility for bladder cancer. OR 0.1200 [NR] kb increase TL.
KRT80 **rs17653722**	Encodes type 2 epithelial keratin, involved in cell differentiation. (Sherry *et al.,*2001).	A genome-wide association study identifies a locus on *TERT* for mean TL in Han Chinese. Risk Allele T. OR 0.12 [NR] kb increase in TL.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
LOC100128993 **rs11787341**	**RNA gene, associated with non-coding RNAs. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele A**
rs10496920		Associated with TL in a juvenile African American and Asian populations.
LRP1B **rs10496920**	This gene encodes a member of the low density lipoprotein (LDL) receptor family. Disruption of this gene has been reported in several types of cancer. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American and Asian populations.
MYOM2 **rs12678295**	Encodes protein myomesin II, which is associated with the tintin molecule in forming the structure of sarcomeres (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American and Asian populations.
CSMD1 **rs12678295**	Encodes CSMD1 protein, which is a potential suppressor of squamous cell carcinomas. Associated with Epilepsy and Smallpox. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American and Asian populations.
NAF1 **rs7675998**	Encodes a protein coding gene that is RNA binding and is involved in ribosome biogenesis. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association between previously identified loci affecting TL and coronary heart disease (CHD) in Han Chinese population. Risk allele A
NCR2 **rs9357354**	Protein Coding gene. Among its related pathways are Hematopoietic Stem Cell Differentiation Pathways and Lineage-specific Markers and Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association with TL in a South Asian population.
NKAIN2 **rs594119**	This gene encodes a transmembrane protein that interacts with the beta subunit of a sodium/potassium-transporting ATPase. A chromosomal translocation involving this gene is a cause of lymphoma. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American and Asian populations.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
OBFC1 **rs9420907**	**Encodes protein that stimulates the activity of DNA polymerase-alpha-primase, the enzyme that initiates DNA replication. OBFC1 also appears to function in a telomere-associated complex with and TEN1. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/.**	**Association between previously identified loci affecting TL and coronary heart disease (CHD) in Han Chinese population. Identification of seven loci affecting mean TL and their association with disease. Risk Allele A. OR 0.07 [0.049-0.089] kb decrease in TL.**
rs9419958	Encodes protein that stimulates the activity of DNA polymerase-alpha-primase, the enzyme that initiates DNA replication. OBFC1 also appears to function in a telomere-associated complex with and TEN1. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/.	Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Risk Allele T. OR 0.08 [0.057-0.108] kb increase TL.
rs4387287		Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Risk Allele A. OR 0.10 [0.06-0.14] kb increase TL.
ONECUT1 **rs28790308**	This gene encodes a member of the Cut homeobox family of transcription factors. This gene may influence a variety of cellular processes including glucose metabolism, cell cycle regulation, and it may also be associated with cancer. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association with TL in a South Asian population.
PARP1 **rs1151814**	This gene encodes a chromatin-associated enzyme, which is involved in the regulation of various important cellular processes such as differentiation, proliferation, and tumour transformation and also in the regulation of the molecular events involved in the recovery of cell from DNA damage. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association with TL in a South Asian population.
PARVB **rs11703393**	This gene encodes a member of the parvin family of actin-binding proteins, which play a role in cytoskeleton organization and cell adhesion. This protein also functions in tumour suppression.	Associated with TL in a juvenile African American and Asian populations.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
PELI2 **rs398652**	**Protein coding gene involved in the pathway protein ubiquitination, which is part of Protein modification (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Leukocyte TL-Related rs621559 and rs398652 Genetic Variants Influence Risk of HBV-Related Hepatocellular Carcinoma. Risk allele G. OR 0.1200 [NR] kb increase in TL.**
PRR20 **rs9537514**	This gene is one of five identical loci in a cluster on chromosome 13q21.1. The predicted protein is proline-rich and contains several dopamine D4 receptor signatures and PRINTS domains (Sherry *et al.,*2001).	Association with TL in a South Asian population.
PXK **rs6772228**	This gene encodes a phox (PX) domain-containing protein which may be involved in synaptic transmission and the ligand-induced internalization and degradation of epidermal growth factors (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	A genome-wide association scan (GWAS) for mean TL within the COGS project: identified loci show little association with hormone-related cancer risk.
SYT16 **rs4902100**	Protein coding gene, which may be involved in the trafficking and exocytosis of secretary vesicles in non-neuronal tissues. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study. Risk Allele G
RTEL1 **rs2297439**	Encodes Regulator Of Telomere Elongation Helicase 1which functions in the stability, protection and elongation of telomeres and interacts with proteins in the shelterin complex. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association study of TL among South Asians identifies a second RTEL1 association signal.
rs6010620		Genetic variants in telomerase-related genes are associated with an older age at diagnosis in glioma patients: evidence for distinct pathways of glioma genesis. Variants near *TERT* and *TERC* influencing TL are associated with high-grade glioma risk. Association between regulator of telomere elongation helicase 1 polymorphism and susceptibility to glioma. Risk Allele G.
rs755017		Association between previously identified loci affecting TL and coronary heart disease (CHD) in Han Chinese population.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
TERC **rs12638862**	**Telomerase RNA Component is an RNA Gene. Forms part of the protein component of telomerase, which is a ribonucleoprotein polymerase that maintains telomere ends by addition of the telomere repeat TTAGGG. Telomerase expression plays a role in cellular senescence, as it is normally repressed in postnatal somatic cells resulting in progressive shortening of telomeres (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Association with TL in a South Asian population.**
rs1920116		Variants near *TERT* and *TERC* influencing TL are associated with high-grade glioma risk.
rs10936599		*TERC* polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. Identification of seven loci affecting mean TL and their association with disease. Risk Allele T. OR 0.10 [0.081-0.113] kb decrease in TL.
rs1317082		Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. Risk Allele G. OR 0.07 [0.046-0.089] kb increase TL.
rs10936601		A genome-wide association scan (GWAS) for mean TL within the COGS project: identified loci show little association with hormone-related cancer risk. Association of *TERC* and OBFC1 Haplotypes with Mean Leukocyte TL and Risk for Coronary Heart Disease. Risk Allele T
rs12696304		Common variants near *TERC* are associated with leukocyte TL in the Chinese Han population. Associations of *TERC* SNPs with Human Leukocyte TL and the Risk of Type 2 Diabetes Mellitus. Risk Allele G. OR 0.11 [0.08-0.14] kb decrease in TL.

Table 6.1 SNPs and genes in AWI-Gen imputed data… continued

Gene and SNP	Role/function	Association and biological significance
TERT **rs7705526**	**Telomerase Reverse Transcriptase) is a Protein Coding gene. Forms part of the protein component of telomerase, with reverse transcriptase activity, encoded by this gene. Telomerase expression plays a role in cellular senescence, as it is normally repressed in postnatal somatic cells resulting in progressive shortening of telomeres. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/**	**Multiple independent variants at the *TERT* locus are associated with TL and risks of breast and ovarian cancer. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with *TERT* expression.**
rs2736100		A genome-wide association study identifies a locus on *TERT* for mean TL in Han Chinese. Variants near *TERT* and *TERC* influencing TL are associated with high-grade glioma risk. Risk Allele C. OR 0.08 [NR] kb increase TL.
TMPRSS7 **rs16859140**	Protein coding gene that encodes a protein that hydrolyses peptides with Arginine. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte TL: the Long Life Family Study.
TRDMT1 **rs10904887**	This gene encodes a protein responsible for the methylation of aspartic acid transfer RNA. It is distinct and highly conserved in its function among taxa. Involved in macrophage differentiation and tRNA processing. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Associated with TL in a juvenile African American and Asian populations. Risk Allele T
WDR65 **rs621559**	The protein encoded by this gene belongs to the WD repeat-containing family of proteins, thought to function in craniofacial development, possibly in the fusion of lip and palate. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Leukocyte TL-related genetic variants in 1p34.2 and 14q21 loci contribute to the risk of oesophageal squamous cell carcinoma.
ZNF208 **rs8105767**	Zinc finger proteins (ZNFs), such as ZNF208, bind DNA and, through this binding, regulate gene transcription. Among its related pathways is gene expression. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Association between previously identified loci affecting TL and coronary heart disease (CHD) in Han Chinese population.
ZNF676 **rs1975174**	Encodes a zinc finger protein, implicated in gene expression pathways. (Sherry *et al*.,2001) available at https://www.ncbi.nlm.nih.gov/SNP/	Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. Risk Allele T.

6.4 Appendix D: Ethics Clearance Certificate

R14/49 Ms Sebentile Hleli

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1703101

NAME: (Principal Investigator)	Ms Sebentile Hleli
DEPARTMENT:	Human Genetics University of the Witwatersrand
PROJECT TITLE:	Telomere Length Associated with Type 2 Diabetes in a Sub-Saharan African Population
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-Study (M1403107)
SUPERVISOR:	Ms Cassandra Soo
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	12/04/2017

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/2nd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review in March and will therefore be due in the month of March each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 12 April 2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

6.5 Appendix E: Plagiarism declaration

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Sebentile Hleli Mthimkulu (Student number: 1521476) am a student registered for the degree of Master of Medicine (Human Genetics) in the academic year 2016.

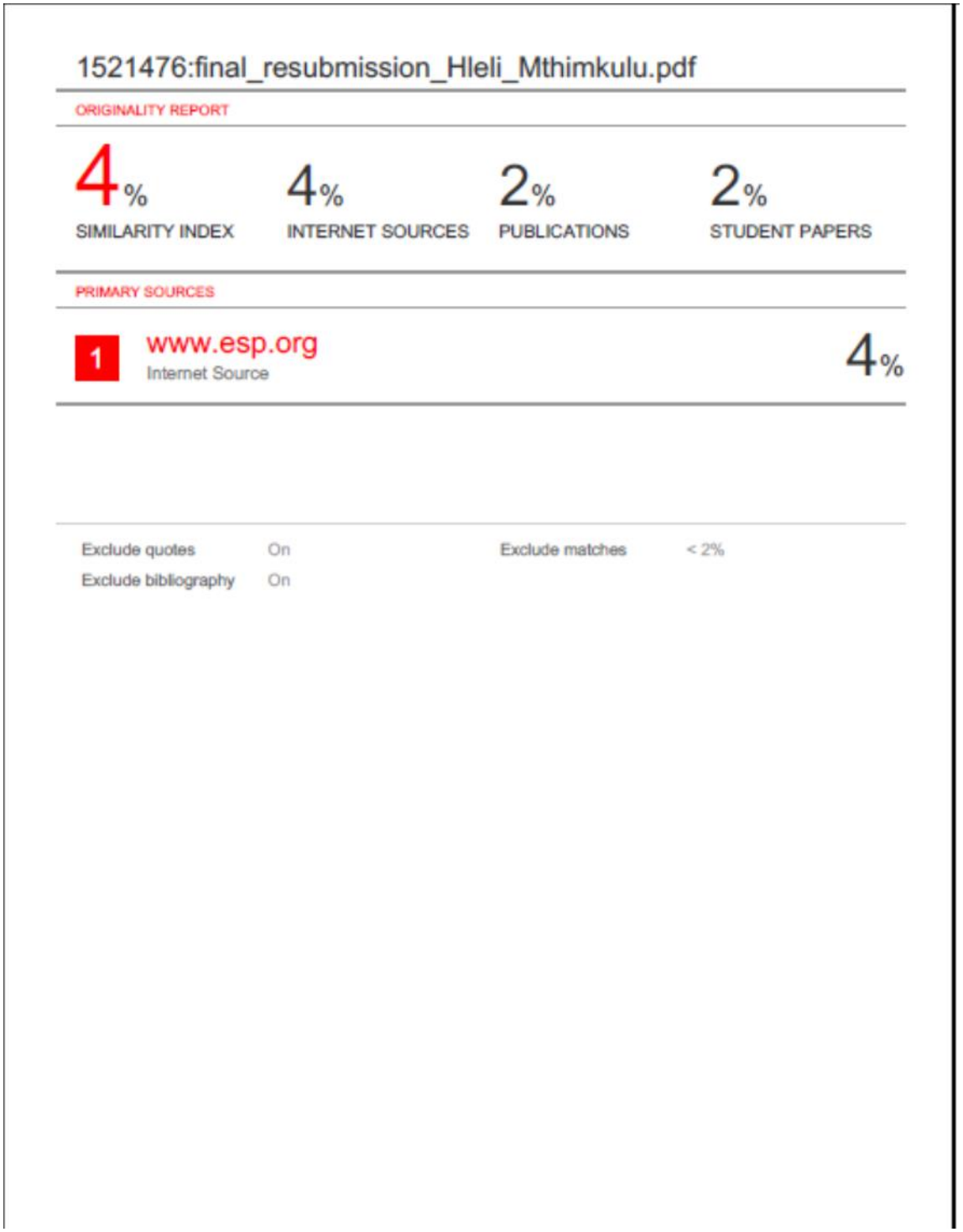
I hereby declare the following:

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

Signature: _____ Date: 28 March 2019

1

6.6 Appendix F: Turnitin report



1521476:final_resubmission_Hleli_Mthimkulu.pdf

ORIGINALITY REPORT

4%	4%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.esp.org Internet Source	4%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		