Investigating the genetic factors for gestational diabetes mellitus (GDM) in a black South African cohort

Nadine Botha

Student number: 1469585

Supervisors: Prof Zané Lombard and Dr Shelley Macaulay

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

Johannesburg, 2020

1 DECLARATION

I, Nadine Botha, 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 this or any other University.

NBotha

Nadine Botha

Date: 2nd day of July 2020 in Johannesburg

2 DEDICATION

To my loving family - Johann, Mariette, Charné, and Jacques Botha

3 PRESENTATIONS ARISING FROM THIS RESEARCH

Faculty of Health Sciences Research Day and Postgraduate Expo, 6 September 2018, University of the Witwatersrand, Johannesburg, South Africa

Poster presentation: Investigating the genetic factors for Gestational Diabetes Mellitus (GDM) in a black South African cohort

Southern African Society of Human Genetics (SASHG) Conference, 3-6 August 2019, Century City Conference Centre, Cape town, South Africa

Poster presentation: Investigating the genetic factors for Gestational Diabetes Mellitus (GDM) in a black South African cohort

4 ABSTRACT

Gestational diabetes mellitus (GDM) is defined as any degree of glucose intolerance first diagnosed during pregnancy. Many adverse pregnancy outcomes and long-term health implications exist for mothers with GDM, as well as their offspring. There is evidence of a genetic contribution to the risk of developing GDM; single nucleotide polymorphisms (SNPs) in Maturity Onset Diabetes of the Young (MODY) genes, and genes shown to influence type 2 diabetes (T2D) susceptibility, attribute 4- 10% of GDM cases. This study focused on selected SNPs from five MODY genes and one T2D-associated gene, and aimed to investigate if these variants were associated with GDM in a black South African cohort. Genotyping was carried out for 23 SNPs in DNA samples from 80 GDM-positive and 160 GMD-negative women, and the correlations were statistically assessed using PLINK. Analysis revealed that, rs4581569, an intronic SNP in the Pancreatic and Duodenal Homeobox 1 *(PDX1)* gene was significantly associated with a decreased risk of GDM and low fasting glucose levels. Since rs4581569 tags one other SNP (rs9512918) in the gene region, and may be linked to other SNPs, the SNP might only be indirectly associated with GDM. The associated SNP is not specific to the South African population as the minor allele frequency was similar globally when compared to publicly available genetic variation data from the 1000 Genomes Project. The association found is a novel discovery and prompts further investigation to establish the significance of this SNP as a protective factor against GDM development.

5 ACKNOWLEDGEMENTS

First and foremost, I want to thank my supervisors, Prof Zané Lombard and Dr Shelley Macaulay for all the time and effort you have offered while having me as your student. You have continually given me guidance and support to conduct novel research and to grow as a scientist/ researcher. You welcomed me with open arms, and with your influence, I was able to become fully integrated into the Division. I will forever be grateful for this opportunity and for all that I have learnt.

Thank you to all the students, the ones who have done (or are still doing) their studies on a full-time basis (Heather Seymore, Michaella Hulley, Patracia Nevondwe, Jessica Levesley, and Stephan Wessels), and also the part-time students, doing their it concurrently with routine laboratory work (Odirile Tabane, Maria Mudau, and Mahlatse Moremi). I value the friendship that we have built over the past few years. You have kept me motivated to keep going, and made me feel like we are in this together.

To my family. Thank you for always believing in me, and with your support allowing me to pursue my degree. Your love, emotional, as well as financial support is what got me through this journey.

6 FINANCIAL ACKNOWLEDGEMENTS

This work is based on the research supported in part by the National Research Foundation (NRF) of South Africa for the grant unique grant no. 94007, linked to the grant holder, Zane Lombard. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

TABLE OF CONTENTS

7 LIST OF ABBREVIATIONS AND SYMBOLS

AADM	Africa America Diabetes Mellitus
BMI	Body mass index
CADD	Combined Annotation Dependent Depletion
CNV	Copy number variant
DNA	Deoxyribose nucleic acid
DOHaD	Developmental Origins of Health and Disease
DPHRU	Developmental Pathways for Health Research Unit
dsDNA	Double-stranded DNA
EDTA	Ethylenediamine tetraacetic acid
EMP1	Pointwise *p* value
EMP2	*p* value corrected for multiple testing
eQTLs	Expression quantitative trait loci
FPG	Fasting plasma glucose
FRET	Fluorescence resonance energy transfer
GDM	Gestational diabetes mellitus
GRCh37	Genome Reference Consortium Human Build 37
GWAS	Genome-wide association Studies
HWE	Hardy-Weinberg Equilibrium
IQR	Interquartile range
LD	Linkage disequilibrium
MAF	Minor allele frequency
MALDI-TOF	Matrix-assisted laser desorption/ionization-time of flight
MODY	Maturity Onset Diabetes of the Young
NCBI	National Center for Biotechnology Information
OGTT	Oral Glucose Tolerance Test
OR	Odds ratio
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
S1000	Soweto First 1000 Days Study

SAP	Shrimp Alkaline Phosphatase
SBE	Single-base extension
SBIMB	Sydney Brenner Institute for Molecular Bioscience
SES	Socioeconomic status
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes mellitus
TBE	Tris-borate-EDTA
TBE	Tris-borate-EDTA
TE	Tris-EDTA
TF	Transcription factor
UCSC	University of California Santa Cruz
UV	Ultraviolet
VEP	Variant Effect Predictor
WHO	World Health Organization

Gene List

HNF4A	Hepatocyte nuclear factor 4α
TCF14	Transcription Factor 14
ABCC8	ATP binding cassette transporter subfamily C member 8
ADIPOQ	Adiponectin, C1Q and collagen domain containing
APPL1	Adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 1
BETA2	Beta-cell E-box transactivator 2
BLK	BLK Proto-oncogene, Src family tyrosine kinase
CAPN10	Calpain 10
CDK2A/B	Cyclin-dependent kinase inhibitor 2A/B
CDKAL1	CDK5 regulatory subunit associated protein 1 like 1
CEL	Carboxyl Ester Lipase
GCK	Glucokinase
GLIS3	GLIS family zinc finger 3
GPSM1	G protein signaling modulator 1

HHEX	Hematopoietically expressed homeobox
HMGA2	High mobility group AT-hook 2
HNF1A	Hepatocyte nuclear factor 1α
HNF1B	Hepatocyte nuclear factor 1β
IGF2BP2	Insulin like growth factor 2 mRNA binding protein 2
INS	Insulin
IPF1	Insulin promoter factor 1
IRS1	Insulin receptor substrate 1
IRS2	Insulin receptor substrate 2
KCNJ11	Potassium inwardly rectifying channel subfamily J member 11
KCNQ1	Potassium voltage-gated channel subfamily Q member 1
KLF11	Kruppel Like Factor 1
MTNR1B	Melatonin receptor gene 1 B
NeuroD1	Neuronal Differentiation 1
PAX4	Paired Box 4
PDX1	Pancreatic and Duodenal Homeobox 1
PPARG	Peroxisome proliferator activated receptor gamma
RREB1	Ras responsive element binding protein 1
SLC30A8	Solute carrier family 30 member 8
TCF1	Transcription factor 1
TCF7L2	Transcription factor 7 like 2
WFS1	Wolframin ER transmembrane glycoprotein
ZRANB3	Zinc finger RANBP2-type containing 3

Population names

CEU	Northern Europeans from Utah
ESN	Esan in Nigeria
GWD	Gambian in Western Divisions in the Gambia
LWK	Luhya in Webuye, Kenya
MSL	Mende in Sierra Leone
YRI	Yoruba in Nigeria

8 LIST OF TABLES

9 LIST OF FIGURES

1 LITERATURE REVIEW

Chapter 1 provides a background on aspects of gestational diabetes mellitus (GDM), including the definition/ diagnosis, prevalence, health implications, pathophysiology and the role of genetics in the development thereof. In the second section, genes found to be associated with GDM are discussed and more information is given on genes causative of Maturity Onset Diabetes of the Young (MODY), as well as, genes associated with type 2 diabetes mellitus (T2D). The third section presents important concepts and considerations for conducting genetic association studies. The last section describes the principles and advantages of the genotyping technologies used in this study. Lastly, the study rationale, aims, and objectives, are stated at the end of the chapter.

1.1 GESTATIONAL DIABETES MELLITUS

Gestational diabetes mellitus is defined as "any degree of glucose intolerance diagnosed for the first time during pregnancy" (Buchanan and Xiang, 2005). The only exception is that the level of hyperglycaemia should not fall within the overt diabetes range, which refers to pre-existing diabetes that is only identified during pregnancy (WHO, 2013). The Oral Glucose Tolerance Test (OGTT) is the standard test for diagnosing GDM. The World Health Organization (WHO) recommend a two-hour, 75 g OGTT and the criteria state that one or all of the following must be equalled or exceeded in order for GDM to be diagnosed: fasting plasma glucose (FPG) 5.1-6.9 mmol/l, 1-h plasma glucose ≥ 10.0 mmol/l and 2-hour plasma glucose 8.5-11.0 mmol/l (WHO, 2013). Significant risk factors for GDM include having a family history of type 2 diabetes mellitus, belonging to a particular ethnic group (e.g. African, Asian, Hispanic), being over the age of 25 years, being obese, and having had a previous stillbirth or a macrosomic baby (birth weight of ≥ 4.0 kg) (American Diabetes Association, 2014).

1.1.1 PREVALENCE OF GESTATIONAL DIABETES MELLITUS

Gestational diabetes mellitus is regarded as one of the most common maternal conditions that have become a global epidemic (Guariguata *et al.*, 2014). Global rates for GDM are shown in Table 1.1. Limited research has been performed on GDM in Africa, highlighting the necessity for further studies. In a systemic review on GDM in Africa, only six countries (11% of the African continent) had recorded prevalence rates of GDM (Macaulay *et al.*, 2014). A recent study reported a 9.1% GDM prevalence amongst black South African women in Johannesburg, South Africa (Macaulay *et al.*, 2018).

Table 1.1: Mean GDM prevalence in geographical regions and countries. Obtained from Zhu and Zhang (2016)

Geographic regions and countries	Mean prevalence
Middle East and North Africa	13%
Southeast Asia	11.7%
Western Pacific	11.7%
South and Central America	11.2%
Africa	8.9%
North America and Caribbean	7.0%
Europe	5.8%

1.1.2 HEALTH IMPLICATIONS OF GESTATIONAL DIABETES MELLITUS

Many adverse pregnancy outcomes and long-term health implications exist for the mother with GDM and her offspring. A primary short-term outcome is having a baby that is large for gestational age (>90th percentile) or one with macrosomia (≥4 kg). Secondary outcomes includes pre-eclampsia and other hypertensive-associated conditions (Coustan *et al.*, 2010). Independently, these conditions could result in delivery complications such as preterm delivery and Caesarean section. Birth trauma is also a risk; shoulder dystocia or birth injury can occur as a result of labour obstruction. After delivery, medical complications of the infant may include respiratory distress syndrome, cardiomyopathy, hypoglycaemia, hypocalcaemia, polycythaemia and death (Beischer *et al.*, 1996, Miller, 1946, Schmidt *et al.*, 2001). Altogether this forms a group of conditions that significantly increase the health and financial burden in many countries (Ferrara, 2007).

Offspring born to mothers with GDM, are at a higher risk of developing metabolic disorders, which manifest as T2D, obesity, and cardiovascular disease later in life (Dabelea, 2007). The exposure to high levels of glucose *in utero* is thought to have an impact on normal fetal development that can lead to permanent changes and altered fetal programming in glucose homeostasis. This concept has been derived from the Developmental Origins of Health and

Disease (DOHaD) hypothesis (Barker, 2007), introduced by Barker (1990), who found a link between low birth weight and increased rates of cardiovascular disease and hypertension. Since then, many studies have shown that non-communicable diseases, including T2D, are related to over- or under-nutrition in the unborn baby. The mechanism by which altered fetal programming occurs is possibly through epigenetic modifications of the fetal genome that alters the normal pattern of gene expression (Monteiro *et al.*, 2016). Differential methylation signals between diabetes-exposed and unexposed babies have been observed in genes involved in satiety/appetite, energy regulation, pancreatic development and β-cell function, providing evidence for the involvement of epigenetics in obesity and T2D in adulthood (Bouchard *et al.*, 2010, Carolan-Olah *et al.*, 2015, del Rosario *et al.*, 2014).

Women who have GDM are at risk of developing T2D in the future even though glucose metabolism usually reverts back to normal after delivery of their babies(Daly *et al.*, 2018, Kitzmiller *et al.*, 2007). Within five to 16 years of having been diagnosed with GDM, women have a 17-63% risk of developing T2D (Bellamy *et al.*, 2009, Hanna and Peters, 2002, Li *et al.*, 2018). Gestational diabetes mellitus can be controlled and the consequences to the mother and the offspring can be prevented through lifestyle modification (weight control, diet, and exercise) and/or medication (Lindsay *et al.*, 2017). Identifying women at increased risk for GDM is therefore important so that early preventative measures can be put into place as soon as hyperglycaemia is detected.

1.1.3 PATHOPHYSIOLOGY OF GESTATIONAL DIABETES MELLITUS

In pregnancy there is a normal increase in insulin resistance, due to the release of insulin-desensitizing hormones to accommodate for fetal nutritional demands (Reyes-López *et al.*, 2014, Sonagra *et al.*, 2014). Gestational diabetes mellitus arises when women are unable to lower the rising blood glucose through insulin secretion, which generally results from dysfunction of the pancreatic β-islet cells (Buchanan and Xiang, 2005). Pre-existing abnormal insulin resistance may also cause a greater change in insulin sensitivity. The precise pathophysiology of GDM, however, is still not fully understood. Nevertheless, insulin resistance and insulin secretion remains the two known factors responsible for the development of GDM (Retnakaran, 2017).

1.1.4 THE ROLE OF GENETICS IN THE DEVELOPMENT OF GESTATIONAL DIABETES MELLITUS

Gestational diabetes mellitus is established as a multifactorial condition. Thus, numerous genetic variants, together with environmental factors, collectively contribute to the development of the condition. Since a family history of diabetes increases the risk for developing GDM, a genetic contribution towards GDM development is plausible (Solomon *et al.*, 1997). Gestational diabetes mellitus also has a tendency to reoccur in subsequent pregnancies. A recurrence of 30% has been estimated amongst Hispanic women and a higher rate amongst other population groups (Latina, African American, Japanese and Asian); suggesting that these women belonging to particular ethnicities may be genetically predisposed to GDM (Kim *et al.*, 2007). Heritability estimates of a Danish twin study showed that both etiological factors, insulin secretion and insulin action, can be explained by 75-84% and 53-55% of genetic components, respectively (Poulsen *et al.*, 2005). Moreover, significant associations have been found between GDM and genetic loci in several genes that are also associated with T2D development (Robitaille and Grant, 2008). In genetic association studies, candidate genes are chosen based on biological plausibility (Lowe Jr *et al.*, 2016). For example, genes identified to be associated with GDM are known to be involved in insulin secretion, insulin resistance, and lipid and glucose metabolism. Many variants in the MODY genes and T2D-associated genes are also proven to occur in association with GDM (Shaat *et al.*, 2006, Watanabe, 2011).

1.2 GENES ASSOCIATED WITH GESTATIONAL DIABETES MELLITUS

The most recent systemic reviews, meta-analyses, and case-control studies have reported variants in four MODY genes, including *HNF1A*, *GCK*, *HNF4A*, and *KCNJ11*, that are significantly associated with an increased risk for GDM (Table 1.2). Most of these variants are recognized as T2D-associated variants, rather than MODY-associated variants, even though they are present in one of the many genes known to be responsible for monogenic diabetes (Vaxillaire *et al.*, 2012). Strong evidence also suggests that the *PDX1* gene, also known as *IPF1,* might be associated with GDM since four coding variants have been found in pregnant women with GDM and diabetes, and within individuals from families with clinical phenotypes of GDM, MODY, and T2D (Doddabelavangala Mruthyunjaya *et al.*, 2017, Gragnoli *et al.*, 2005a, Weng *et al.*, 2002). A common practice in finding variants associated with GDM is to choose candidate genes that have already been associated with T2D, as they share an identical pathophysiology (Ding *et al.*, 2018).

Table 1.2: MODY-linked gene functions (Firdous *et al.*, 2018, Wheeler *et al.*, 2013) and variants associated with gestational diabetes mellitus

Gene	Gene Function	Variants	References
HNF4A	Nuclear transcription factor that regulates hepatic and pancreatic beta cell gene expression	rs4812829	(Kanthimathi *et al.*, 2017)
GCK	Enzyme that catalyses the conversions of glucose to glucose-6-phosphate	rs1799884 (-30 G > A)	(Rosta *et al.*, 2017, Yang and Du, 2014, Zhang *et al.*, 2013)
		rs4607517	(Mao *et al.*, 2012)
HNF1A	Nuclear transcription factor that regulates insulin gene transcription and glucose transport metabolism	rs1169288 (I27L)	(Shaat *et al.*, 2006)
PDX1	Regulates transcription of genes: insulin, glucagon, glucose transporter *(GLUT2)* and *GCK* enzymes.	rs137852787 (Glu224Lys)	(Doddabelavangala Mruthyunjaya *et al.*, 2017)
		rs199644078 (Pro239Glu)	(Weng *et al.*, 2002)
		rs192902098 (Pri33Thr)	(Gragnoli *et al.*, 2005a)
KCNJ11	Regulate the potassium inward rectifier current and, thereby, beta cell depolarisation, the trigger for insulin release.	rs5219	(Mao *et al.*, 2012, Zhang *et al.*, 2013)

Table 1.3 shows T2D-related genes and variants, that have been reported to be associated with GDM in more than one study. Only one genome-wide association study (GWAS) for GDM has been conducted thus far, on a Korean population, where rs10830962 near *MTNR1B* and rs7754840 in *CDKAL1* achieved genome-wide significance (Kwak *et al.*, 2012). A more recent genetic association study on GDM in Europeans tested GWAS-confirmed T2D-associated genes, in two independent cohorts, consisting of 8722 women. Out of the 112 single nucleotide polymorphisms (SNPs), eleven variants (eight novel and three known GDM-associated variants) were identified in the following genes: *HNF1A, GLIS3, SLC30A8, RREB1*, *TCF7L2*, *MTNR1B*, and *GPSM1* (Ding *et al.*, 2018). Among the previously GDM-associated variants were rs10830963 (*MTNR1B*), rs1387153 (*MTNR1B*), and rs4506565 (*TCF7L2*); as well as rs7903146 (*TCFL2*), which was incorrectly determined as a novel discovery as it has been found before in association with GDM (Table 1.3). More and stronger association signals have been observed for genes related to the insulin secretory function, rather than those involved in insulin resistance.

This is in support of the understanding that a defective β-cell function is what causes the development of GDM, as reduced insulin secretion would be insufficient to overcome the increase in insulin resistance (Ding *et al.*, 2018, Voight *et al.*, 2010, Zhang *et al.*, 2013).

Table 1.3: Type 2 diabetes-associated genes associated with gestational diabetes

Gene	Variants	References
Genes and genetic variants related to insulin secretion		
TCF7L2	rs7903146	(Chang *et al.*, 2017, Ding *et al.*, 2018, Rosta *et al.*, 2017, Wu *et al.*, 2016, Zhang *et al.*, 2013)
	rs12255372	(Chang *et al.*, 2017, Zhang *et al.*, 2013)
	rs7901695	(Chang *et al.*, 2017)
MTNR1B	rs10830963	(Ren *et al.*, 2014, Rosta *et al.*, 2017, Wu *et al.*, 2016, Zhang *et al.*, 2013)
	rs1387153	(Liu *et al.*, 2016, Zhang *et al.*, 2013, Zhang *et al.*, 2014)
	rs10830962	(Kim *et al.*, 2011, Kwak *et al.*, 2012)
CDKAL1	rs7754840	(Kanthimathi *et al.*, 2017, Rosta *et al.*, 2017, Zhang *et al.*, 2013)
	rs7756992	(Cho *et al.*, 2009, Kanthimathi *et al.*, 2017)
KCNQ1	rs2237892	(Ao *et al.*, 2015, Huerta-Chagoya *et al.*, 2015)
	rs2237895	(Fatima *et al.*, 2016, Shin *et al.*, 2010, Zhou *et al.*, 2009)
IGF2BP2	rs4402960	(Kwak *et al.*, 2012, Mao *et al.*, 2012, Zhang *et al.*, 2013)
Genes and genetic variants related to insulin resistance		
IRS1	rs1801278	(Wu *et al.*, 2016, Zhang *et al.*, 2013, Zhang *et al.*, 2014)
SLC30A8	rs13266634	(Cho *et al.*, 2009)
	rs3802177	(Ding *et al.*, 2018, Kwak *et al.*, 2012)
ADIPOQ	rs266729	(Kasuga *et al.*, 2017, Pawlik *et al.*, 2017)

1.2.1 MODY GENES

Fourteen MODY types have been characterized thus far (Firdous *et al.*, 2018). They account for 1–2% of all cases of diabetes. The MODY subtypes are numerically categorized, from MODY 1 to MODY 14 (Table 1.4), each representing a single gene containing identified mutations causal to diabetes. The most commonly described MODY genes are the glucokinase gene (*GCK*), and two hepatocyte nuclear factor genes, *HNF1A* (also known as *TCF1*) and *HNF4A*, accounting for 32%, 52%, and 10% of MODY cases in the UK, respectively (Shields *et al.*, 2010). Other MODY types are uncommon in studied populations.

Table 1.4: Categories of MODY types and the single genes associated with each type (adapted from Firdous *et al.*, 2018)

Type	Gene	Chromosomal locus	Year of recognition
MODY1	Hepatocyte nuclear factor 4A *(HNF4A)/* Transcription Factor 14 *(TCF14)*	20q13	1991
MODY2	Glucokinase *(GCK)*	7p13	1993
MODY3	Hepatocyte nuclear factor 1A *(HNF1A*)/ Transcription factor 1 *(TCF1)*	12q24	1996
MODY4	Pancreatic and Duodenal Homeobox 1 *(PDX1)/* Insulin promoter factor 1 *(IPF1)*	13q12.2	1997
MODY5	Hepatocyte nuclear factor 1B *(HNF1B)*	17q12	1997
MODY6	Neuronal Differentiation 1 *(NeuroD1)/* Beta-cell E-box transactivator 2 *(BETA2)*	2q31	1999
MODY7	Kruppel Like Factor 11 *(KLF11)*	2p25	2005
MODY8	Carboxyl Ester Lipase *(CEL)*	9q34	2006
MODY9	Paired Box 4 *(PAX4)*	7q32	2007
MODY10	Insulin *(INS)*	11p15	2008
MODY11	BLK Proto-oncogene, Src family tyrosine kinas*e (BLK)*	8p23.1	2009
MODY12	ATP binding cassette transporter subfamily C member 8 *(ABCC8)*	11p15	2012
MODY13	Potassium inwardly rectifying channel subfamily J member 11 *(KCNJ11)*	11p15.1	2012
MODY14	Adaptor protein, phosphotyrosine interacting with PH domain and leucine zipper 1 *(APPL1)*	3p14.3	2015

Recent reviews have reported over 414 different *HNF1A* mutations in 1247 families, 103 *HNF4A* mutations in 173 families, and 620 *GCK* mutations in 1441 families of white European ethnicity (Colclough *et al.*, 2013, Colclough *et al.*, 2014, Osbak *et al.*, 2009). Approximately 60-65% of mutations identified in MODY genes are novel with only a handful of common mutations found within mutational hotspots and founder populations (Colclough *et al.*, 2014). These genes have been associated with T2D and GDM, they are highly polymorphic and many variants within these genes are of uncertain clinical significance (Vaxillaire and Froguel, 2008). With the advent of next generation sequencing, large amount of genetic variation data has become available, bringing about challenges in pathogenicity prediction and characterization, and disruption in the previous understanding of Mendelian disease inheritance. Some variants previously thought to be rare and disease-causing have recently been discovered to be more prevalent and sometimes harmless when studied in larger cohorts, including previously understudied populations (Auer *et al.*, 2012, Karki *et al.*, 2015, Myles *et al.*, 2008). When taking all the possible biophysical mechanisms responsible for disease into account, it can be difficult to classify genetic variants as neutral or deleterious. Upon re-evaluation of previously reported MODY causing mutations, some have been found to be less penetrant and related to a more complex form of diabetes with varying expressivity within the context of the whole population rather than within a specific family (Althari and Gloyn, 2015, Flannick *et al.*, 2016).

1.2.2 TYPE 2 DIABETES-ASSOCIATED GENES

To date, more than 80 genetic variants have shown robust signals in association with T2D (Fuchsberger *et al.*, 2016, Prasad and Groop, 2015). These loci explain only a limited part of the expected heritability of T2D. Effect sizes are determined to be relatively small with some significant variants only reaching an maximum odds ratio (OR) of 1.3, with the exception of variants within *TCF7L2* and *KCNQ1*, having estimated ORs at 1.37 and 1.40, respectively (Ali, 2013). Some of the most important T2D-associated genes includes: *TCF7L2*, *HHEX*, *SLC30A8*, *CDKN2A/B*, and *IGF2BP2*, which have been identified through genome-wide association studies. Linkage studies have only revealed two T2D-associated genes, namely: *TCF7L2* and *CAPN10.* Candidate gene association studies focusing on genes known to be involved in glucose metabolism, insulin secretion, insulin receptors, post-receptor signalling and lipid metabolism, have also produced some strong associated genes: *PPARG, IRS1*, *IRS2, WFS1,* and some MODY genes; *KCNJ11 HNF1A*, *HNF1B*, and *HNF4A* (Ali, 2013). Current research efforts trying to replicate these associations in different populations have been relatively successful. Many associations initially detected in Caucasian populations have been replicated in Asian populations, and vice versa, showing fairly good transferability of T2D-association loci across populations (Prasad and Groop, 2015).

A genome-wide analysis on T2D has only recently been published for sub-Saharan Africans from Nigeria, Ghana, and Kenya as part of the Africa America Diabetes Mellitus (AADM) study (Adeyemo *et al.*, 2015, Adeyemo *et al.*, 2019). The first analysis on 1035 cases and 740 controls, showed transferability of 11 loci, including rs7903146 SNP within *TCF7L2*, a genetic variant which is also associated with GDM. The *TCF7L2* SNP rs7903146 showed the strongest association ($p = 1.61 \times 10^{-8}$, OR 1.50, 95% CI 1.26–2.15) from all the variants assessed that have been genotyped by the Affymetrix Axiom® PanAFR array and imputed into the 1000 Genomes phase 1v3 reference panel (Adeyemo *et al.*, 2015). In the follow up analysis of ~18 million SNPs in 5231 individuals, *TCF7L2* was again the most strongly associated SNP (Adeyemo *et al.*, 2019). This T2D-associated SNP, along with two others; a novel genome-wide significant locus for T2D – the Zinc Finger RANBP2-Type Containing 3 (*ZRANB3*), and *HMGA2*, a known T2D-associated gene in Europeans and African Americans, achieved genome-wide significance (Adeyemo *et al.*, 2019).

1.3 GENETIC ASSOCIATION STUDIES

Genetic association studies are used to study the genetic contributors of a disease by finding a statistical correlation between a genetic variant and clinical disease phenotype (Balding, 2006, Bush and Moore, 2012, Carlson *et al.*, 2004). By finding a significant difference in allele frequencies between cases and controls, an allele's link to disease susceptibility can be inferred. Case-control studies can be used to explore the genetic associations for complex disease, and can be either family- or population based. Family-based study designs are usually employed for detecting genes or genomic regions linked to a disease that have an identifiable segregation pattern within a family. Studying disease in related individuals eliminates the chance of population stratification, as the family would have the same genetic background and disease susceptibility loci (Ott *et al.*, 2011). This type of design, however, falls short in pinpointing genetic variants that have a small to moderate effect on the disease phenotype. The sample size obtained from family pedigrees is generally too small to achieve significant power necessary to detect such associations. Therefore, to increase numbers, a sample is rather taken from unrelated individuals in a population-based case control design (Witte, 2010). Alternatively, when categorical variables of the binary case-control are not well defined, a quantitative trait-based approach would be more appropriate to follow in order to avoid arbitrary dichotomization. Using continuous phenotype measurements in this case is more informative, especially when the disease diagnosis is made based on quantitative trait cut-off values, such as body mass index (BMI), blood glucose levels, or blood pressure measurements (Newton-Cheh and Hirschhorn, 2005).

1.3.1 CONSIDERATIONS IN THE DESIGN OF A CASE-CONTROL STUDY

There are a number of important aspects to consider when designing a case-control study. These include the selection criteria for cases and controls, the sample size, the type of genetic variation, functionally significant genetic variation, and the extent to which the genome will be studied (candidate gene or a genome-wide association approach) (Zondervan and Cardon, 2007).

Strict criteria for the selection of case-control participants should be applied to ensure nearly homogenous groups; different and comparable only in two aspects - disease status, and the genetic difference under investigation. This is important in order to avoid selection bias and population stratification (or population admixture); the two main causes leading to false positive results or reduced power to detect genetic-disease associations. Hence, confounding factors (such as age, weight, ethnicity) potentially involved in disease susceptibility, should be excluded by cohort selection, case to control matching, or accounted for by statistical analysis methods. A less obvious phenotypic variable is that of ethnicity and the population substructure. Since the allele frequencies differ in various populations, spurious results can be obtained if the cases and controls are selected disproportionately from groups with different genetic ancestry, or groups that have undergone recent admixture (Clarke *et al.*, 2011). In terms of selecting the two groups, the cases selected should have tested positive for the disease and vice versa for controls in order to exclude the bias of misclassification (McCarthy *et al.*, 2008). Only in the situation where the disease is rare, and/or clearly identifiable, can the controls be randomly selected from the population.

1.3.2 GENETIC VARIATION

There are several types of genetic variation. The most common variations are SNPs, which are defined as loci with alleles that differ at a single base (Brookes, 1999). Biallelic SNPs are the predominant type within the genome. However, tri-allelic SNPs has also been found to present at sites beyond the expectant frequency (Hodgkinson and Eyre-Walker, 2010). To be able to identify a SNP within a population, the rarer allele should have a frequency of at least 1% in a random set of individuals. The rare allele is referred to as the minor allele, whereas the common allele is referred to as the major allele. The minor or major allele frequency is generally compared between case and control groups to assess the SNP's possible relation to disease. Due to the abundance of SNPs within the genome, high throughput genotyping technologies, and the amount of catalogues of SNP data available, SNPs are the simplest, and also the most commonly studied genetic variation in genetic association studies. Other variants that exist include: Micro-duplications and -deletions, microsatellite markers (or short tandem repeats, which are repetitive units of one to six base pairs), larger copy number variants (CNVs), transposable elements (e.g. *Alu* elements), and complex structural rearrangements (such as inversions) (Ku *et al.*, 2010).

1.3.3 LINKAGE DISEQUILIBRIUM AND TAGSNPS

Variation within the genome is structured in blocks of linkage disequilibrium (LD), meaning that variants within these regions are found together and thus inherited together more often than would occur by chance (Teare and Barrett, 2005). Linked variants usually occur in close proximity to one another, and the combination of alleles (termed haplotypes) are therefore less likely to be disrupted by recombination or mutational events that occur over time. The measurement of LD is based on the frequencies of alleles observed for the population, and the probability of the variants being dependent (in linkage disequilibrium) versus the probability of them being independent from one another (or randomly associated). Mathematically, the strength of LD is commonly calculated by D' or r^2. The equation of D (below) simplistically describes the definition of LD as the difference between the frequency of the pair of alleles at two loci (p_{AB}), and the product of the two frequencies (p_A and p_B). A calculated value of zero (0) implies that the alleles are independent, while a value of one (1) indicates that they are in linkage disequilibrium and have not been separated by recombination (Slatkin, 2008).

$$\mathrm{DAB} = \mathrm{pAB} - \mathrm{pApB}$$

The information on LD-blocks has brought some important applications for the study of variation in association studies. Since the linked variants co-occur within a specific population, it is possible to select only a couple of SNPs, called tagSNPs, to represent the variation present within a gene. Selecting tagSNPs is a cost-effective way to reduce the number of SNPs necessary to genotype (de Bakker *et al.*, 2005, Jorgensen *et al.*, 2009). TagSNPs, however, should be chosen from the LD data generated for a specific population, because populations have different degrees and patterns of LD (Sawyer *et al.*, 2005). Populations of African descent, for example, have undergone more recombination events, and therefore have more and smaller linkage equilibrium blocks than founder populations or more recently established populations. In this case, more tagSNPs would be required to account for all the genetic variation found within the population (Bush and Moore, 2012)

1.3.4 CANDIDATE GENE VERSUS GENOME-WIDE ASSOCIATION STUDIES

Two approaches, either a candidate gene or GWAS, are most often used when conducting an association study. With the candidate gene approach, genes are chosen based on prior knowledge of the genes' biological functions related to the disease of interest. In comparison, a GWAS is a hypothesis-free approach that studies variants commonly seen within the genome of the population (hence the word "genome-wide") (Amos *et al.*, 2011). Both approaches are based on the "common disease, common variant" hypothesis and make use of tagSNPs, representative of the common variation within the genome. A common SNP present in more than 1-5% of the population is identified through various sources, such as Hapmap and 1000 Genomes (Bush and Moore, 2012, Schork *et al.*, 2009, Shameer *et al.*, 2016). The consequence of the "common disease, common variant" concept is that important rare variants, that may explain remaining genetic contributors or the missing heritability of disease, are largely left unexplored (Wilkening *et al.*, 2009).

The candidate gene approach is advantageous in prioritizing a limited number of variants suited to the budget of a study. The disadvantage, however, is that potential causal genes that have limited biological information could be excluded when selecting only the most biologically plausible genes. Hence, novel gene-disease associations are unlikely to be discovered through the use of a candidate gene approach, and more likely by a GWAS. Candidate gene association studies are useful in replicating and validating genetic associations with disease in unexplored populations (Jorgensen *et al.*, 2009).

A GWAS on the other hand, also has some study design implications and limitations. Due to the multiple comparison nature of a GWAS, a high number of false-positive and false-negative results are likely. Therefore, more stringent significance thresholds allowing for multiple testing (such as the Bonferonni correction), are applied, and a larger sample size is required in order to achieve significance in GWAS compared to candidate gene association studies (Bush and Moore, 2012).

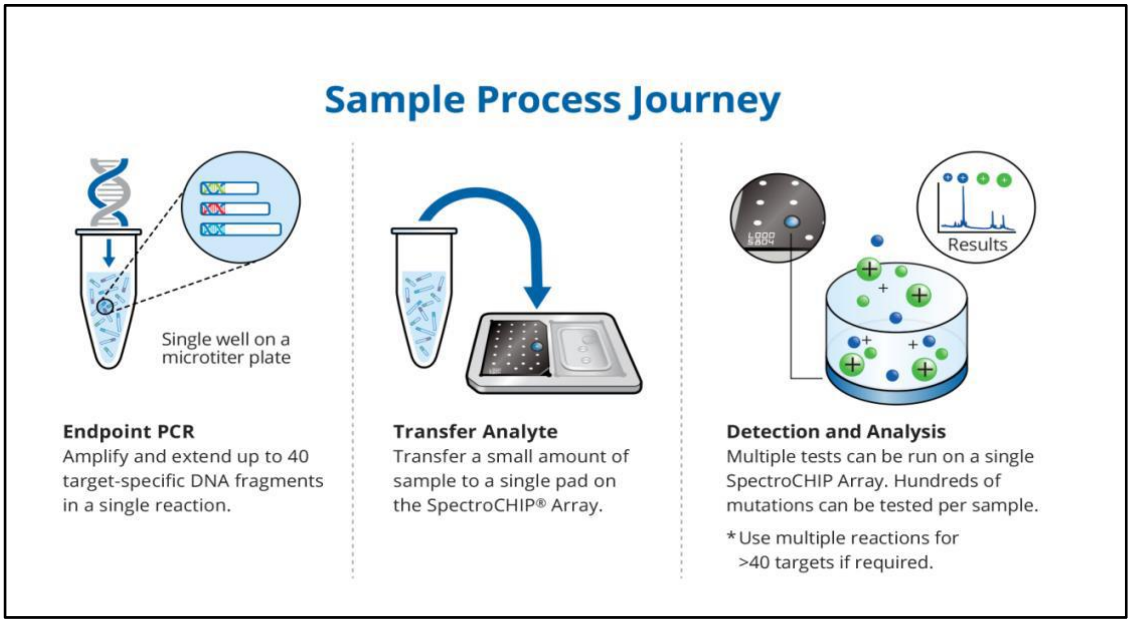
Currently, most GWAS are performed using data obtained by SNP arrays. Commercial arrays are constructed to include both common and rare variants (Tam *et al.*, 2019). However, the reference sequences used to select genotypes for the construction of commercial arrays are mostly of European descent. GWAS by the use of these arrays are thus limiting for the study of non-European populations (Wilkening *et al.*, 2009). For instance, coverage of only 43.3% of common variants was achieved for a T2D GWAS conducted on African Americans. Even though progress has been made to shift the focus of research to unrepresented populations, the increase in coverage and the inclusion of non-European samples have remained minimal (Bentley *et al.*, 2017).

1.4 GENOTYPING TECHNOLOGIES

Multiple technologies are available for genotyping SNPs in genetic association studies. For most technologies, an initial amplification of the target sequence by polymerase chain reaction (PCR) is required to achieve sensitivity and specificity. Thereafter, the methods by which the alleles are discriminated and detected differs between technologies (Kim and Misra, 2007). Critical factors that determine suitability of the technology involves the cost, the level of throughput (or multiplexing), accuracy, and the time required for assay design and optimization. No single technology satisfies the needs of every study. However, high throughput technologies, such as microarray, are usually utilized by GWAS for the genotyping of many SNPs within a large sample population. In contrast, technologies with lower through-put that allow for customization are preferred for smaller targeted studies, especially for replication and validation of associations found (Ellis and Ong, 2017). Custom assays, such as Taqman and MassARRAY, have the advantage of being robust and cost-effective with a sufficient level of through-put, flexible design and assay conversion rate for the genotyping of user-defined SNPs (Ragoussis, 2009).

1.4.1 MASSARRAY MALDI-TOF

MassARRAY differentiates alleles based on the mass of the allele-specific products that are generated by the methodology. This technology employs an initial amplification of the target region containing the SNP of interest, followed by single base extension (SBE) for allele discrimination, and matrix-assisted laser desorption/ionization time-of flight (MALDI-TOF) for allele detection. The workflow is outlined in Figure 1.1.



Endpoint PCR
Amplify and extend up to 40 target-specific DNA fragments in a single reaction.

Transfer Analyte
Transfer a small amount of sample to a single pad on the SpectroCHIP® Array.

Detection and Analysis
Multiple tests can be run on a single SpectroCHIP Array. Hundreds of mutations can be tested per sample.

*Use multiple reactions for >40 targets if required.

Figure 1.1: Genotyping of SNPs using Agena MassARRAY SpectroCHIP workflow (http://agenabio.com/products/massarray-system/).

The SBE reaction, or the discrimination assay, is constructed to incorporate a single mass-modified nucleotide at the polymorphic region, immediately upstream of the designed primer. The mass-extended products are then spotted onto a matrix-containing chip, which is loaded onto the MassARRAY instrument. The steps of MALDI-TOF involve vaporization and ionization of the sample, which is triggered by a short laser pulse. The matrix assisted sample molecules in vacuum are then electrostatically transferred, allowing for the separation of the sample molecules from the matrix ions, and acceleration towards the detector. The time taken for the sample ions to reach the detector (time-of-flight), are proportional to the square root of the molecules mass-to charge (m/z) ratio, which are the units recorded by the analysis software. By design, the mass of each allele is expected to occupy a unique position within the mass-spectrum (Gabriel *et al.*, 2009, Nakai *et al.*, 2002, Storm *et al.*, 2003).

MassARRAY is advantageous technique when it comes to multiplexing, since a wide mass spectrum range can be used to simultaneously distinguish and detect up to 40 SNPs (Gabriel *et al.*, 2009). With the utilization of the MassARRAY® Analyzer 4, oligonucleotides with a minimum difference of 16 Da can be detected within a mass range, ranging from 4500 Da to 9000 Da; illustrating the theoretical infinite multiplexing capability (Ellis and Ong, 2017). Due to the development of improved algorithms for multiplex assay design, the assay conversion rate is >80%, providing a success rate of 95% or higher, with an accuracy of more than 99% (depending on the sample quality). The iPlex technology also comes with an automated system, utilizing liquid handling instruments and PCR blocks, necessary for increased through-put, and reducing the risk of contamination (Ragoussis, 2009). The technology is very price competitive as a genotyping technology that could also be used for other applications, such as DNA methylation analysis, expression profiling, and proteomics (Jurinke *et al.*, 2002).

1.4.2 QUANTITATIVE REAL-TIME POLYMERASE CHAIN REACTION

Quantitative polymerase chain reaction (qPCR), or real-time polymerase chain reaction (real-time PCR), is a technology used to quantify and detect DNA sequence fragments, using fluorescence probes as they are being amplified in real time. Two reporter systems exist, namely, the intercalating SYBR (Synergy Brands Inc.) Green assay (Wittwer *et al.*, 1997) and the TaqMan probe system (Holland *et al.*, 1991, Livak *et al.*, 1995). The SYBR Green assay is based on a fluorescent probe binding to double-stranded DNA (dsDNA) of any sequence being amplified, and is more unspecific in comparison to the TaqMan method. However, specificity is achievable through proper primer design and optimization. The Taqman method is generally more expensive due to increased specificity obtained through the Taqman probe being used. Taqman is based on the complementary hybridization of fluorescent probes to the target sequence, and the exonuclease activity of the Taq polymerase enzyme.

Quantitative PCR has the advantage of precise quantification of nucleic acids, even in the case where the starting material is of low concentration. The time it takes for the fluorescence intensity to reach the detection threshold, correlates to the amount of starting material. Therefore, this technology is not only used for distinguishing one base pair differences in similar sequences, but is also used for quantifying viral load in patients (Ward *et al.*, 2004), assessing gene copy number in cancer tissue (Bieche*et al.*, 1998; Kindich*et al.*, 2005; Konigshoff*et al.*, 2003), and for studying gene expression levels when coupled with reverse transcription PCR.

1.4.2.1 TAQMAN 5’EXONUCLEASE ASSAY

The TaqMan® assay utilizes sequence-specific probes that hybridize to the DNA target region containing either the wild-type or the alternative allele. The probes contain a fluorescent dye on the 5’end and a quencher molecule at the 3’end (Figure 1.2). Before and during the hybridization step, while the probe is still intact, the fluorescence emitted by the flurophore is absorbed by the quencher molecule by fluorescence resonance energy transfer (FRET). The quencher is chosen based on the spectra fluorescence it can absorb, and is positioned from the fluorescence at a specific distance to maximize the amount of light captured. Different types of fluorescent dyes can be utilized, which permits allele detection of multiple SNPs in one reaction. During the synthesis of the complementary strand, the Taq polymerase cleaves the probe hybridized to the target site. At this stage, the reporter molecule gets separated from the quencher, resulting in the fluorescence being emitted and recorded by the detection instrument. The cycling of amplification causes an accumulation of fluorescence signal. Genotypic calls can be made at endpoint, by comparing the total fluorescence emitted by each probe (Holland *et al.*, 1991, Koch, 2004, Livak *et al.*, 1995).

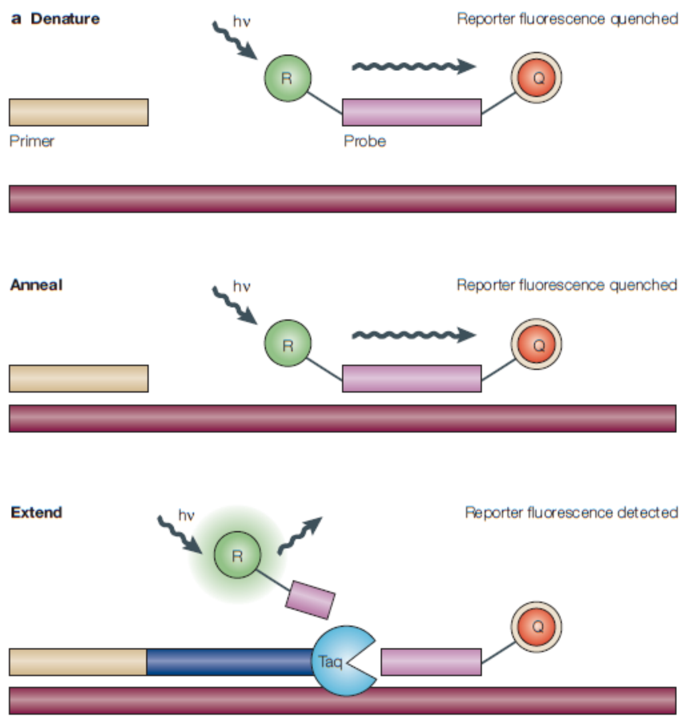

a Denature
hν
Reporter fluorescence quenched
R
Q
Primer
Probe
Anneal
hν
Reporter fluorescence quenched
R
Q
Extend
hν
Reporter fluorescence detected
R
Q
Taq

Figure 1.2: Schematic of the TaqMan SNP genotyping assay process. After initial denaturation, the intact probe anneals to the target region. The thermostable polymerases extend the primer and cleave the hybridize probes with 5'-nuclease activity. The reporter fluorophore (R) separated from the quencher (Q) and emit fluorescence when it is excited by an external light source (hv) (Obtained from Koch, 2004).

1.5 STUDY RATIONALE

Genetic variants associated with GDM development have been studied in many populations, especially European and Asian populations. Limited studies have been conducted on the black South African population to identify genes and genetic variants associated with GDM. Therefore, this study focussed on previously identified GDM-associated genes, especially genes causing MODY, and their association with GDM development in the black South African population. MODY genes were good candidates, since pathogenic variants in the MODY genes are known to cause the monogenic form of diabetes, and therefore there was a higher likelihood of finding an association with a strong genetic effect. A better understanding of the putative genetic factors contributing to GDM risk and impaired glucose tolerance amongst black South African women could possibly aid in the screening and identification of at risk individuals.

1.6 AIM AND OBJECTIVES

This project aimed to identify genetic factors associated with GDM risk in an urban black South African cohort. To achieve this the following objectives were set:

i) To select candidate genes to investigate, through literature review.

ii) To select SNPs to investigate, including African specific tagSNPs in candidate genes.

iii) To genotype GDM patients (n=80) and healthy controls (n=160) for the selected variants in the specific candidate genes.

iv) To assess the association of genotypes with GDM, and fasting blood glucose concentrations from the 75 g 2h OGTT as a measure of GDM risk.

v) To characterize any associated variants identified in terms of frequency, type of mutation, and putative functional effect, and compare to other global populations.

2 METHODOLOGY

This chapter describes the sample collection and selection used for this study, and the main methodology employed to achieve the objectives. An in-depth description is given on the methods used for selecting tagSNPs, the genotyping platforms, the quality control measures, and the statistical tests used. The approach to characterize the SNP found to be significantly associated is also described. The ethical considerations regarding the permission to use biological specimens and phenotypic data from participants, as well as using African-specific SNP data, are addressed. The laboratory work and analyses were executed at the Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, South Africa. Each step of the methodology executed by the relevant parties is outlined in the flow diagram (Figure 2.1).

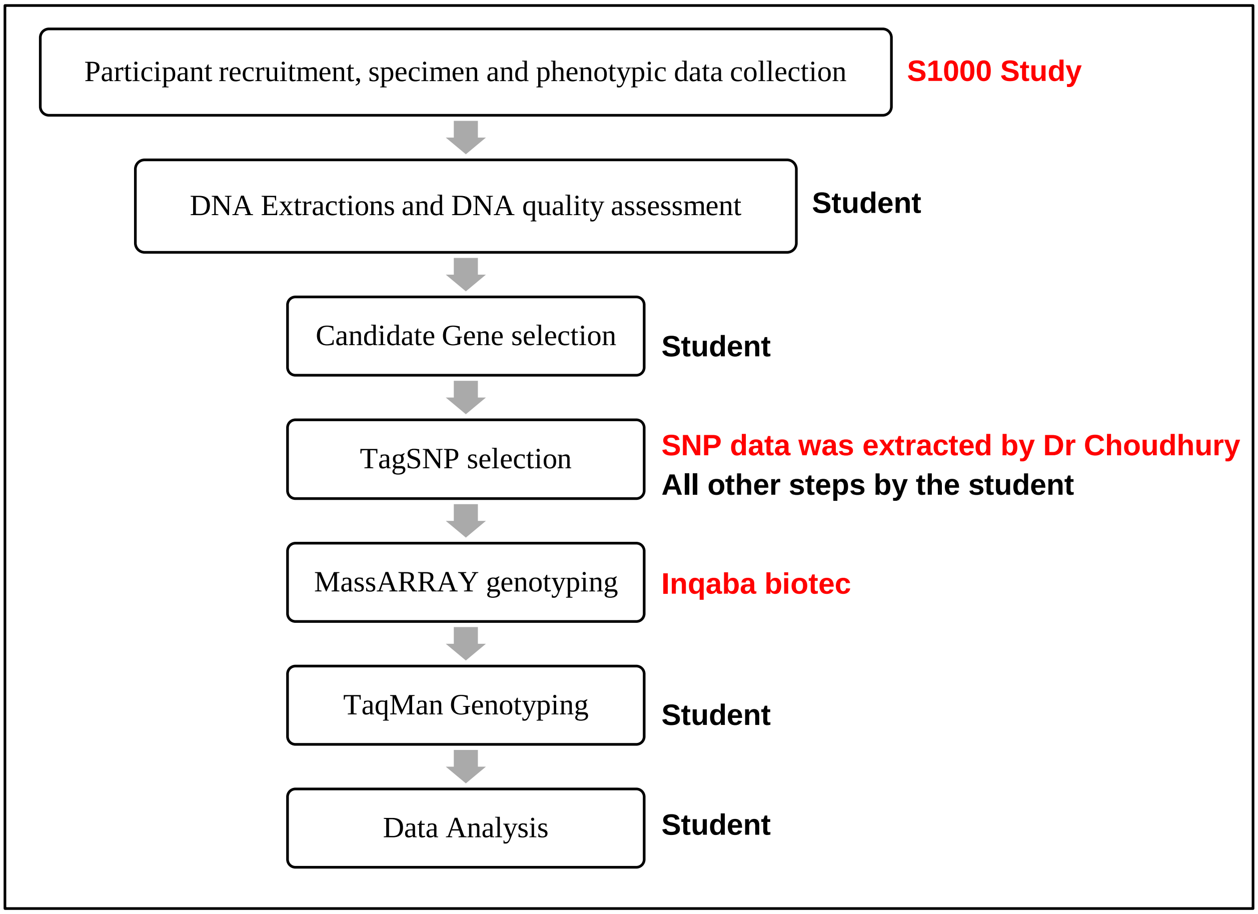

Participant recruitment, specimen and phenotypic data collection
S1000 Study
DNA Extractions and DNA quality assessment
Student
Candidate Gene selection
Student
TagSNP selection
SNP data was extracted by Dr Choudhury
All other steps by the student
MassARRAY genotyping
Inqaba biotec
TaqMan Genotyping
Student
Data Analysis
Student

Figure 2.1 Flow diagram of methodology executed by different parties.

2.1 STUDY SAMPLES

Study samples and phenotypic data were obtained from the Soweto First 1000 Days Study (S1000), a study conducted at the MRC/Wits Developmental Pathways for Health Research Unit (DPHRU) in Soweto (Johannesburg, South Africa) from June 2013 to April 2017. The overarching aim of the S1000 study was to investigate maternal factors influencing fetal development and birth outcomes. Participants were recruited early in pregnancy (≤14 weeks, but no later than 20 weeks pregnant) and followed up closely through to delivery. At 24-28 weeks' gestation all women underwent a two-hour 75 g OGTT as per the WHO 2013 diagnostic criteria for the diagnosis of GDM (Metzger *et al.*, 2010). Participants were self-reported black South African females residing in Soweto, an urban metropolitan area in Johannesburg (Macaulay *et al.*, 2018). The full set of criteria for the inclusion and exclusion of participants are outlined in Table 2.1.

Table 2.1: The inclusion and exclusion criteria of participants in the S1000 study

Inclusion Criteria	Exclusion Criteria
Black South African females Residing in Soweto ≥18 years of age ≤20 weeks pregnant at time of recruitment Pregnant with singleton pregnancies Non-diabetic at time of recruitment Able to give informed consent	Known diabetic women (at the time of recruitment) Multiple pregnancies (twins, triplets etc) Fetal abnormalities detected on ultrasound

Phenotypic data obtained from the women in the S1000 study included the woman's age and body mass index (BMI) at enrolment, number of previous pregnancies (gravidity), level of education and household socioeconomic status (SES). The household asset score was used as a proxy to determine a woman's SES, and was based on the number of household items a woman owned (a total of nine items described in more detail under Section 2.2.4). Body mass index was classified into the WHO categories for underweight (< 18.5 kg/m2), normal weight (≥ 18.5-24.9 kg/m2), overweight (≥ 25-29.9 kg/m2) and obese (≥ 30 kg/m2) (WHO, 2017). Whole blood samples for DNA extraction were available on 80 women with GDM and 160 controls (women who tested negative for GDM).

2.2 ETHICAL CONSIDERATIONS

This study received clearance by the Human Research Ethics Committee (Medical) (HREC (M)) of the University of the Witwatersrand (certificate no. M170851, Appendix A). Similarly, the S1000 study was also approved (certificate no: M120524), and a letter of permission for the use of data and specimens collected were given by the director of the DPHRU, Professor Shane Norris (Appendix B). All participants from the S1000 study had given informed consent for blood extraction and DNA analysis. Samples had been code labelled by the S1000 study to maintain the confidentiality of the participants. Any other information was strictly limited for the use by the investigator and supervisors. To achieve the objective of selecting African-specific SNPs, the Zulu dataset from the African Genome Variation Project (AGVP) was accessed under pre-approved data access permission given to the Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand.

2.3 POWER ANALYSIS

An initial power analysis was performed using Quanto (version 1.2.4, May 2009). The power for this study, that consisted of 80 cases and 160 controls (1:2 ratio), was calculated for the study population under the log-additive inheritance model and using a population-level GDM risk value of 9%, based on the GDM prevalence observed in the study by Macaulay *et al.* (2018). As the minor allele frequency (MAF) and the effect size of the alleles were unknown at the time, alleles with a frequency of 0.05 to 0.45 at a two-sided α of 0.05 (i.e. opposite direction of effect) for odds ratio (OR) ranging from 0.1 to 6.0, were used to survey the spectrum of power expected. Figure 2.2 shows the graph depicting the power against the ORs obtained for alleles with different MAF in the population. In this study, there was a power of 80% to detect an association of an allele with an effect (or OR) of 2 with an allele frequency of at least 0.16. In the opposite direction, a power of 80% could be achieved for a SNP with an effect of 0.4 with an allele frequency greater or equal to 0.15. In order to achieve a power of 80% for rarer alleles, the allele must have a large effect on the disease phenotype, as indicated by extreme ORs away from 1.

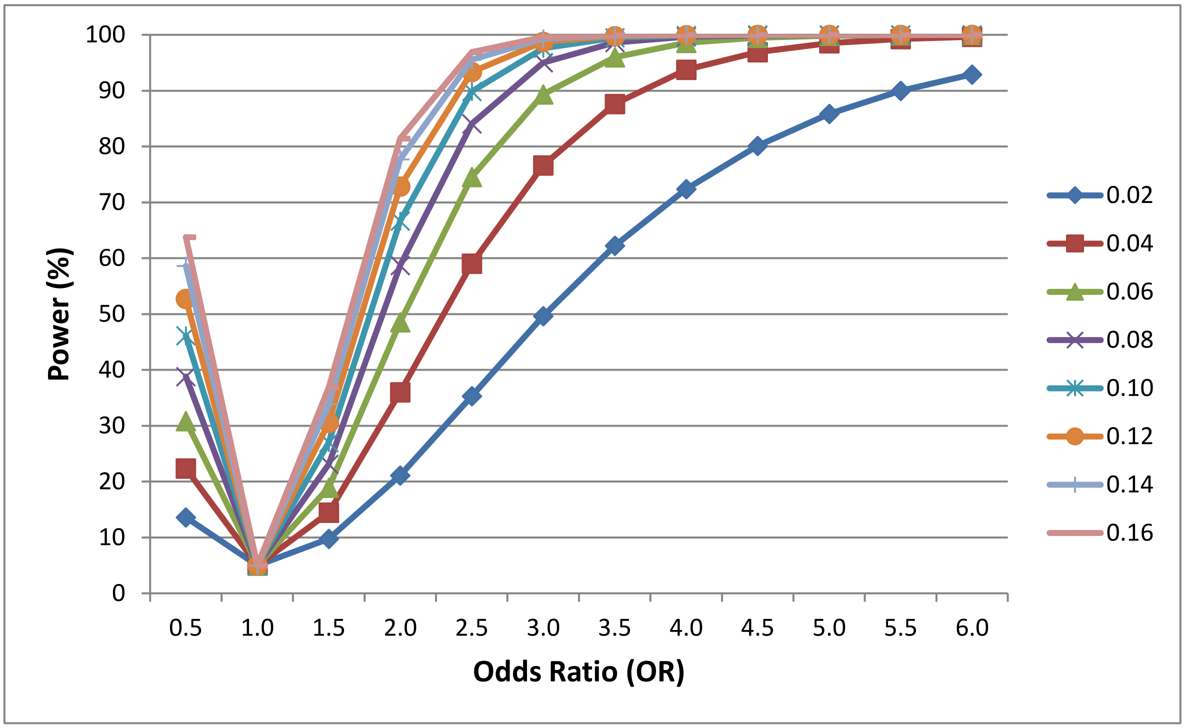


Figure 2.2: A graph depicting the power (%) over odds ratio for SNPs with different allele frequencies.

2.4 COHORT CHARACTERIZATION

Data analysis tools from the free add-in, Real Statistics (http://www.real-statistics.com/) on Microsoft© Excel© for Office 365 were used to summarize and assess the distribution of the phenotypic data (Table 2.2) obtained from the study participants. Variables were compared by the use of R Project for statistical computing v.3.5.3 (R Core Team, 3.5.3), and the user-friendly interphase RStudio (RStudio Team, 2015). Body mass index, the number of previous pregnancies (gravidity), and the level of education were represented as categorical variables as described in Table 2.2. Body mass index was assessed as both a categorical and a continuous variable. The Shapiro-Wilk test (Shapiro and Wilk, 1965) was used to test for normality of the phenotypic data. Normally distributed continuous data were presented as means and standard deviations, and non-normally distributed data were presented as medians and interquartile ranges (IQRs). For categorical variables, the data were presented as frequencies and percentages. The continuous data that were normally distributed and those that were not normally distributed for GDM positive women and women without GDM (control group) were compared using the Student's t-test, and the nonparametric Mann–Whitney U test, respectively. The differences between the categorical variables between cases and controls were assessed using the Pearson's Chi Square test. A statistically significant finding was defined as $p < 0.05$.

Table 2.2: Phenotypic data used to describe the study participants

Characteristics	Description (Continuous or categorized variables)
Age	In years
BMI at enrolment in early pregnancy	0 Normal (≤24.9 kg/m^2) 1 Overweight (25-29.9 kg/m^2) 2 Obese (≥30 kg/m^2)
Gravidity	0 None 1 One to two 2 Three or more
Level of Education	0 No schooling/ primary school 1 Secondary school 2 Tertiary education
Socio-economic status: Sum of household items	1 Electricity 2 Radio 3 Television 4 Refrigerator 5 Cellphone 6 Personal computes 7 Farm animals 8 Agricultural land 9 Bicycle 10 Motorcycle/ scooter 11 Car/truck/tractor

2.5 DNA EXTRACTION

The isolation of DNA was performed from approximately 5 ml of whole blood in ethylenediamine tetraacetic acid (EDTA) tubes. Two different extraction methods were employed, a magnetic bead-based technology using the Chemagic 360 instrument (Perkin Elmer, Baesweiler, Germany) and a single-tube precipitation method technology utilized by the Flexigene DNA kit (Qiagen, Novato, CA). The kit-based method was only used for a few extractions, whereas the automated method was used for the majority of samples.

2.5.1 CHEMAGIC 360

The purification protocol and kit (Product no CMG-703) for 5 ml of human blood for the Chemagic^{TM} 360 instrument from PerkinElmer (Germany) was used for DNA extraction. The instrumentation is equipped with liquid handling robots, magnetic rods, and QA software in an enclosed automated system. The PerkinElmer Chemagen Technology employs a unique separation method whereby nucleic acids bind to functionalized magnetic particles (M-PVA Magnetic Beads), that are subsequently attracted to transiently magnetized metal rods (Hübner *et al.*, 2017). The bound nucleic acids are then transferred through the different process solutions and finally resuspended and washed by the deactivated electromagnet and the concerted rotation of the magnetic rods (Berensmeier, 2006).

Only a few manual steps were required for sample and equipment preparation, and final purified sample retrieval. Set-up included placing rod covers, empty 50 ml, and pre-filled 50 ml tubes and 30 ml elution tubes in their respective positions on tube racks on the apparatus. Blood samples were decanted into 50 ml tubes with the addition of 20 µl protease. 600 µl Magnetic beads were added to the second set of 50 ml tubes, and 300 µl elution buffer were added to the 30 ml elution tubes. Buffers in Table 2.3 were automatically resuspended during the operation, with the supply being checked and refilled as required. The minimum filling volumes for each buffer container is also stipulated in Table 2.3.

Table 2.3: Minimum filing volume for each buffer container to process 12 samples

Buffer	Position	Minimum Filling Volume for 12 samples
Lysis Buffer 1	1	175 ml
Binding Buffer 2	2	300 ml
Wash Buffer 3	3	200 ml
Wash Buffer 4	4	200 ml
Wash Buffer 5	5	200 ml
Wash Buffer 6	6	200 ml

The final manual step was to pipette the purified product out from the 13 ml elution tubes, placed on a magnetic block, to allow for the remaining magnetic beads to stick to the bottom of the vessel. A schematic diagram showing the steps of this technique is represented in Figure 2.3.

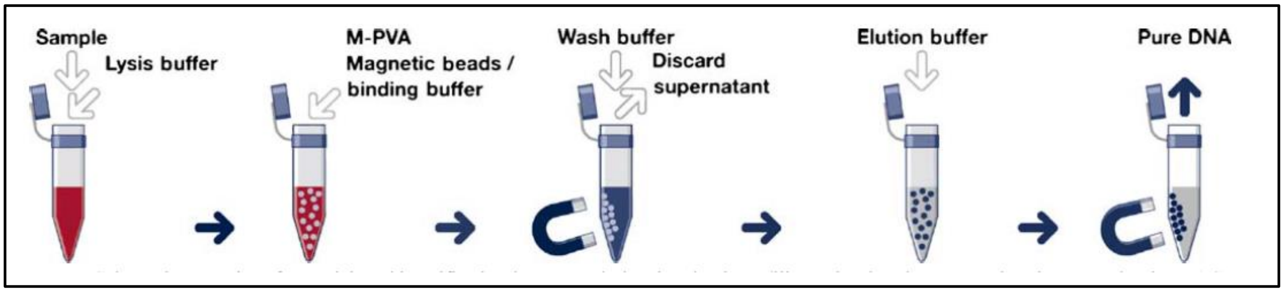


Figure 2.3: A schematic procedure for the isolation of nucleic acids by magnetic bead-based technology (illustration by Chemagen Biopolymer-technology AG, Germany).

2.5.2 FLEXIGENE DNA KIT

The Flexigene DNA kit (Qiagen, Novato, CA, USA) uses a single tube precipitation method technology, requiring centrifugation steps as indicated in Figure 2.4. In the first step, cell membrane lipids were broken down by the addition of 10 ml lysis buffer to blood decanted into 50 ml tubes, and mixed by vortexing and inverting the tubes. Thereafter, the samples were centrifuged for 5 min at 2000g at room temperature with the Allegra® X-30 Series Benchtop Centrifuge (Beckman Coulter, Brea, CA, USA) to pellet cell nuclei and mitochondria. The supernatant was then gently discarded, preventing backflow of supernatant while recovering the pellet. To remove proteins, the pellet was resuspended in freshly prepared 2 ml denaturation buffer, containing QIAGEN Protease and chaotropic salts. The samples were immediately vortex after each single addition of denaturation buffer until completely homogenized. The samples were incubated for 25 min at 65°C, until a colour change from red to olive green was observed. After incubation, 2 ml 100% isopropanol was added and mixed by inversion, which caused DNA precipitate to appear. The DNA precipitate was pelleted by centrifuging the samples for 5 min at 2000g at room temperature and discarding the supernatant. Lastly, the DNA aggregate was washed by adding 2 ml 70% ethanol and repeating the last centrifugation step. DNA was dissolved and stored in 200-400 µl Tris-EDTA (TE) buffer.

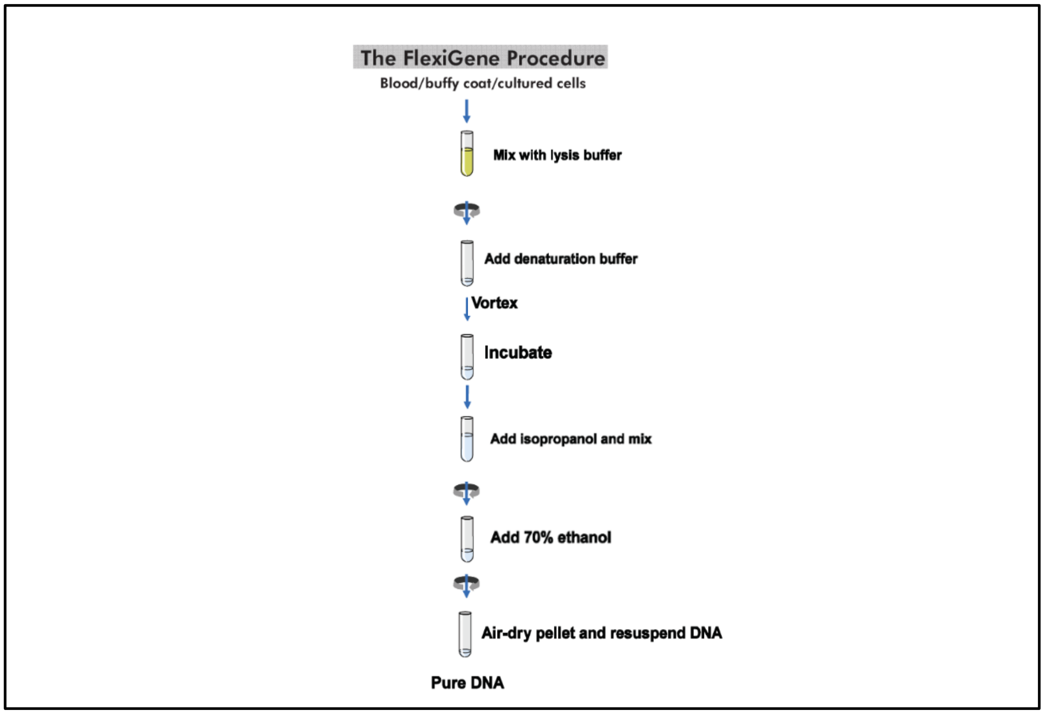


Figure 2.4: The FlexiGene procedure for extracting genomic DNA from variety of samples (Illustration from the FlexiGene Handbook).

2.6 QUALITY ASSESMENT OF EXTRACTED GENOMIC DNA

Spectrophotometric analysis was performed to measure the DNA quality and quantity of the extracted genomic DNA, using the Thermo Scientific NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, USA). Spectroscopy is based on the principle that molecules absorbs light at certain wavelengths. When a DNA sample is exposed to 260 nm ultraviolet (UV) light (the maximum absorbance of DNA), some light will be absorbed, and the other fraction will move through the nucleic acid. The amount of light that reaches the photodetector on the other end depends on the DNA concentration of the sample. By applying the Beer-Lambert Law, the DNA concentration can be determined. To detect unwanted molecules (RNA, proteins, nucleotides, and aromatic compounds) that can also absorb at or near 260 nm, the sample is also excited with light at 230 nm and 280 nm. Before the sample was measured, the spectrophotometer was blanked with TE buffer, to obtain a zero reference to which the sample was compared in the analysis. Thereafter, the optical measurement surface was wiped off with double distilled water and a paper towel, and 1 µl of DNA sample was pipetted onto the surface. The DNA concentration was measured in ng/µl. The ratio of absorbance at 230 nm, 260 nm, and

280 nm (A260/A280 and A260/A230, respectively) were used to assess the purity of the DNA. A A260/A280 ratio of ~1.8 , and a A260/A230 ratio of ~2.0 were regarded as “pure” based on the general practise (Green and Sambrook, 2018, Koetsier and Cantor, 2019). If the ratios were remarkably lower or higher in either case, it indicated the presence of contaminants. For instance, the presence of protein can be inferred based on a low A260/280 ratio, whereas RNA contamination can be inferred based on the increase of this ratio. Contaminants, including chaotropic salts, EDTA, non-ionic detergents, etc. can be inferred by A260/A230 <1.8 (Koetsier and Cantor, 2019).

In addition to verify the integrity and quality of the DNA, agarose gel electrophoresis was also performed. Agarose gel electrophoresis is a method routinely used for separating DNA molecules. When DNA is subjected to an electric field, the negatively charged DNA move through the porous gel to the positive anode at a rate proportional to its size. Small fragments migrate faster, whereas larger DNA molecules migrate slower (Green and Sambrook, 2019). In this study, a 0.8 w/v agarose gel were used, TBE (Tris-borate-EDTA) buffer solution, and ethidium bromide (EtBr), an intercalating agent, were used to stain the DNA. For loading of DNA, 5 µl of loading dye, containing ficoll, were mixed with 1 µl of the DNA sample, and loaded onto the gel submerged in TBE buffer. The DNA ladder, consisting of 1 µl of 1 kb molecular weight marker and 5 ul loading dye, were loaded into the first well on the gel. The gel tank was connected to an electric current of 10 V/cm for about 1-1.5 hours until the DNA have fully migrated out of the wells. The Omega Fluor™ Gel Documentation Systems (Vacutec, Alpegen) was used to illuminate and produce images of the stained gels for visualization. The intensity of the fluorescence observed gave a rough estimate of the amount of genomic DNA present in the samples. A single bright band indicated that DNA extraction was successful in producing good quality DNA of a high concentration. As genomic DNA has a high molecular weight, it moves slower in comparison to smaller DNA fragments, and therefore, its position on the gel was expected to be above the top band of the 1kb molecular weight marker. Any other bands on the gel indicated contamination or DNA fragmentation.

2.7 CANDIDATE GENE SELECTION

An extensive literature review was conducted to identify common MODY genes and MODY gene variants significantly associated with increased GDM risk in other populations. Based on the most recent systemic reviews, meta-analysis and case-control studies, four MODY genes, and one T2D-associated gene were selected for investigation in this study (shown in Table 2.4). Table 2.4 also shows the specific variants within each gene that were previously found to be significantly associated with GDM risk by the respective studies. The positions of the GDM-associated variants were used for the selection of tagSNPs, described in the following section.

Table 2.4: Gene and gene variants significantly associated with increased GDM risk

Type	Gene	Variants	References
MODY 1	Hepatocyte nuclear factor 4α *(HNF4A*)	rs4812829	(Kanthimathi *et al.*, 2017)
MODY 2	Glucokinase *(GCK)*	rs1799884 rs4607517	(Rosta *et al.*, 2017, Yang and Du, 2014, Zhang *et al.*, 2013) (Mao *et al.*, 2012)
MODY 3	Hepatocyte nuclear factor 1α *(HNF1A)*	rs1169288	(Shaat *et al.*, 2006)
MODY 4	Pancreatic And Duodenal Homeobox 1 *(PDX1)*/ Insulin promoter factor 1 *(IPF1)*	rs137852787 rs199644078 rs192902098	(Doddabelavangala Mruthyunjaya *et al.*, 2017) (Weng *et al.*, 2002) (Gragnoli *et al.*, 2005a)
MODY 13	Potassium Voltage-Gated Channel Subfamily J Member 11 *(KCNJ11)*	rs5219	(Mao *et al.*, 2012, Zhang *et al.*, 2013)
T2D-associated gene	Transcription Factor 7-Like 2 *(TCF7L2*)	rs7903146 rs12255372 rs7901695	(Chang *et al.*, 2017, Rosta *et al.*, 2017, Wu *et al.*, 2016, Zhang *et al.*, 2013) (Chang *et al.*, 2017, Zhang *et al.*, 2013) (Chang *et al.*, 2017)

2.8 TAGSNP SELECTION

The SNP data of the selected genes were extracted for the Zulu population from the AGVP dataset (Gurdasani *et al.*, 2015) with assistance from Dr Ananyo Choudhury, at SBIMB at the

University of the Witwatersrand, Johannesburg. Data from the AGVP were obtained from the European Genome-phenome Archive (accession number EGAS00001000363), with permission granted to the SBIMB for use. The GRCh37/hg19 coordinates for the genes were used for data retrieval, and was obtained from the following sources: Ensembl (GRCh37.p13 release 75 - August 2017), National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/gene), University of California Santa Cruz (UCSC) Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly (https://genome.ucsc.edu); all accessed on September 2017. Because the coordinates varied slightly between the sources, the coordinates that spanned all three sources were used for data retrieval (Table 2.5).

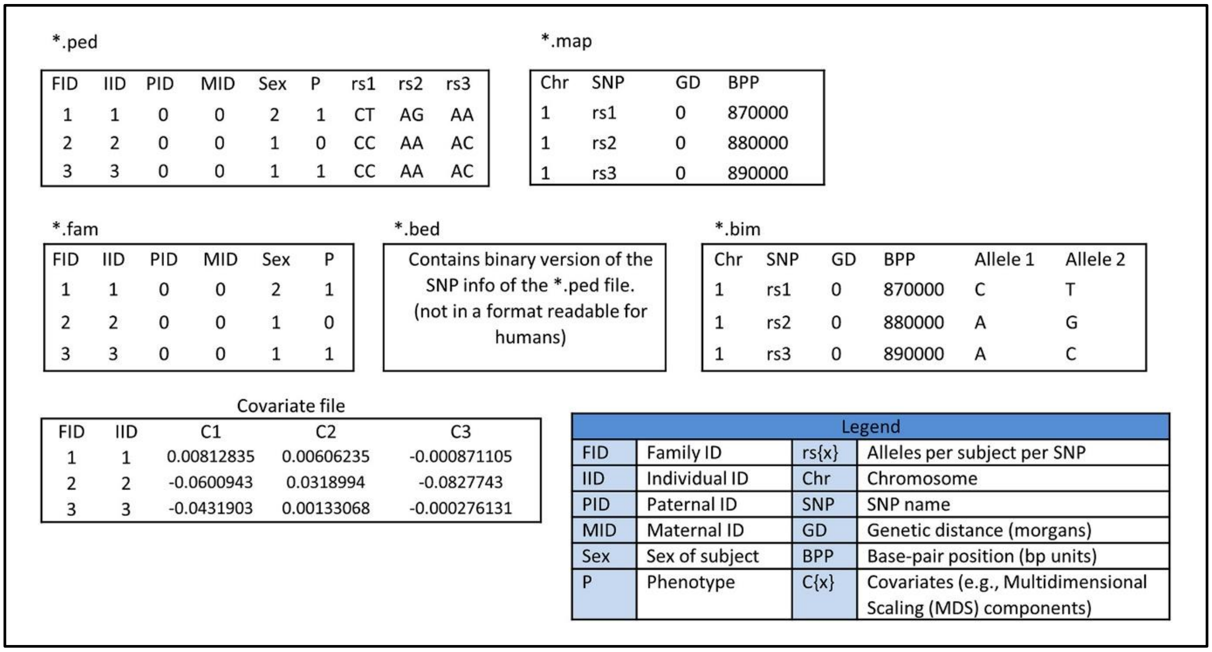
Table 2.5: GRCh37/hg19 coordinates of selected genes

Gene	NCBI	UCSC Genome Browser	Ensembl	Altogether
HNF4A	20: 42 984 441 - 43 061 485	20: 42 984 441 - 43 061 485	20: 42 984 340 - 43 061 485	20: 42 984 340 - 43 061 485
GCK	7: 44 183 870 - 44 229 022	7: 44 182 812 - 44 229 038	7: 44 183 872 - 44 237 769	7: 44 183 870 - 44 237 769
HNF1A	12: 121 415 861 - 121 440 315	12: 121 416 371 - 121 440 314	12: 121 416 346 - 121 440 315	12: 121 416 346 - 121 440 315
PDX1	13: 28 494 168 - 285 00 451	13: 28 494 168 - 28 500 451	13: 28 494 157 - 28 500 368	13: 28 494 157 - 28 500 451
KCNJ11	11: 17 406 795 - 17 410 878	11: 17 406 796 - 17 410 878	11: 17 407 406 - 17 410 878	11: 17 406 795 - 17 410 878
TCF7L2	10: 114 709 978 - 114 927 437	10: 114 886 392 - 114 927 436	10: 114 710 009 - 114 927 437	10: 114 709 978 - 114 927 437

NCBI = National Center for Biotechnology Information University of California Santa Cruz (UCSC), coordinates are provided as chromosome number: starting nucleotide to ending nucleotide within the GRCh37/hg19 genome assembly

The SNP data were originally packaged into individual/sample (.ped) and SNP information (.map) file formats. The individual information (.ped) file contained columns with the following information: Family ID, sample ID, paternal ID, maternal ID, sex (1 = male; 2 = female; other = unknown), disease phenotype (0 = GDM negative; 1 = GDM positive), and genotypes (2 for each marker; 0 = missing). Included in the SNP information (.map) file was the chromosomes on which the SNPs are located, the marker IDs, genetic distances and physical positions. To condense SNP information into appropriately sized files for storage and use in PLINK, standard

PLINK files (.ped and .map) were converted to binary files (.bed, .bim and .fam). The .bed file contained the genotype calls for each individual (from the .ped file), whereas the .fam file contained the first six columns of the .ped file (as stated above). The .bim file contained all the SNP information (obtained from the .map file), including the two SNP alleles (Clarke *et al.*, 2011). Figure 2.5 shows examples of the format of each file type.



*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend

FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

Figure 2.5: Examples of plink files illustrating the format and the type of information included (Marees *et al.*, 2018).

For file conversion of the SNP dataset for each selected gene, PLINK 1.9 and a graphic user interface for common PINK operations, called gPLINK (http://zzz.bwh.harvard.edu/plink/gplink.shtml), was used in conjunction with PLINK command-line. In gPLINK, the “generate fileset” tool under the data management tab was used to perform this function. The exact method was followed for the preparation of files for genotype data generated in this study. In addition to data management, functions such as summary statistics and association analysis from PLINK was also used, as described in Chapter 2, Section 11 Data analysis (Purcell and Chang, 2017, Purcell *et al.*, 2007).

In further preparation of the dataset, SNPs with more than two alleles were excluded from the SNP data. This was done by sorting and exporting the multi-allelic information from the .bim file to a new text file using Excel. Two or more alleles in any of the two allele columns (Allele 1 and

Allele 2) per SNP (or per row) in the .bim file indicated that SNP had multiple alleles; the columns had to be sorted separately. In PLINK, a fileset was generated, excluding the multiallelic SNP and its information compiled in the text file by using the following command:

```
PLINK –bfile <BINARY INPUT FILENAME>--exclude <MULTIALLELIC SNPS
FILENAME> --make-bed –out <OUTPUT FILENAME>
```

Similar to the exclusion of multiple allelic SNPs, SNPs falling outside of a specific selected region within each gene were also excluded. Only the CEU (Northern Europeans from Utah, a European population included in the 1000 Genomes Project) LD-block, containing the GDM-associated SNP, were used as input for tagSNP selection (Table 2.6) to ultimately minimise the amount of tagSNPs captured. Positional coordinates of the LD-blocks were retrieved using the location-based displays for linkage data, specific for the CEU European population, on Ensembl (GRCh37.p13 release 90 - August 2017; accessed on: November 2017). The whole gene of *PDX1* and *KCNJ11* were used for capturing tagSNPs, as they were relatively small genes, in comparison to the others.

Table 2.6: CEU LD Blocks, containing significant GDM-associated SNP, for capturing TagSNPs

Gene	GDM-associated SNP	CEU LD Block Location	Size
HNF4A	rs4812829	20:42988767-42989767	1 kb
GCK	rs1799884 rs4607517	7:44228636-442293207 7:44235536-44235668	684 kb 133 bp
HNF1A	rs1169288	12:121416650-121416988	339 bp
PDX1	No significant associated SNP		6.21 kb
KCNJ11	rs5219		4.08 kb
TCF7L2	rs7903146 rs12255372 rs7901695	10:114758349-114758779 10:114808566-114809149 10:114753800-114754088	431 bp 584 bp 289 bp

*LD Block Location coordinates – chromosome: start-stop position from GRCh37-release 90

After exclusion of multiallelic SNPs and SNPs that fell outside the CEU LD-block region, the binary files were converted to Haploview/Linkage format in the form of a data (.ped) and a locus information file (.info). In gPLINK, the files were generated under data management and selecting the Haploview format option as an output.

In Haploview (Barrett, 2009)**,** the tagger algorithm was used to assesses the level of LD within the SNP dataset, and to choose the minimum amount of SNPs that adequately tags all the genetic information within the region of interest. The parameters used for tagSNP selection was a MAF cut-off of 5%, and SNPs only in complete LD with other SNPs ($r^2 = 1$). The output data produced was a list of positional coordinates of the tagSNPs and the SNPs that each tagSNP captures. To retrieve the SNPs corresponding rs numbers (the universal SNP identification tag assigned by NCBI), a search was conducted on Ensembl in the variant table under genetic variation option on gene-based displays. Table 2.7 shows rs numbers and positional coordinates of the tagSNPs and the SNPs captured by this study.

Table 2.7: The rs numbers and positional coordinates of the TagSNPs selected and the SNPs accounted for by each TagSNP

	TagSNPs		**SNPs Captured**	
	rs numer	Location	rs numer	Location
	HNF4A			
1	rs80276513	20:42989218	rs80276513	20:42989218
2	rs6031551	20:42989714	rs6031551	20:42989714
	GCK			
3	rs112257899	7:44229079	rs112257899, rs111560203	7:44229079,7:44235544
4	rs4607517	7:44235668	rs4607517	7:44235668
5	rs758983	7:44235536	rs758983	7:44235536
6	rs1799884	7:44229068	rs1799884	7:44229068
	HNF1A			
7	rs2244608	12:121416988	rs2244608	12:121416988
	PDX1f			
8	rs61944006	13:28496419	rs61944006, rs57247118, rs60353775, rs4002828	13:28496419, 13:28497828, 13:28498265, 13:28496119
9	rs73169687	13:28498102	rs73169687, rs75034644	13:28498102, 13:28498325
10	rs7981781	13:28499962	rs7981781, rs4430606	13:28499962, 13:28495193
11	rs4581569	13:28497621	rs4581569, rs9512918	13:28497621,13:28494949
12	rs9554205	13:28499741	rs9554205	13:28499741
13	rs4415872	13:28497159	rs4415872	13:28497159
	KCNJ11			
14	rs5210	11:17408251	rs5210, rs2285676, rs5222	11:17408251, 11:17408025, 11:17410283
15	rs5214	11:17408550	rs5214, rs112070496	11:17408550, 11:17409531
16	rs5215	11:17408630	rs5215	11:17408630
	TCF7L2			
17	rs34872471	10:114754071	rs34872471	10:114754071
18	rs7903146	10:114758349	rs7903146	10:114758349
19	rs7901695	10:114754088	rs7901695	10:114754088
20	rs34347733	10:114753800	rs34347733	10:114753800
21	rs12255372	10:114808902	rs12255372	10:114808902
22	rs115626858	10:114758520	rs115626858	10:114758520
23	rs115758892	10:114808835	rs11575889f2	10:114808835

2.9 MASSARRAY GENOTYPING

Genotyping was performed using the MassARRAY® System by Agena Bioscience (San Diego, CA). Genotyping and training on this system are provided as a commercial service by Inqaba Biotec (Pretoria, South Africa). MassARRAY determines the alleles based on the mass of the generated allele-specific products. The technology allows for multiplexing of up to 40 SNPs due to the use of a wide mass spectrum range. This technology employs an initial amplification of the target region containing the SNP of interest, followed by SBE for allele discrimination, and MALDI-TOF for allele detection (Gabriel *et al.*, 2009, Nakai *et al.*, 2002, Storm *et al.*, 2003).

The MassARRAY® Typer 4.0.20 Software (Agena Bioscience, San Diego, CA) was used to assess if all the selected SNPs could be genotyped in a single multiplex reaction through an *in silico* test. This included designing primers (two for PCR, and one for extension) and assessing them for specificity to the target region and for potential interference with one another (Ellis and Ong, 2017). When the assay design was finalized, information was generated on the unique mass that could be expected for each extension primer and the specific nucleotide.

After confirming SNPs that could be genotyped, the process followed as illustrated in Figure 2.6. Firstly, the target regions were amplified in 5 µl PCR reactions. The PCR consisted of 2.5 ng/µl of genomic DNA, 100 nmol of each amplification primer, 500 µM dNTPs, 2 mM $MgCl^2$, and 1 U of HotStarTaq Plus DNA polymerase. The PCR conditions included initial heating at 94°C for 2 min, 45 cycles of amplification (30 sec at 94°C, 30 sec at 56°C,60 sec at 72°C), and a final extension for 5 min at 72°C. Unincorporated dNTPs were removed from the PCR product, by adding 2µl Shrimp Alkaline Phosphatase (SAP, 0.24X SAP buffer and 0.51 U SAP enzyme) and subjecting it to 37˚C for 40 min, and then 85˚C for 5 min. Single base pair extension was performed on the SAP-treated plate to incorporate a single mass-modified nucleotide at the polymorphic region, immediately upstream of the designed extension primer. The extension reaction was made up to a total volume of 9 µl, containing 1X iPlex Buffer Plus, iPlex dNTPs, and iPlex polymerase, 1150 nM of the primer pool and the entire cleaned PCR product. Single base extension was facilitated by five cycles of the following conditions; denaturation at 95˚C for 5 sec, annealing for 52˚C for 3 sec, extension at 80˚C for 3 sec. Initial heating was done at 94˚C for 30 sec, and the final extensions step at 72˚C for 3 min (single cycle). The extension reaction products were treated with a cation exchange resin to remove salts (Na^+, K^+, and Mg^{2+} ions) that can result in high background noise in the mass spectra. Using a nanoliter dispenser, 25 µl of analytes were transferred to the MassARRAY chip, containing matrix that assists in desorption and ionization. At the final step, the chip was loaded onto the mass spectrometer, and detection and analysis were performed by using the real-time detection software (Gabriel *et al.*, 2009).

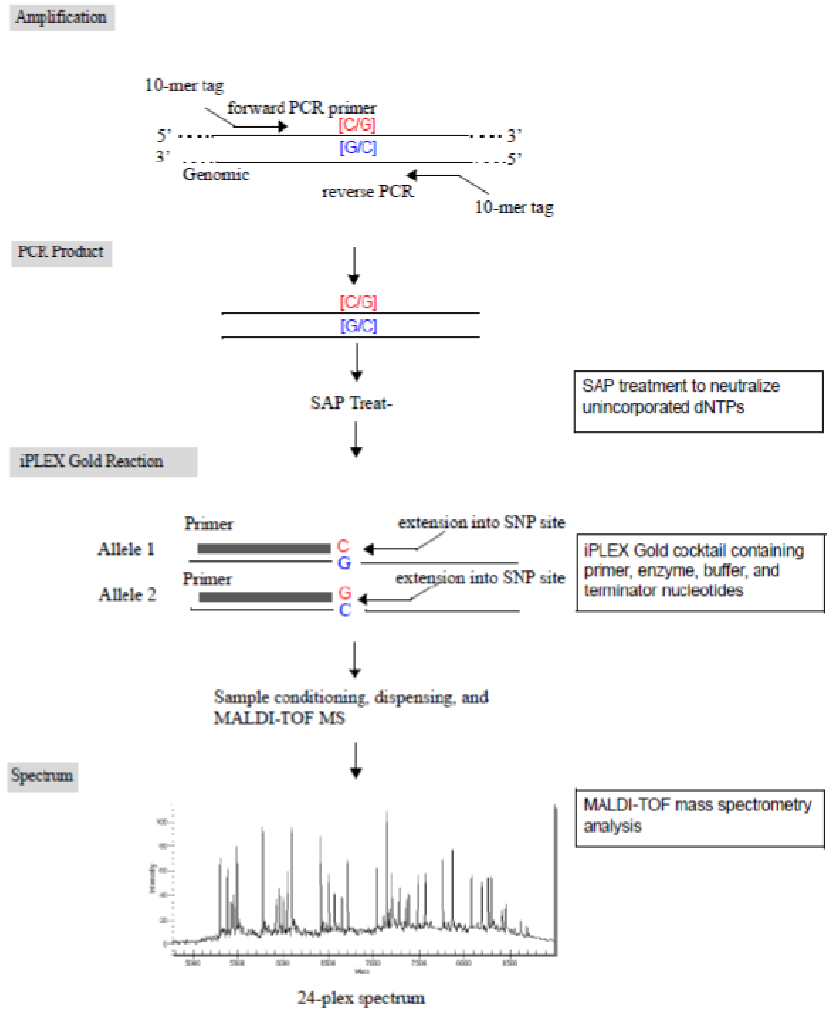


Figure 2.6: Process of MassARRAY genotyping. The target region was amplified, and the unincorporated nucleotides were removed from the PCR product by Shrimp Alkaline Phosphatase (SAP). Subsequently, single base extension (SBE) extended a primer with one nucleotide.

The steps of MALDI-TOF involve vaporization and ionization of the sample, which is triggered by a short laser pulse. The matrix assisted sample molecules in vacuum are then electrostatically transferred, allowing for the separation of the sample molecules from the matrix ions, and acceleration towards the detector. The time taken for the sample ions to reach the detector (time-of-flight), are proportional to the square root of the molecules mass-to charge (m/z) ratio, which were the units recorded by the analysis software. By design, the mass of each allele was expected to occupy a unique position within the mass-spectrum.

2.10 TAQMAN GENOTYPING

The SNPs that were not compatible with the MassARRAY protocol were genotyped by using Applied Biosystems® TaqMan SNP genotyping assays. TaqMan is based on the complementary hybridization of fluorescent probes to the target sequence, and the exonuclease activity of the Taq polymerase enzyme. The technology uses a probe that hybridizes to a specific SNP allele and emits fluorescence upon amplification of the region when released from the quencher molecule (Holland *et al.*, 1991, Livak *et al.*, 1995). Different colours of fluorescent dyes can be used to label the probe, which permits simultaneous detection of different alleles and SNPs.

The assay mixture consisted of PCR primers, and SNP-specific probes labeled with FAM™ or VIC® dye, that were diluted to 20X from the 40X stock concentration before use. Reactions with a total volume of 5 µl were set-up according to TaqMan® SNP Genotyping Assay’s User Guide for single-tube assays, and consisted of 1-20 ng/µl DNA,TaqMan SNP Assay, and Taqpath ProAmp Mastermix. End-point PCR and fluorescence detection was done by using the Quantudio® 5 Real-Time PCR System (Applied Biosystems®). By comparing the total fluorescence emitted by each probe, genotypic calls were made by QuantStudio Design & Analysis Software. Analysis was conducted by using the QuantStudio Design & Analysis Software that generated real time amplification plots, allelic discrimination cluster plots, and quality control summaries. For verification of genotypes, at least 10% of the samples from the first two plates were re-genotyped on the last plate. Each assay had one sample that did not cluster well with the three genotypic groups (n = 4), and were therefore also added in addition to the 10% of samples to the last plate for re-genotyping. The result that fell in a distinct cluster, were taken as the true genotype call. For each plate a non-template control was included as a negative control.

2.11 DATA ANALYSIS

Data quality control and association testing was carried out using PLINK v1.9 (Purcell and Chang, 2017, Purcell *et al.*, 2007), and gPLINK. Genotype data generated was originally compiled into the format of .ped and .map files (as described in Chapter 2, 8. TagSNP selection). Thereafter, SNPs were converted into the appropriate binary. bed, .bim, and .fam files, while also excluding SNPs by applying certain data quality thresholds, as discussed in 11.1 Genotype data quality control. The “Data Management”, “Summary Statistics” and “Association” tools in gPLINK were primarily used for data analysis.

2.11.1 GENOTYPE DATA QUALITY CONTROL

Summary statistics were performed to calculate the genotype failure per SNP and individual, to calculate the MAF of each SNP, and to test for Hardy-Weinberg equilibrium (HWE) for all SNPs. Based on this, per-SNP quality control was performed to avoid false positive results due to poor genotyping of SNPs and systemic genotype errors (Namipashaki *et al.*, 2015). The SNPs that failed to be genotyped in more than 5% of the samples and SNPs that were not in HWE ($p <$ 0.01), were excluded from the analysis. HWE testing was only applied to control samples, as deviation from HWE in cases may indicate selection, an event that is possible for SNPs that have an effect on the phenotype of interest (Anderson *et al.*, 2010). Homogeneity, the accuracy of the genetic representation of the population within the cohort, and therefore also population substructure, has been partly addressed by ensuring that the observed genotypic frequencies of selected SNPs are in accordance with HWE predictions (Namipashaki *et al.*, 2015). Furthermore, SNPs that had a MAF < 0.05 were also excluded, as to avoid for the possibility of false positive findings due to the low frequency and a relatively small sample size. Finally, the allelic discrimination cluster plots were manually inspected to verify the accuracy of the genotypes identified. A plot clustering well within a cluster was regarded as an accurate call, and one that fell outside a distinct cluster was regarded as an incorrectly identified genotype.

2.11.2 ASSOCIATION TESTING

The association of the minor allele of each SNP with GDM was assessed by using logistic regression analysis. Logistic regression analysis was also used to assess the genotypic association of individual SNPs (in reference to the minor allele) with GDM. The association analysis between SNPs and fasting glucose levels was performed by using linear regression. Since the core assumptions of linear regression is normally distributed residuals, and the fasting glucose measurements were not significantly normally distributed, the data was log-transformed before proceeding with the analysis (Fusi *et al.*, 2014).

The files included for the analysis were covariate, .cov files (containing phenotypic continuous variables of age and BMI) together with the binary PLINK files (.bed, .bim and .fam). Permutation testing for 1000 permutations were performed to empirically generate pointwise *p* values (*EMP1*) for an individual SNP, and *p* values (EMP2) that are corrected for multiple testing.

Logistic regression analysis was applied as it is the method used to evaluate the relationship between one or more independent variables (SNPs and/or confounders that are either continuous or categorical), with a dichotomous dependent variable (such as GDM positive women and controls). In contrast linear regression is used when the dependent variable is continuous (such as fasting glucose), and when there is also more than one independent variable. These statistical methods are predictive as they can estimate the odds or probability of the outcome, based on the genotype data modelled under the logistic or the linear function (Bush and Moore, 2012, Walker and Duncan, 1967). In comparison to the univariate Chi-square based tests, regression analyses has the advantage of incorporating covariates into the analysis in an additive (logistic/ linear) manner, making it a method by which one can control for numerous confounders (Beyene and Pare, 2013, Clarke *et al.*, 2011, Pourhoseingholi *et al.*, 2012).

The OR and the beta coefficient, numerical values generated by logistic and linear regression, respectively represented the direction and the magnitude of the effect the allele has on the phenotype. The beta regression coefficient is defined as the log odds of the outcome associated with a one-unit increase in the exposure (Szumilas, 2010). The OR/ regression coefficient thus represents the relationship of the two variables, and therefore predict the phenotype outcome based on the increase in allele or genotype. The OR is the odds of having GDM when exposed to a variable (minor allele) in comparison to the odds of the occurrence without the exposure variable (the alternative or major allele). An OR=1 indicates that there is no difference in the odds of having GDM between the groups of women carrying two different SNP alleles. An OR > 1 indicates the minor allele is associated with higher odds of having GDM. For the opposite, OR <1.0, indicates a decrease in the odds of having GDM when carrying the minor allele (Ranganathan *et al.*, 2015, Szumilas, 2010).

Permutation is a re-sampling approach used to create new datasets, assumed to be under the null hypothesis by randomly rearranging labels (i.e. breaking up phenotype-genotype relationships) of the non-permuted study data. The number of datasets, as specified by the number of permutations, are then used for testing against the observed test statistic, thereby generating significance levels empirically. The empirical significance level is estimated as the proportion of randomization samples with a test statistic at least as large as the one observed. Labels for data are interchangeable under the null hypothesis of no association. A significant p value ($p < 0.5$) would therefore indicate that label is accordant with the data showing an association and that it is not interchangeable (Bush and Moore, 2012, Knijnenburg *et al.*, 2009, Marees *et al.*, 2018).

For permutation, the resolution of obtainable p values is determined by the number of permutations performed ($1/N$). A number of a 1000 permutations was chosen, as the smallest achievable p value would be 0.001 for rejecting the null hypothesis (Knijnenburg *et al.*, 2009). Due to this procedure being computationally intensive, permutation is preferred for studies of smaller sample sizes such as this one (Clarke *et al.*, 2011).

Multiple testing should be corrected for in an association testing, since more than one comparison is conducted. For a single test, the probability of detecting a false positive is usually set to 0.05. In other words, five percent of the time the null hypothesis of no association is rejected when it is true (Type I error). As the number of tests increases, the cumulative probability of a Type I error also increases (Beyene and Pare, 2013). In this study, permutation was the method used to correct for multiple testing, as it is less stringent than the Bonferonni correction.

2.12 SNP CHARACTERIZATION

The functional significance of the SNPs found to be associated with GDM were explored by using two online bioinformatic tools: Variant Effect Predictor (VEP) from Ensembl (McLaren *et al.*, 2016), and RegulomeDB (Boyle *et al.*, 2012). These tools asses and predict the functional effect of the variants based on multiple input reference datasets from various sources. The rs IDs of SNPs are recognized on both platforms and were thus used as input to retrieve the information specific to the SNPs of interest.

2.12.1 VARIANT EFFECT PREDICTOR

Variant Effect Predictor can be used for most types of genomic variation in the coding and non-coding regions of the genome, and provide actively curated information on the function, mutation type, location and clinical significance. Variant Effect Predictor annotate variants based on their positions within the genome, and thus, VEP may provide multiple annotations for a single variant that overlap more than one gene, transcript, or genomic feature. Variant Effect Predictor classifies variants according to consequence or variant types, and provide the predefined impact ranked according to the specific consequent type (i.e. low, moderate and high). Variant consequences are described by a standardized set of annotation terms as defined by sequence ontology (Figure 2.7). These terms help to describe the features related to the sequences taking part in biological processes. To enhance prediction capabilities, the VEP web interface can be configured to incorporate several pathogenicity prediction programmes. In this study, Combined Annotation Dependent Depletion (CADD) v1.4 (Kircher *et al.*, 2014, Rentzsch *et al.*, 2019) was used to annotate the coding and noncoding variants found to be associated with GDM or fasting glucose levels in this study.

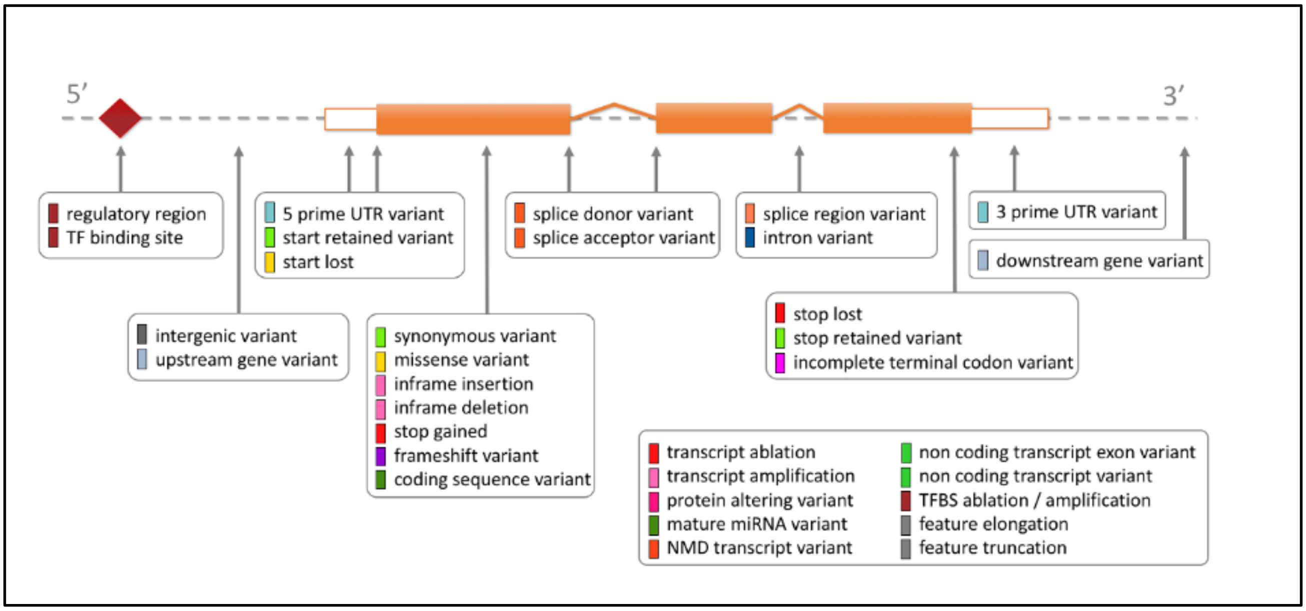


Figure 2.7: A diagram showing the set of consequence terms for variants located in certain gene regions (obtained from https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html).

2.12.2 COMBINED ANNOTATION DEPENDENT DEPLETION

Combined Annotation Dependent Depletion is a scoring tool used to measure the deleteriousness of SNPs and small insertions or deletions by integrating diverse genome annotations. The framework is based on identifying differences between observed human derived changes, that are assumed to be fixed or nearly fixed, and simulated variants, serving as controls. The “fixed” changes in the human genome were identified by comparing the human genome to an ancestral genome constructed from human and chimpanzee sequences. Variation in regions deemed to be fixed or nearly fixed due to selective constraint is considered to be more functional in comparison to variation in regions that have not been conserved throughout evolution. Therefore, a deleterious SNP will likely be amongst the simulated variation rather than the “fixed” variation. By applying various annotations methods and integrating information from regulatory datasets, such as ENCODE, all possible variants are assigned a raw C-score. For interpretation purposes, a C-score of each variant is further ranked relative to all other possible variants and are called phred-like C scores. Variants scoring a value of 10 or greater are within the top 10% of all scores, meaning that they are least likely to be observed in humans under this model and thus amongst the variants that are the most deleterious. C score values of 20 or greater, are the top 1%, and C scores of 30 and greater are within the top 0.1%. The higher the C score value, the greater the functional impact of the variant. Table 2.8 shows the prediction score interpretation for CADD.

Table 2.8: CADD score prediction and interpretation

Score	Prediction	Interpretation
30 < x < 40	Top 0.1% most deleterious	Greatest functional impact
20 < x < 20	Top 1% most deleterious	Greater functional impact
10 < x < 20	Top 10% most deleterious	Functional impact
x < 10	Irrelevant score	Unpredictive range 25% - 75%

2.12.3 REGULOMEDB

To assist in the functional consequence prediction of non-coding variants, RegulomeDB was used in addition to the other bioinformatic tools. RegulomeDB is a database that annotates variants based on the known and predicted regulatory elements in the intergenic regions of the genome. Multiple data sources are used to identify regions containing signatures/elements indicating a potential function in regulation. These signatures include sites with DNAse hypersensitivity, transcription factor (TF) binding sites, and promoter regions with biochemical evidence of regulation. A heuristic scoring system has also been developed that represent the confidence that the variant of interest lies within a regulator region that could have a functional impact when altered. The top ranked variants in Category 1 are known expression quantitative trait loci (eQTLs) for genes, and thus have been found to be associated with expression. Other unknown eQTLs not associated with regulation are classified in Category 6. Categories are also further subcategorized according to the confidence in regulatory information as shown in Table 2.9.

Table 2.9: RegulomeDB variant classification scheme (Boyle *et al.*, 2012)

Category Scheme	
Category	Description
	Likely to affect binding and linked to expression of gene target
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak
	Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
	Less likely to affect binding
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
	Minimal binding evidence
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

2.13 POPULATION ALLELE FREQUENCY COMPARISON

PLINK was used to calculate MAF data for each SNP from the generated genotype data. This was done by using the summary statistics function in gPLINK. The .ped and .map files previously formatted for genotype quality assessment were used as input. The allele frequency data was then compared to other publicly available allele frequency data available from the 1000 Genomes Project (Abecasis *et al.*, 2010**)**. Allele frequency data were retrieved from Ensembl (GRCh38.p12 release 95 - Jan 2019) for the European CEU population and the following African populations: Gambian (GWD) in Western Divisions in the Gambia, the Mende (MSL) in Sierra Leone, Yoruba (YRI) and Esan (ESN) in Nigeria, and the Luhya (LWK) in Webuye, Kenya, as shown in Figure 2.8. The Chi-square test and the Fisher’s exact test were used to assess if the allele frequencies were significantly different using R Project for statistical computing v.3.5.3 (R Core Team,3.5.3), and RStudio (RStudio Team, 2015). The Fisher’s exact test was used in the case of rare variants, where the allele count was less than five.

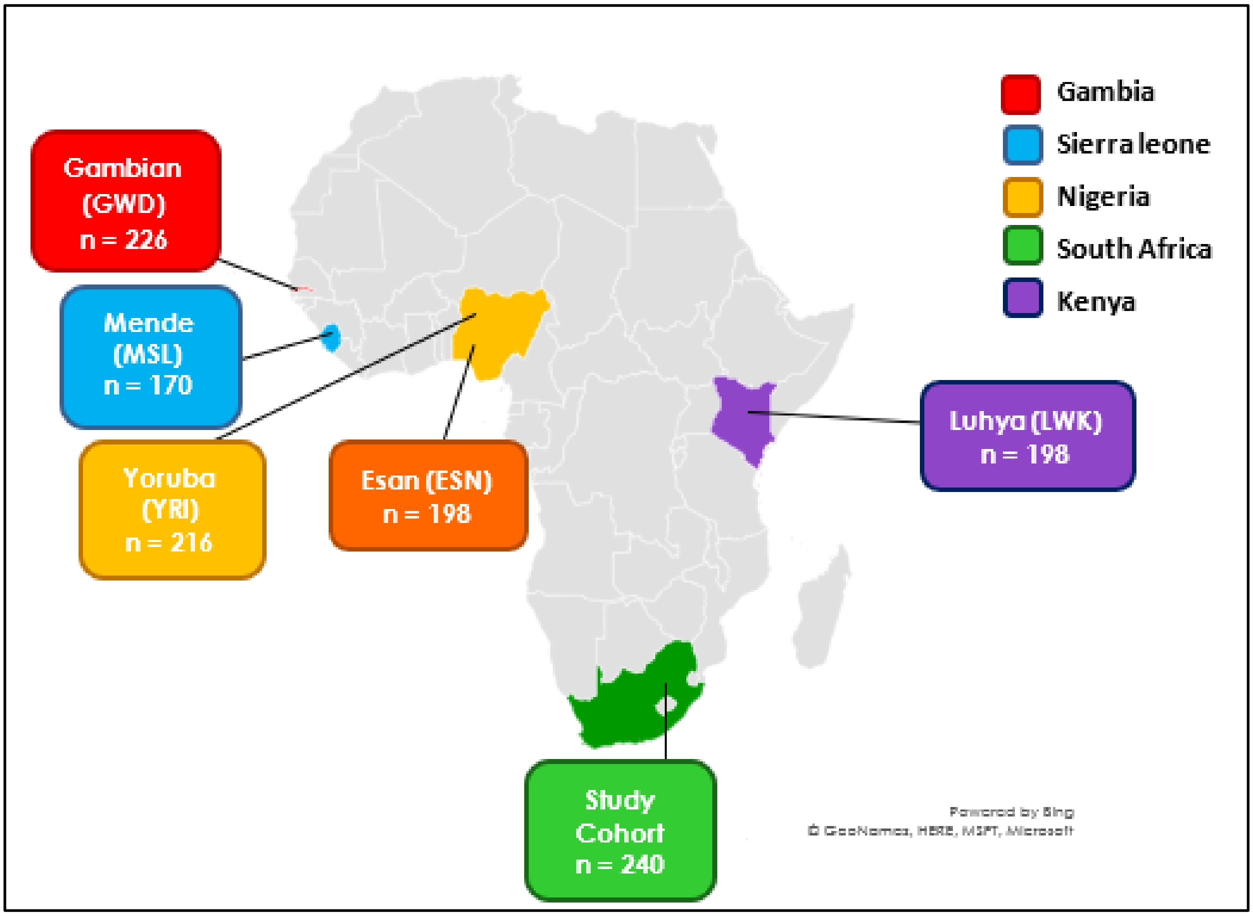


Figure 2.8: Geographic origin of African populations used in this study for allele frequency comparison

3 RESULTS

In this chapter, a summary of the characteristics of the cohort are provided, as well as, a description of the selected SNPs that were genotyped, and the SNPs excluded from the association analyses. In the following sections, the summary statistics of the assessed SNPs, and results of the association analyses, are presented. The association analysis section includes the results for logistic regression analyses conducted to assess the allele and genotype associations with GDM, and the linear regression analysis conducted to assess the association of each SNP with fasting glucose levels. Lastly, the information obtained on the SNPs found to be associated with GDM and fasting glucose levels, are presented, including their annotated or predicted functional impact, and the allele frequencies of the tagSNP within other populations.

3.1 COHORT CHARACTERIZATION

Table 3.1 summarizes the phenotypic variables for the whole cohort, as well as for the cases and controls separately. The continuous phenotypic variables (age, BMI, fasting glucose, and household asset score) were not normally distributed, and were therefore presented as medians and IQRs. The categorical variables were described in terms of frequencies and percentages. There were no significant differences between the cases and controls for any of the characteristics, except for fasting glucose. As expected, most women with GDM had a significantly higher fasting glucose concentration in comparison to women without GDM. A high BMI was seen for the whole cohort with most women (42.9%) falling into the obese category.

Table 3.1: Participant characteristics

Characteristics	GDM Cohort (n=240)	Women with GDM (n=80)	Women without GDM (n =160)	*p* value
Age (years)	31.0 (27.0 - 36.0)	31.0 (27.0 - 36.0)	31.0 (27.0 - 36.0)	0.989
BMI (kg/m^2)	28.6 (25.0 - 32.4)	29.1 (25.4 - 34.2)	28.1 (24.3 - 32.0)	0.088
Fasting glucose (mmol/L)	**4.4 (3.9 - 5.1)**	**5.2 (5.1 - 5.5)**	**4.1 (3.7 - 4.5)**	**< 2.2 x 10^{-16}**
Household asset score	5.0 (5.0 - 6.0)	6.0 (5.0 - 6.3)	5.0 (5.0 - 6.0)	0.129
BMI categories				
Normal (≤24.9 kg/m^2)	60 (25.0%)	17 (21.3%)	43 (26.9%)	0.609
Overweight (25-29.9kg/m^2)	77 (32.1%)	26 (32.5%)	51 (31.9%)	
Obese (≥ 30 kg/m^2)	103 (42.9%)	37 (46.3%)	66 (41.3%)	
Previous pregnancies				
None	8 (3.3%)	3 (3.8%)	5 (3.1%)	0.597
One to two	152 (63.3%)	47 (58.8%)	105 (65.6%)	
Three or more	80 (33.3%)	30 (37.5%)	50 (31.3%)	
Education				
No schooling/ primary school	4 (1.7%)	3 (3.8%)	1 (0.6%)	0.226
Secondary school	183 (76.3%)	60 (75.0%)	123 (76.9%)	
Tertiary education	53 (22.1%)	17 (21.3%)	36 (22.5%)	

3.2 SNPS SELECTED FOR GENOTYPING

Due to the criteria of selecting tagSNPs, only seven tagSNPs were selected. An additional 16 SNPs (that could not be tagged by a proxy SNP) were also included and were selected for genotyping to ensure coverage of the CEU block region containing the GDM-associated SNP. The following five SNPs had previously been found to be significantly associated with GDM and were also assessed in this study: rs1799884 and rs4607517 within *GCK*; and rs7901695, rs7903146, and rs12255372 within *TCF7L2* (indicated in bold in Table 3.2).

Table 3.2: List of SNPs and tagSNPs selected for genotyping

Captured SNPs	tagSNPs
HNF4A rs80276513	*GCK* rs112257899
HNF4A rs6031551	*PDX1* rs73169687
***GCK* rs4607517**	*PDX1* rs7981781
GCK rs758983	*PDX1* rs4581569
***GCK* rs1799884**	*PDX1* rs61944006
HNF1A rs2244608	*KCNJ11* rs5210
PDX1 rs9554205	*KCNJ11* rs5214
PDX1 rs4415872	
KCNJ11 rs5215	
TCF7L2 rs34872471	
TCF7L2 rs7903146	
TCF7L2 rs7901695	
TCF7L2 rs34347733	
TCF7L2 rs12255372	
TCF7L2 rs115626858	
TCF7L2 rs115758892	

3.3 QUALITY CONTROL

The quality of the genotyped data per SNP was assessed based on their genotype success in individuals, HWE, and MAF. Five SNPs (rs758983, rs34347733, rs7981781, rs73169687, and rs4415872) were removed from analyses because they failed to be genotyped in more than 5% of the samples (> 0.5 failure rate). A sixth SNP (*TCF7L2* rs7903146), was removed because it was not in HWE when looking only at the HWE *p* value in the controls. The SNPs that were excluded were all genotyped using Mass-ARRAY System by Agena Bioscience (San Diego, CA). All SNPs had a MAF > 0.01 and were thus not excluded based on the MAF cut-off of 0.05. Table 3.3 shows the SNPs excluded due to a high genotype failure rate in the participants, and having a HWE *p* value < 0.01.

The 10% of SNPs that have been genotyped and re-genotyped using Taqman assays, were verified to be the correct genotype call with no discrepancies in their genotype cluster position between the initial results and repeated results. Of the four repeated samples with initial uninformative results (falling outside of a distinct cluster), three had results that clustered well within a distinct genotype cluster and were regarded as the true genotype call.

Table 3.3: SNPs removed due to failing quality control measures

Gene	SNPs	Number of failed samples (n)	Failure rate	HWE *p* value
GCK	rs758983	14	0.06	1.00
TCF7L2	rs34347733	14	0.06	1.00
	rs7903146 [a]	8	0.03	$6.21e^{-23}$
PDX1	rs7981781	14	0.06	0.03
PDX1	rs73169687	15	0.06	1.00
PDX1	rs4415872	209	0.87	0.63

[a] previously reported GDM-associated SNP, HWE = Hardy-Weinberg Equilibrium

Of the five SNPs that had a > 0.5 failure rate, one SNP (rs4415872) had no results for most of the samples (209/240) and indicated that the assay failed for that particular SNP. As for the other SNPs, the service provider reported that genotyping failed due to poor quality of at least 15 samples with a failure rate of > 0.10 (sample failed for more than two out of 22 SNPs). Table 3.4 shows the number of SNPs with missing data and the DNA quality for each participant sample that could have contributed to genotyping failure. The A260/230 ratios were missing for some samples, whereas A260/280 ratios were available for all the samples.

Table 3.4: Participant samples with a high genotype failure rate (>10%) and their A230/A280 and A260/A230 ratios indicating DNA purity

Sample ID	Number of failed SNPs	Failure rate	260/280	260/230
14-10413	3	0.14	1.94	-
14-10562	3	0.14	1.76	1.93
14-10569	3	0.14	1.76	1.82
14-10203	4	0.18	2.01	-
14-10256	6	0.27	1.72	1.21
3G5913407	6	0.27	1.71	1.52
SFG1301	8	0.36	1.75	1.76
14-10508	10	0.45	1.83	2.17
14-10501	11	0.50	1.80	2.13
14-10487	13	0.59	1.64	2.23
SFG1310	13	0.59	1.83	1.85
14-10329	15	0.68	2.05	-
SFG1240	16	0.73	1.81	2.08
14-10464	18	0.82	1.77	-
SFG1062	18	0.82	1.42	-

A260/A280 ratio of ~1.8 and A260/A230 ratio of ~2 is considered good quality and "pure"

In total, the genotype data of seven of the 23 SNPs originally selected were not tested for association with GDM due to failing quality control and/ or design error. Only 18 SNPs of the initial 23 variants were compatible with the MassARRAY system and thus genotyped using the technology in a single multiplex reaction. Taqman assays were used to genotype the remaining four SNPs; two SNPs by pre-designed assays and two by custom designed assays. One Taqman assay failed to be designed by the manufacturer for the genotyping of rs9554205. Figure 3.1 shows a flow-diagram of the steps leading up to the statistical analyses.

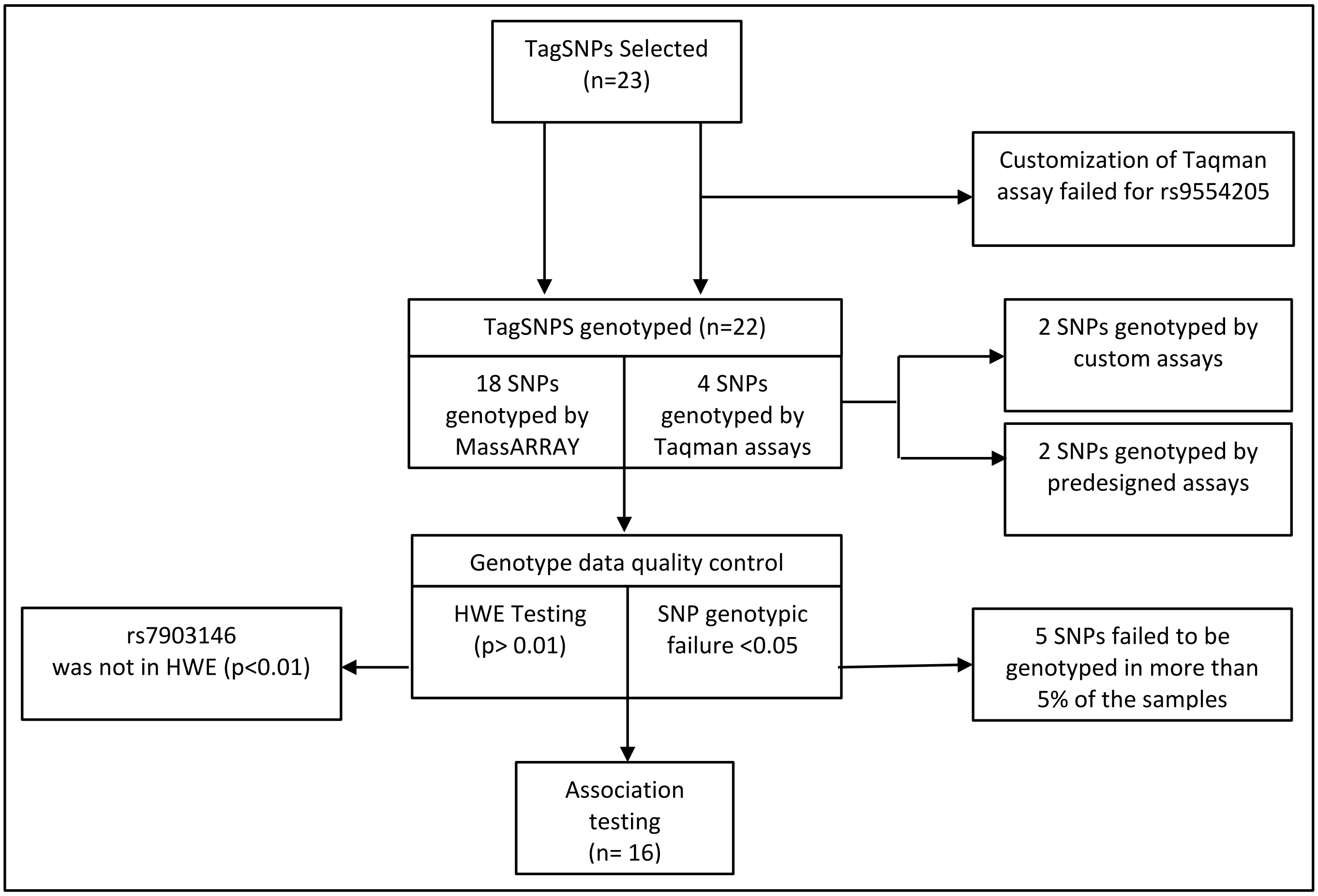

TagSNPs Selected
(n=23)
Customization of Taqman assay failed for rs9554205
TagSNPS genotyped (n=22)
18 SNPs genotyped by MassARRAY
4 SNPs genotyped by Taqman assays
2 SNPs genotyped by custom assays
2 SNPs genotyped by predesigned assays
Genotype data quality control
HWE Testing
(p> 0.01)
SNP genotypic failure <0.05
rs7903146
was not in HWE (p<0.01)
5 SNPs failed to be genotyped in more than 5% of the samples
Association testing
(n= 16)

Figure 3.1: Flow-diagram illustrating the steps before the association analysis

3.4 GENOTYPING

The genotype data for SNPs included for statistical analyses is summarized in Table 3.5. For the complete summary of genotype data, refer to Appendix C.

.

Table 3.5: Summary statistics of the genotype results obtained from the 22 SNPs analysed

					GDM Positive Cases				GDM Negative Controls				
Gene	SNP ID	Failure Rate	No. of failed samples	Minor Allele (A1)	A1/A1 (n)	A1/A2 (n)	A2/A2 (n)	MAF	A1/A1 (n)	A1/A2 (n)	A2/A2 (n)	MAF	HWE
GCK	rs1799884[a]	0.03	7	T	2	28	48	0.21	7	48	100	0.20	0.62
	rs112257899	0.00	0	T	1	15	64	0.11	0	22	138	0.07	1.00
	rs4607517[a]	0.03	8	A	0	6	71	0.04	1	22	132	0.08	1.00
TCF7L2	rs34872471	0.00	1	C	12	39	29	0.39	27	79	53	0.42	0.87
	rs7901695[a]	0.00	0	C	15	49	16	0.49	39	82	39	0.50	0.87
	rs115626858	0.05	12	T	1	14	63	0.10	1	23	126	0.08	1.00
	rs115758892	0.03	7	A	0	9	70	0.06	3	15	136	0.07	0.02
	rs12255372[a]	0.03	7	T	6	28	45	0.25	13	56	85	0.27	0.41
KCNJ11	rs5210	0.05	11	G	11	33	33	0.36	21	70	61	0.37	0.86
	rs5214	0.00	1	C	0	6	74	0.04	1	24	134	0.08	1.00
	rs5215	0.03	6	C	0	10	68	0.06	0	16	140	0.05	1.00
HNF1A	rs2244608	0.05	12	G	0	6	70	0.04	0	23	129	0.08	1.00
PDX1	rs61944006	0.02	4	C	4	40	36	0.30	6	62	88	0.24	0.27
	rs4581569	0.02	5	T	3	27	48	0.21	18	63	76	0.32	0.36
HNF 4A	rs80276513	0.01	3	A	4	6	70	0.09	2	15	140	0.06	0.10
	rs6031551	0.03	7	C	2	21	55	0.16	5	41	109	0.17	0.57

SNP = single nucleotide polymorphism, SNP ID = universal SNP identification tag (rs number) assigned by National Center for Biotechnology Information (NCBI), A1 = minor allele, A2 = major allele, A1/A1, A1/A2, A2/A2 (n) = genotype counts, MAF = Minor Allele Frequency, HWE = Hardy-Weinberg Equilibrium, [a] previously reported GDM-associated SNP

3.5 ASSOCIATION ANALYSIS

When assessing the allelic association of 16 SNPs with GDM risk through logistic regression analysis, only one SNP (*PDX1* rs4581569) was significantly associated with GDM, with a *p* value (*EMP1*) of < 0.05, even after adjusting for covariates. The OR result for the significantly associated SNP was lower than 1, which indicated that this SNP was associated with a lower risk of GDM. However, when the association was adjusted for multiple testing by permutation, the *p* value (EMP2) was no longer significant. Table 3.6 shows the association results for the SNPs tested, with the SNP significantly associated with GDM indicated in bold. As only SNPs and no samples were excluded from the analyses, the number of successfully genotyped samples (n out of a total of 240 participants) for each SNP is also indicated in the table.

Table 3.6: Allelic association results for SNPs obtained via logistic regression analysis

			Unadjusted for BMI & Age			Adjusted for BMI & Age		
Gene	SNP ID	n	OR [95% CI]	*EMP1*	*EMP2*	OR [95% CI]	*EMP1*	*EMP2*
GCK	rs1799884[a]	233	1.03 [0.63-1.67]	0.878	1.000	1.02 [0.62 - 0.67]	0.935	1.000
	rs112257899	240	1.64 [0.83-3.24]	0.110	0.964	1.71 [0.86 - 3.43]	0.104	0.911
	rs4607517[a]	232	0.48 [0.19-1.21]	0.130	0.917	0.48 [0.19 - 1.21]	0.127	0.901
TCF7L2	rs34872471	239	0.90 [0.61-1.33]	0.609	1.000	0.92 [0.62 - 1.39]	0.709	1.000
	rs7901695[a]	240	0.97 [0.65-1.45]	0.862	1.000	1.00 [0.67 - 1.51]	0.983	1.000
	rs115626858	228	1.26 [0.65-2.42]	0.465	1.000	1.14 [0.58 - 2.23]	0.706	1.000
	rs115758892	233	0.85 [0.40-1.80]	0.644	1.000	0.75 [0.34 - 1.63]	0.456	1.000
	rs12255372[a]	233	0.94 [0.61-1.44]	0.784	1.000	0.92 [0.60 - 1.43]	0.719	1.000
KCNJ11	rs5210	229	0.95 [0.64-1.42]	0.806	1.000	0.94 [0.63 - 1.41]	0.777	1.000
	rs5214	239	0.43 [0.17-1.09]	0.062	0.756	0.44 [0.17 - 1.12]	0.088	0.813
	rs5215	234	1.29 [0.55-2.99]	0.582	1.000	1.36 [0.57 - 3.25]	0.461	1.000
HNF1A	rs2244608	228	0.48 [0.19-1.24]	0.134	0.929	0.48 [0.18 - 1.25]	0.115	0.921
PDX1	rs61944006	236	1.45 [0.91-2.30]	0.118	0.909	1.34 [0.84 - 2.15]	0.209	0.993
	rs4581569	**235**	**0.59 [0.38-0.93]**	**0.015**	**0.32**	**0.62 [0.40 - 0.98]**	**0.031**	**0.505**
HNF4A	rs80276513	237	1.35 [0.72-2.53]	0.402	0.999	1.31 [0.70 - 2.48]	0.418	1.000
	rs6031551	233	0.97 [0.58-1.63]	0.915	1.000	0.99 [0.59 - 1.67]	0.977	1.000

SNP = single nucleotide polymorphism, SNP ID = universal SNP identification tag (rs number) assigned by National Center for Biotechnology Information (NCBI), BMI = body mass index, OR = odds ratio, CI = confidence interval, *EMP1* = pointwise *p* value, EMP2 = *p* value corrected for multiple testing, [a] previously reported GDM-associated SNP, the SNP significantly associated with GDM is indicated in bold

Figure 3.2 is a forest plot that visually illustrates the ORs and 95% confidence interval values of the logistic regression (Table 3.6). The forest plot shows that the minor allele of *PDX1* rs4581569 was significantly associated with GDM as the confidence interval for this SNP does not cross the OR threshold of 1. The rest of the SNPs had confidence intervals that ranged over OR=1, indicating no significant association to either low or high GDM risk.

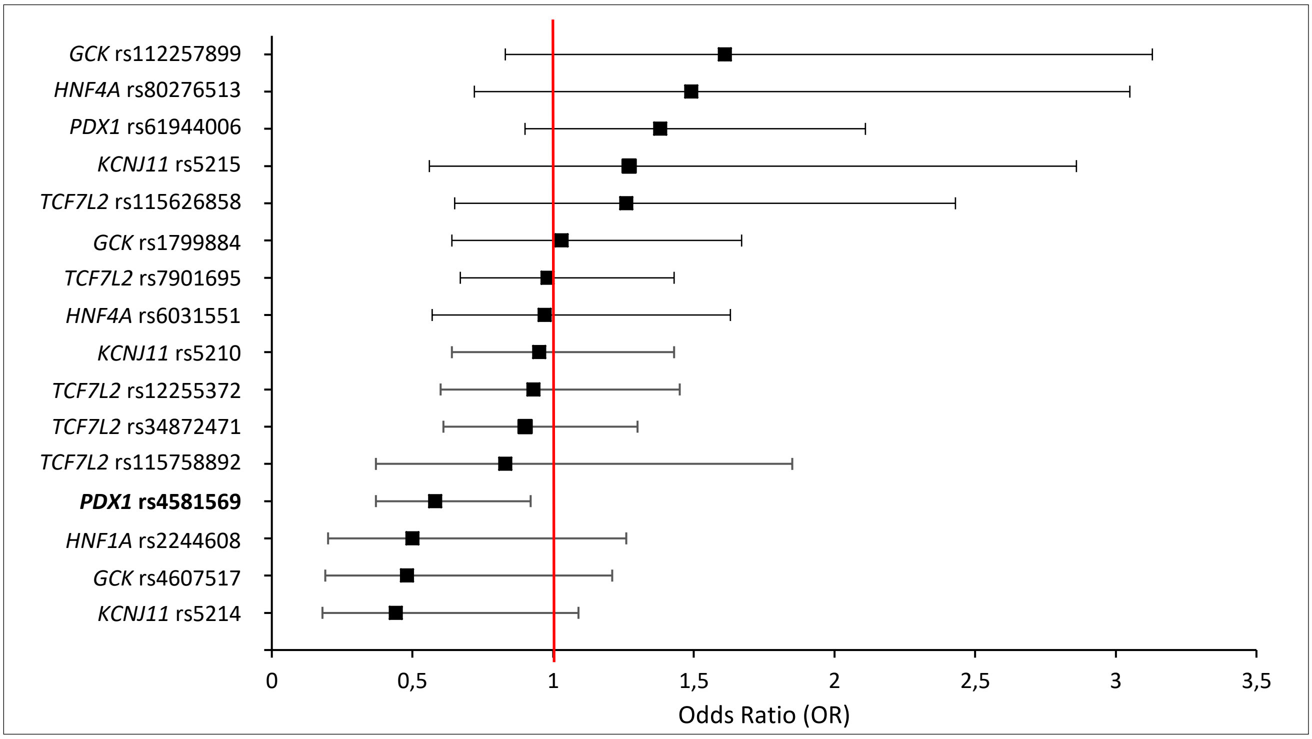


Figure 3.2: Forest plot of odds ratios (ORs) for the SNPs genotyped and tested for association with GDM. The SNP significantly associated with GDM is indicated in bold.

Table 3.7 shows the genotype association results for the SNPs analyzed under the additive model. The genotypes (T/T and T/C) of rs4581569 (indicated in bold in Table 3.7), including the minor allele T, were significantly associated with GDM after adjusting for BMI and age. The calculation of OR values failed, due to insufficient distributions for genotype combinations for some SNPs, where there were zero participants in either the case or control group having at least one genotype.

Table 3.7: Genotype association results for SNPs obtained via logistic regression analysis

			Unadjusted for BMI & Age			Adjusted for BMI & Age		
Gene	SNP ID	n	OR [95% CI]	*EMP1*	*EMP2*	OR [95% CI]	*EMP1*	*EMP2*
GCK	rs1799884[a]	233	0.77 [0.34-1.73]	0.606	1.000	0.74 [0.33-1.69]	0.487	0.998
	rs112257899	240	-	0.037	1.000	-	0.326	1.000
	rs4607517[a]	232	-	0.429	1.000	-	0.316	1.000
TCF7L2	rs34872471	239	0.01 [0.60-1.36]	0.601	1.000	0.93 [0.21-0.62]	0.752	1.000
	rs7901695[a]	240	0.97 [0.64-1.47]	0.843	1.000	1.00 [0.65-1.52]	0.998	1.000
	rs115626858	228	1.41 [0.35-5.70]	0.239	1.000	1.33 [0.32-5.45]	0.448	1.000
	rs115758892	233	-	0.839	1.000	-	0.739	1.000
	rs12255372[a]	233	0.93 [0.56-1.57]	0.798	1.000	0.90 [0.53-1.53]	0.687	1.000
KCNJ11	rs5210	229	0.99 [0.65-1.50]	0.921	1.000	0.96 [0.62-1.48]	0.851	1.000
	rs5214	239	-	0.350	1.000	-	0.359	1.000
	rs5215	234	-	1.000	1.000	-	1.000	1.000
HNF1A	rs2244608	228	-	1.000	1.000	-	1.000	1.000
PDX1	rs61944006	236	1.28 [0.66-2.47]	0.483	0.999	1.22 [0.63-2.40]	0.551	1.000
	rs4581569	**235**	**0.51 [0.27-0.97]**	**0.029**	**0.208**	**0.52 [0.28-1.00]**	**0.033**	**0.267**
HNF4A	rs80276513	237	2.00 [0.85-4.73]	0.077	0.577	1.90 [0.80-4.54]	0.106	0.697
	rs6031551	233	0.89 [0.39-2.05]	0.792	1.000	0.84 [0.36-1.95]	0.624	1.000

SNP = single nucleotide polymorphism, SNP ID = universal SNP identification tag (rs number) assigned by National Center for Biotechnology Information (NCBI), BMI = body mass index, OR = odds ratio, CI = confidence interval, *EMP1* = pointwise *p* value, EMP2 = *p* value corrected for multiple testing,
[a] previously reported GDM-associated SNP, the SNP significantly associated with GDM is indicated in bold

Assessing the association of SNPs with fasting glucose measurements yielded a significant result for the SNP, rs4581569 (indicated in bold in Table 3.8), which was also found to be associated with a low GDM risk in this study. However, the *p* value was no longer significant when adjusting for covariates, causing the *p* value to fall to 0.050, a value on the significance threshold. The SNP, rs61944006, also assessed within the same gene (*PDX1*) as the significantly associated SNP, was the closest to the significance threshold in comparison to all the other SNPs assessed with a *p* value of 0.085.

Table 3.8: Linear regression results for assessing SNPs with fasting glucose measurements from the 75 g 2h OGTT

			Unadjusted for BMI & Age			Adjusted for BMI & Age		
Gene	SNP ID	n	OR [95% CI]	*EMP1*	*EMP2*	OR [95% CI]	*EMP1*	*EMP2*
GCK	rs1799884[a]	233	0.005 [-0.012 - 0.023]	0.551	1.000	0.005 [-0.012 - 0.022]	0.570	1.000
	rs112257899	240	0.013 [-0.013 - 0.038]	0.309	1.000	0.014 [-0.011 - 0.039]	0.262	0.999
	rs4607517[a]	232	0.000 [-0.028 - 0.028]	0.981	1.000	0.000 [-0.027 - 0.028]	0.986	1.000
TCF7L2	rs34872471	239	0.005 [-0.019 - 0.009]	0.477	1.000	0.004 [-0.018 - 0.010]	0.550	1.000
	rs7901695[a]	240	0.002 [-0.017 - 0.012]	0.747	1.000	0.001 [-0.016 - 0.013]	0.845	1.000
	rs115626858	228	0.002 [-0.026 - 0.022]	0.879	1.000	0.006 [-0.030 - 0.018]	0.669	1.000
	rs115758892	233	0.012 [-0.038 - 0.015]	0.419	1.000	0.016 [-0.042 - 0.010]	0.231	0.998
	rs12255372[a]	233	0.001 [-0.014 - 0.016]	0.931	1.000	0.000 [-0.015 - 0.015]	0.977	1.000
KCNJ11	rs5210	229	0.001 [-0.013 - 0.015]	0.913	1.000	0.000 [-0.014 - 0.014]	0.971	1.000
	rs5214	239	0.021 [-0.049 - 0.006]	0.136	0.928	0.021 [-0.049 - 0.007]	0.126	0.942
	rs5215	234	0.012 [-0.019 - 0.043]	0.459	1.000	0.012 [-0.019 - 0.044]	0.457	1.000
HNF1A	rs2244608	228	0.023 [-0.052 - 0.007]	0.127	0.935	0.022 [-0.051 - 0.007]	0.129	0.942
PDX1	rs61944006	236	0.015 [-0.001 - 0.032]	0.085	0.801	0.012 [-0.005 - 0.029]	0.157	0.965
	rs4581569	**235**	**0.016 [-0.031 - 0.001]**	**0.036**	**0.487**	**0.015 [-0.029 - 0.000]**	**0.050**	**0.683**
HNF4A	rs80276513	237	0.004 [-0.028 - 0.020]	0.744	1.000	0.005 [-0.029 - 0.018]	0.668	1.000
	rs6031551	233	0.005 [-0.014 - 0.023]	0.613	1.000	0.005 [-0.013 - 0.024]	0.567	1.000

SNP = single nucleotide polymorphism, SNP ID = universal SNP identification tag (rs number) assigned by National Center for Biotechnology Information (NCBI), BMI = body mass index, OR = odds ratio, CI = confidence interval, *EMP1* = pointwise *p* value, EMP2 = *p* value corrected for multiple testing, [a] previously reported GDM-associated SNP, the SNP significantly associated with GDM is indicated in bold

3.6 SNP FUNCTIONALITY PREDICTION

The function annotation and the predicted functional information gathered for rs4581569 and the linked SNP are shown in Table 3.9. The variant/consequence type from RegulomeDB and VEP for both variants is jointly displayed in the table. If the variant was predicted to be in a regulatory region it was also indicated in Table 3.9 as a “regulatory_variant”. The impact, as defined by Ensembl’s VEP, was “MODIFIER” for both SNPs which is usually assigned to non-coding variants where the impact is unknown or difficult to determine. These variants have also not been reported on Clinvar. There is thus no information with regards to their pathogenicity and clinical health impact on individuals, classifying them as variants of uncertain clinical significance.

According to the interpretation of the CADD and RegulomeDB scores, both SNPs fell under the threshold values, which indicate that there is little to no information related to their functional impact. The CADD scores were below 10, which fell into the irrelevant/ unpredictive score range. Similarly, RegulomeDB yielded a score of four with the following interpretation; “minimal binding evidence”.

Table 3.9: Functional prediction information obtained from Variant Effect Predictor (VEP) and RegulomeDB

SNPs	Variant (from RegulomeDB)/ Consequence Type (from VEP)	Impact (VEP)	Clinical Significance	CADD Score	RegulomeDB Score
rs4581569	**intron_variant**	MODIFIER	unknown	0.224	4
rs9512918	**intron_variant**. regulatory_region_variant	MODIFIER	unknown	2.900	4

3.7 POPULATION ALLELE FREQUENCIES OF THE SIGNIFICANTLY ASSOCIATED SNP

Figure 3.3 shows the allele frequency across a set of global populations for rs4581569, the SNP found to be associated with GDM risk in this study (labelled as GDM cohort). When comparing the allele frequency of the South African cohort to that of other populations, only the ESN population differed from the current cohort (*p* = 0.016).

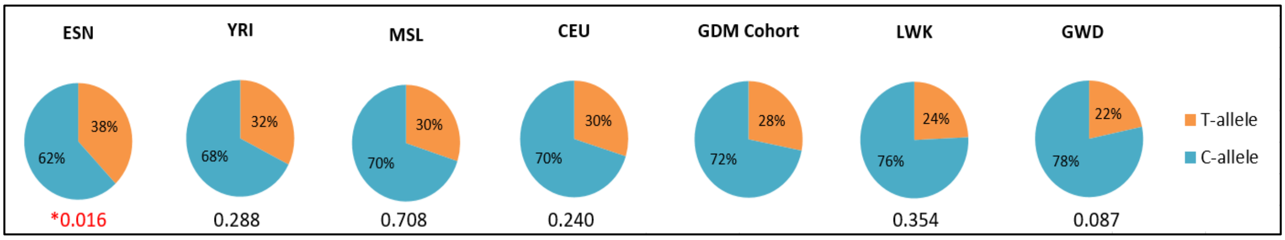


Figure 3.3: Pie charts illustrating the allele frequencies (T/ C-allele) for rs4581569. The *p* values were generated from comparing the allele frequencies between GDM cohort and each respective African population (ESN, ARI, MSL, LWK or GWD) and the CEU European population.

4 DISCUSSION

The aim of this study was to investigate the genetic risk of GDM in a South African cohort. Genes with variants previously reported as being associated with GDM in other populations were selected and investigated to see if they were similarly associated with GDM in a black South African cohort. Due to the scarcity of GDM genetic studies, and the similarity in pathophysiology of GDM to T2D, this study focused on genes that were previously linked to monogenic diabetes, and genes that have been repeatedly and strongly associated with T2D. After an extensive literature review, five MODY genes (*HNF4A, GCK, HNF1A, PDX1,* and *KCNJ11*) and one T2D-associated gene (*TCF7L*) were selected for investigation.

4.1 COHORT CHARACTERIZATION

No significant differences were detected between GDM positive and GDM negative women for all the continuous and categorical variables, except for fasting glucose levels. The characteristics for which no significant differences were found include age, continues BMI measurements, BMI categories, household asset scores, number of previous pregnancies, and education. All these characteristics, especially BMI and age, are known factors that influence GDM risk. Hence, these cofactors can be excluded as having an impact on the genetic association results.

Most women in the cohort were overweight and obese (75%). The prevalence of overweight and obesity is increasing in populations, such as the Sowetan population, that are experiencing rapid urbanization (Kruger *et al.*, 2002, Mfenyana *et al.*, 2006, Micklesfield *et al.*, 2018). The most evident reason for this increase in BMI is the transition to a lifestyle that includes reduced levels of physical activity and increased intake of energy-dense foods (Popkin and Gordon-Larsen, 2004). This trend makes the population more vulnerable to non-communicable diseases, such as diabetes, that might be further exacerbated by genetic factors (Tibazarwa *et al.*, 2009).

4.2 SIGNIFICANT FINDINGS BETWEEN THE *PDX1* GENE, GDM, AND FASTING GLUCOSE LEVELS

The *PDX1* gene was found to be significantly associated with women who tested negative for GDM (controls). The *PDX1* gene is known to have a major role in the development of the pancreas and the functioning of the pancreatic cells for glucose metabolism. During embryogenesis, PDX1 is expressed in proliferative progenitor cells, giving rise to the pancreatic buds, and eventual differentiation into the exocrine, endocrine and ductal portions of the pancreas (Cox and Kushner, 2017, Guz *et al.*, 1995, Jonsson *et al.*, 1994, Offield *et al.*, 1996). After development, the gene is primarily responsible for the maintenance and the function of the islet cells, through regulating related genes, such as *insulin*, *glucose transporter 2*, *GCK*, *somatostatin*, and *islet amyloid polypeptide* (Ashizawa *et al.*, 2004, Kropp *et al.*, 2018).

In this study, a significant association was detected between the minor allele of rs4581569 in the *PDX1* gene and GDM risk after adjusting for BMI and age (*EMP1* = 0.031). The genotypic association for this SNP and GDM was also found to be significant under the additive model (adjusted *EMP1* = 0.033). The OR values of <1, shows that the minor T-allele, and the homozygous T/T, and heterozygous T/C genotypes, were negatively associated with the risk of GDM. The SNP was determined to have a moderate effect on GDM, with the genotypic association having a slightly stronger effect in comparison to the allelic association (adjusted OR = 0.51 vs 0.62). The overall result, however, showed that carrying a minor T-allele of rs4581569 decreased the odds of having GDM for women in this cohort.

The association was still significant after adjusting for BMI and age. Only a minor change was observed after correcting for the two cofactors for both the allele (*EMP1* = 0.015 vs 0.031) and genotypic association (*EMP1* = 0.029 vs 0.033). The allelic and genotypic association with GDM are therefore more likely to be due to genetic variation, captured by the rs4581569 tagSNP, within *PDX1* than the combination of cofactors, BMI and age.

This is a novel discovery for GDM-association studies considering the gene and genetic variant involved, and the inverse association with GDM detected. Genetic variation within the *PDX1* gene has not been reported before in association with a lower risk of GDM. With regard to lower GDM-risk, only one other study has reported associated variants; one in the *LOC646736*/*IRS1* region and another in the *SLC30A8* gene (Rosta *et al.*, 2017). *PDX1* variants reported to be implicated in diabetes within the literature are either rare and highly penetrant, segregating in families with a strong family history of diabetes, or are contained within coding regions, which have been functionally proven to result in a defective PDX1 protein (Doddabelavangala Mruthyunjaya *et al.*, 2017, Gragnoli *et al.*, 2005b, Weng *et al.*, 2002, Yang and Chan, 2016). Other studies have not found GDM-associated *PDX1* variants with moderate effect sizes similar to what have been determined by this study.

After finding the rs4581569-GDM association, an association with fasting glucose could have been expected, as the majority of women with GDM in the cohort were diagnosed on fasting glucose alone (Macaulay *et al.*, 2018). As such, this SNP was significantly associated with women who had lower fasting glucose readings, however, only with a minor effect (beta = -0.015). The inverse of the beta value (due to data that has been transformed), showed that carrying a T-allele decreased fasting glucose readings by ~1 mmol/L (0.966) for women in the GDM cohort.

Once covariates were added to the association testing model, there were no significant findings between rs4581569 and fasting glucose. The *p* value was borderline significant (at 0.050), and could be due to the fact that the study might have been underpowered to detect the small effect of the SNP while accounting for other factors. BMI, however, has been suggested to mediate the genetic association of the *PDX1* gene with fasting glucose. Within the Meta-Analyses of Glucose- and Insulin-related traits Consortium (MAGIC), the *PDX1* gene has been identified as a factor effecting fasting glucose, and was included in a joint meta-analysis, which simultaneously tested for the genetic main effect, adjusted for BMI, and the potential interaction with BMI. In the meta-analysis of 52 studies, comprising of more than 96,496 non-diabetic individuals of European decent, the *PDX1* gene and its interaction with BMI, achieved a genome-wide significance at $P<5\times10^{-8}$ (Manning *et al.*, 2012). In a follow-up study in Korean individuals, the index SNP identified in *PDX1* (rs2293941) was significantly associated in lean participants (*p* value = 0.001), and marginally associated with dichotomized BMI (*p* value = 0.04) (Hong *et al.*, 2014). The majority of women in our cohort were overweight or obese, and thus BMI might be a contributing factor influencing fasting glucose levels in black South African women.

The SNP, rs4581569, found to be associated with GDM and fasting glucose, was selected as a tagSNP, and was completely linked (r^2 =1) to one other SNP, rs9512918. Since rs4581569 is linked to rs9512918, the genotyped SNP might only be indirectly associated with GDM and fasting glucose. The linked SNP could therefore be inferred to be associated with a lower risk of GDM and lower fasting glucose, and potentially have a protective role against the development of GDM.

4.3 SNP CHARACTERIZATION

The functions of the associated SNPs (rs4571569, and the linked SNP, rs9512918) and their relevance to disease or phenotypes, have not been reported before on Clinvar and Ensembl. These two SNPs are known intronic SNPs with functions or consequences that are either non-existent or not yet discovered, as annotated by VEP with the impacted described as “MODIFIER”. The chance of having a functional impact was also predicted to be low, with a CADD score (<10) in the irrelevant/unpredictive range, and a regulomeDB score of 4, indicating that there is no evidence of binding with elements and factors for gene regulation. The variants may very well be of no biological significance, or information is just lacking in regard to their possible functional impact.

The little information on functionality for these SNPs is not unusual, the functionality of the majority of non-coding SNPs is difficult to determine due to incomplete knowledge on the variety of mechanisms by which they are considered to regulate gene expression (Rojano *et al.*, 2018). Regulatory effects are also specific to certain cell- and tissue types, development stages or environmental conditions, making them harder to predict and to experimentally identify in functional studies (Nishizaki and Boyle, 2017, Zappala and Montgomery, 2016).The SNPs associated with a protective effect provides another challenge, as functional prediction tools are specifically designed to distinguish pathogenic SNPs from those that are benign. Ultimately, the tools used have their own limitations, which involve the availability of information, and/or the algorithms used to rate the strength of evidence (Nishizaki and Boyle, 2017).

The linked SNP, rs9512918, has been annotated by VEP to be located within a promoter, a regulatory element, whereas the tagSNP, rs4571569, has been annotated to be contained in an intronic region only. The linked SNP is therefore more likely to be functional (or be the causative SNP) than the tagSNP genotyped in this study. The higher CADD score between the two, also indicates that rs9512818 has a greater chance of having a functional impact. Prioritizing rs9512818 with better evidence of being functional is recommended to reduce time and resources spent to explore the biological significance of the association found even further (Nishizaki and Boyle, 2017, Shen *et al.*, 2010).

4.4 POPULATION ALLELE FREQUENCY

Allele frequency data of rs4581569 has revealed that the minor allele is common amongst global populations, including populations across the African continent (GWD in Western Divisions in the Gambia, MSL in Sierra Leone, the ESN and YRI in Nigeria, and the Luhya in Webuye, Kenya) and the CEU population, representing people of European ancestry. The allele frequency for the T-allele ranged between 0.38 and 0.22, with no significant differences detected between the South African cohort and the other examined populations, except for the ESN population from Nigeria (p = 0.016). The SNP is therefore not specific to the black South African population, and we could expect the SNP to be present in most populations. Since the SNP is common, and GDM is relatively prevalent worldwide, the association found is in agreement with the "common variant, common disease" hypothesis. If the SNP indeed has a protective effect against GDM, it could have a selective advantage, which could have resulted in the allele to rise in frequency to become a common variant (Butler *et al.*, 2017). The allele frequency data are therefore in support of assessing this SNP in association with GDM in other populations.

Common and naturally occurring SNPs, such as rs4581569, could be particularly important for developing genetic therapies, as it is relevant to a greater portion of individuals and could be assumed to be harmless and advantageous due to its high allele frequency. The S447X variant of the lipoprotein lipase (*LPL*), is a well-documented example of a common variant identified to significantly lower risk of cardiovascular diseases and hypertension, which has been utilized in a genetic therapy approach to treat people with LPL-deficiency (Gaudet *et al.*, 2013, Niu and Qi, 2011).

4.5 INSIGNIFICANT FINDINGS

Sixteen SNPs, including four previously GDM-associated SNPs within the literature (rs1799884 and rs4607517 in *GCK,* and rs7901695 and rs12255372 in *TCF7L2*) were not significantly associated with GDM or fasting glucose levels in this study. Insignificant findings and replication failure are evident in small-scale gene association studies that have some disadvantages with regard to small sample size, and the under evaluation of gene-gene and gene-environment risk factors, which is different for various populations (Buzdugan *et al.*, 2016, Patnala *et al.*, 2013). Therefore, either the black South African population does not harbour these associations, or there is a chance that true associations have been missed (the possibility of a type II error) (Jorgensen *et al.*, 2009).

The power to detect SNPs with minor effects and small allele frequencies is low in studies having a small sample size (Jorgensen *et al.*, 2009, Kraft *et al.*, 2009). This study, consisting of 240 women (80 cases and 160 controls) was underpowered to detect SNPs with a MAF < 15% and a minor effect between 0.6 and 2.0 (close to an OR of 1), as demonstrated in Section 2.3 in Chapter 2. Power decreased slightly more for SNPs that failed to be genotyped in some individuals, causing the reduction in sample size for each SNP association analysis (the lowest being a 5% reduction).

Genetic epistasis and architecture are also a cause for non-reproducibility and an increased type II error, which includes factors such as the allele frequency and counter effects of other genetic variants (Chen *et al.*, 2019, Greene *et al.*, 2009). A difference in allele frequency of less than 0.01 at a second interacting variant has been shown to dramatically reduce the power to replicate the main effect of a significantly associated variant. Even a reversal of allelic effects has been noted, where an allele identified as being protective becomes associated with increased risk in follow up replication (Greene *et al.*, 2009). The interplay between genetic variants are thus important to take into consideration when conducting genetic association studies, especially in African populations that have a high degree of genetic variation (Bryc *et al.*, 2010, Tishkoff *et al.*, 2009).

4.6 THE MATTER OF CORRECTING FOR MULTIPLE TESTING

The association between rs4581569, GDM, and fasting glucose, became insignificant after correcting for multiple testing (*EMP2* > 0.05). Multiple testing correction is used to filter out possible false positive associations (thus the occurrence of a type I error), and is usually performed in GWAS that test thousands of SNPs simultaneously. In this study it was used as a measure of the validity of the associations found. If the association was able to uphold significance, this would have provided evidence of a true and strong detected genetic association.

Even though multiple testing correction is a standard approach, it is not always applied in candidate gene association studies (Qu *et al.*, 2010). Correcting for multiple testing is sometimes seen as overly conservative when only testing a few candidate SNPs, and when there is no known major gene effects within the disease investigated (Patnala *et al.*, 2013). Gene effects that are assessed for association with complex diseases are known to be very modest, which is also a concern for large studies that remain underpowered to reach very stringent levels of significance (Panagiotou *et al.*, 2011). Therefore, there is a fine line between avoiding false positive associations and not missing a true association, that could be affected by the power of the study. The associations found were therefore regarded as a notable finding despite not being supported by multiple testing correction.

4.7 STRENGHTS AND LIMITATIONS

The study had several strengths, but also some limitations. On the positive side, this study was one of the first to assess genetic factors in association with GDM in a black South African cohort. Investigating MODY genes increased the novelty of this research as these genes have been relatively understudied in people of African descent. The second strength was the selection of South African-specific tagSNPs which may be more relevant than using data from proxy populations. Using the Zulu population data from the AGVP (Gurdasani *et al.*, 2015) provided additional specificity, an improvement from other South African-based studies that, in the past, only had HapMap reference data from geographically and genetically different populations, such as the Yoruba and Luhya, to use as a proxy (May *et al.*, 2013, Teo *et al.*, 2010).

There were four limitations to this study. The main limiting factor in this study was the sample size that consequently resulted in less power to detect certain associations. Nonetheless, the study was able to identify a genetic association, in a gene that can be explored further through studies on GDM in African populations.

The second limitation was the approach taken for the selection of SNPs. The intent of using the Tagger algorithm was to select common tagSNPs (with a frequency of >5%). In order to reduce the amount of SNPs suited to the budget of the study, the CEU-block region, containing a SNP that has previously been significantly associated with GDM, were used as input. Following this approach yielded only seven tagSNPs, including rs112257899 in *GCK*, rs73169687, rs7981781, rs4581569, rs61944006 in *PDX1*, as well as, rs5210 and rs5214 in *KCNJ11*. Sixteen SNPs were not in LD with any other SNP. The small amount of tagSNPs captured was conceivable as LD stretches in African populations are shorter and more fragmented compared to that of Caucasian populations (Bush and Moore, 2012). The advantage of including non-tagSNPs however, was that it facilitated direct association testing (Jorgensen *et al.*, 2009).

The third limitation was the high genotyping failure rate and assay design challenges that caused seven SNPs not be analysed from the 23 SNPs that had originally been chosen for investigation. Failure in genotyping SNPs in some individuals also caused a reduction in sample size for each allele and genotype association analysis. Taking these missing genotype and allele results into account could have had an influence on the results, and therefore to confirm the statistical findings, it is advisable to troubleshoot and thereby re-genotype the SNPs with a high genotyping failure rate of >0.5.

Finally, the fourth limitation has to do with bioinformatic tools currently available for function prediction. There is still a large gap in our understanding of the functional impact of genetic variants found in the human genome. Bioinformatic tools are unable to distinguish most SNPs from being non-functional from those only lacking in functional information. Moreover, the tools also seem to be inadequate for the functional prediction of intronic variants, and variants possibly having protective functions. Combining a few functional prediction tools was thus necessary to pull as much information as possible and to confirm the prediction outputs. As the body of evidence grows, and functional prediction and annotation tools become more sophisticated, confidence in the biological significance or insignificance of rs4581569 and rs9512818 will likely become more calculable.

4.8 FUTURE PROSPECTS

To further, and more thoroughly, evaluate the contribution of genetic variation from the six genes of interest, each entire gene should be used for tagSNP selection. This would then determine more accurately what the contribution of genetic variation from each gene is relative to the risk of GDM. For this purpose, a larger budget would be necessary, and maybe the use of higher throughput technologies, to make it economically feasible. Sequencing could genotype a more comprehensive set of SNPs, also including the SNPs previously found to be associated with GDM in other populations.

Increasing the sample size would have a beneficial impact on the power of future studies to identify rare variants of low or modest effect on the phenotype. Further investigating rs4581569 in a larger cohort could validate the genetic association found with GDM and fasting glucose. This is particularly necessary since the significance of the association with fasting glucose did not pass multiple testing correction. The contribution of the genetic variation, relative to the influence of BMI on the association with fasting glucose levels, would also become clearer in a larger study. A better representation of women of the population would be obtained in a bigger sample size, including women with characteristics, such as high and low BMI, in almost equal proportion. This would allow for increased comparison potential.

Considering the fact that BMI might mediate the genetic effect on GDM and fasting glucose, the genetic association and its interaction with BMI should be further investigated. It would be interesting to examine if the genetic association with fasting glucose is only present in a certain group of women having either normal (≤24.9 kg/m^2) or high BMIs (>25 kg/m^2). Further extending the research to the study of gene-gene and gene-environment effects could give comprehensive perspective on all the biological factors influencing GDM risk in the black South African population.

5 CONCLUSION

In this genetic association study, 16 SNPs in five MODY genes (*HNF4A, GCK, HNF1A, PDX1,* and *KCNJ11*) and one T2D-associated gene (*TCF7L*) were assessed in relation to GDM risk. In the black South African cohort, one tagSNP, rs4581569 in the *PDX1* gene, was significantly associated with GDM and fasting glucose levels. Regression analyses revealed that carrying the minor T-allele of rs4581569 was associated with decreased GDM risk and low fasting glucose levels. The genetic variation captured by the tagSNP, including the linked SNP rs4581569, may have a potential protective effect in black South African women, by regulating glucose metabolism. The association with regard to GDM risk is a novel discovery, but due to a relatively small sample size and the finding not being supported by multiple testing correction, this finding requires validation. Studying this SNP in a bigger cohort is advised to increase the power to detect the moderate effect identified.

The black South African population has been relatively understudied with regard to genetic factors contributing to the development of GDM, as well as T2D. This study has therefore raised awareness of the scarcity of genetic studies in this particular population group. The findings of this study and the allele frequency data generated for rs4581569 add to the research efforts that have recently begun to elucidate the genetic determinants of diabetes in Sub-Saharan African populations. In the future of personalised medicine, this could be used for developing treatment or preventative measures tailored to an individual’s genetic profile.

6 Reference list

Abecasis, G. R., Altshuler, D., Auton, A., *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature,* 467 (7319)**,** 1061-1073. doi: 10.1038/nature09534

Adeyemo, A. A., Tekola-Ayele, F., Doumatey, A. P., *et al.* (2015) Evaluation of Genome Wide Association Study Associated Type 2 Diabetes Susceptibility Loci in Sub Saharan Africans. *Frontiers in Genetics,* 6**,** 335. doi: 10.3389/fgene.2015.00335

Adeyemo, A. A., Zaghloul, N. A., Chen, G., *et al.* (2019) ZRANB3 is an African-specific type 2 diabetes locus associated with beta-cell mass and insulin response. *Nature communications,* 10 (1)**,** 3195-3195. doi: 10.1038/s41467-019-10967-7

Ali, O. (2013) Genetics of type 2 diabetes. *World Journal of Diabetes,* 4 (4)**,** 114-123. doi: 10.4239/wjd.v4.i4.114

Althari, S. & Gloyn, A. L. (2015) When is it MODY? Challenges in the Interpretation of Sequence Variants in MODY Genes. *Review of Diabetic Studies,* 12 (3-4)**,** 330-348. doi: 10.1900/RDS.2015.12.330

Amos, W., Driscoll, E. & Hoffman, J. (2011) Candidate genes versus genome-wide associations: which are better for detecting genetic susceptibility to infectious disease? *Proceedings of the Royal Society B: Biological Sciences,* 278 (1709)**,** 1183-1188. doi: 10.1098/rspb.2010.1920

Anderson, C. A., Pettersson, F. H., Clarke, G. M., *et al.* (2010) Data quality control in genetic case-control association studies. *Nature protocols,* 5 (9)**,** 1564-1573. doi: 10.1038/nprot.2010.116

Ao, D., Wang, H.-J., Wang, L.-F., *et al.* (2015) The rs2237892 Polymorphism in KCNQ1 influences gestational diabetes mellitus and glucose levels: A case-control study and meta-analysis. *PloS ONE,* 10 (6)**,** e0128901. doi: 10.1371/journal.pone.0128901

Ashizawa, S., Brunicardi, F. C. & Wang, X.-P. (2004) PDX-1 and the pancreas. *Pancreas,* 28 (2)**,** 109-120. doi: 10.1097/00006676-200403000-00001

Auer, P. L., Johnsen, J. M., Johnson, A. D., *et al.* (2012) Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *American Journal of Human Genetics,* 91 (5)**,** 794-808. doi: 10.1016/j.ajhg.2012.08.031

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet,* 7 (10)**,** 781-791. doi: 10.1038/nrg1916

Barker, D. J. (1990) The fetal and infant origins of adult disease. *BMJ (Clinical Research Ed.),* 301 (6761)**,** 1111. doi: 10.1136/bmj.301.6761.1111

Barker, D. J. (2007) The origins of the developmental origins theory. *Journal of Internal Medicine,* 261 (5)**,** 412-417. doi: 10.1111/j.1365-2796.2007.01809.x

Barrett, J. C. (2009) Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor Protocols,* 2009 (10)**,** pdb. ip71. doi: 10.1101/pdb.ip71

Beischer, N. A., Wein, P., Sheedy, M. T., *et al.* (1996) Identification and treatment of women with hyperglycaemia diagnosed during pregnancy can significantly reduce perinatal mortality rates. *Australian and New Zealand Journal of Obstetrics and Gynaecology,* 36 (3)**,** 239-247. doi: 10.1111/j.1479-828X.1996.tb02703.x

Bellamy, L., Casas, J.-P., Hingorani, A. D., *et al.* (2009) Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *Lancet,* 373 (9677)**,** 1773-1779. doi: 10.1016/S0140-6736(09)60731-5

Bentley, A. R., Callier, S. & Rotimi, C. N. (2017) Diversity and inclusion in genomic research: why the uneven progress? *Journal of Community Genetics,* 8 (4)**,** 255-266. doi: 10.1007/s12687-017-0316-6

Berensmeier, S. (2006) Magnetic particles for the separation and purification of nucleic acids. *Applied Microbiology and Biotechnology,* 73 (3)**,** 495-504. doi: 10.1007/s00253-006-0675-0

Beyene, J. & Pare, G. (2013) Statistical genetics with application to population-based study design: a primer for clinicians. *European Heart Journal,* 35 (8)**,** 495-500. doi: 10.1093/eurheartj/eht272

Bouchard, L., Thibault, S., Guay, S. P., *et al.* (2010) Leptin gene epigenetic adaptation to impaired glucose metabolism during pregnancy. *Diabetes Care,* 33 (11)**,** 2436-2441. doi: 10.2337/dc10-1024

Boyle, A. P., Hong, E. L., Hariharan, M., *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research,* 22 (9)**,** 1790-1797. doi: 10.1101/gr.137323.112

Brookes, A. J. (1999) The essence of SNPs. *Gene,* 234 (2)**,** 177-186. doi: 10.1016/S0378-1119(99)00219-X

Bryc, K., Auton, A., Nelson, M. R., *et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences,* 107 (2)**,** 786-791. doi: 10.1073/pnas.0909559107

Buchanan, T. A. & Xiang, A. H. (2005) Gestational diabetes mellitus. *The Journal of Clinical Investigation,* 115 (3)**,** 485-491. doi: 10.1172/JCI24531

Bush, W. S. & Moore, J. H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology,* 8 (12)**,** e1002822. doi: 10.1371/journal.pcbi.1002822

Butler, J. M., Hall, N., Narendran, N., *et al.* (2017) Identification of candidate protective variants for common diseases and evaluation of their protective potential. *BMC Genomics,* 18 (1)**,** 575. doi: 10.1186/s12864-017-3964-3

Buzdugan, L., Kalisch, M., Navarro, A., *et al.* (2016) Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics,* 32 (13)**,** 1990-2000. doi: 10.1093/bioinformatics/btw128

Carlson, C. S., Eberle, M. A., Rieder, M. J., *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics,* 74 (1)**,** 106-120. doi: 10.1086/381000

Carolan-Olah, M., Duarte-Gardea, M. & Lechuga, J. (2015) A critical review: early life nutrition and prenatal programming for adult disease. *Journal of Clinical Nursing,* 24 (23-24)**,** 3716-3729. doi: 10.1111/jocn.12951

Chang, S., Wang, Z., Wu, L., *et al.* (2017) Association between TCF7L2 polymorphisms and gestational diabetes mellitus: A meta-analysis. *Journal of Diabetes Investigation,* 8 (4)**,** 560-570. doi: 10.1111/jdi.12612

Chen, A. H., Ge, W., Metcalf, W., *et al.* (2019) An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers. *Heredity,* 122 (5)**,** 660-671. doi: 10.1038/s41437-018-0162-2

Cho, Y., Kim, T., Lim, S., *et al.* (2009) Type 2 diabetes-associated genetic variants discovered in the recent genome-wide association studies are related to gestational diabetes mellitus in the Korean population. *Diabetologia,* 52 (2)**,** 253-261. doi: 10.1007/s00125-008-1196-4

Clarke, G. M., Anderson, C. A., Pettersson, F. H., *et al.* (2011) Basic statistical analysis in genetic case-control studies. *Nature Protocols,* 6 (2)**,** 121-133. doi: 10.1038/nprot.2010.182

Colclough, K., Bellanne-Chantelot, C., Saint-Martin, C., *et al.* (2013) Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha and 4 alpha in maturity-onset diabetes of the young and hyperinsulinemic hypoglycemia. *Human Mutation,* 34 (5)**,** 669-685. doi: 10.1002/humu.22279

Colclough, K., Saint-Martin, C., Timsit, J., *et al.* (2014) Clinical utility gene card for: Maturity-onset diabetes of the young. *European Journal of Human Genetics,* 22 (9)**,** 1153. doi: 10.1038/ejhg.2014.14

Coustan, D. R., Lowe, L. P., Metzger, B. E., *et al.* (2010) The HAPO Study: Paving The Way For New Diagnostic Criteria For GDM. *American Journal of Obstetrics and Gynecology,* 202 (6)**,** 654.e1-654.e6. doi: 10.1016/j.ajog.2010.04.006

Cox, A. R. & Kushner, J. A. (2017) Area IV Knockout Reveals How Pdx1 Is Regulated in Postnatal β-Cell Development. *Diabetes,* 66 (11)**,** 2738-2740. doi: 10.2337/dbi17-0036

Dabelea, D. (2007) The predisposition to obesity and diabetes in offspring of diabetic mothers. *Diabetes Care,* 30 (Suppl. 2)**,** S169-S174. doi: 10.2337/dc07-s211

Daly, B., Toulis, K. A., Thomas, N., *et al.* (2018) Increased risk of ischemic heart disease, hypertension, and type 2 diabetes in women with previous gestational diabetes mellitus, a target group in general practice for preventive interventions: A population-based cohort study. *PLoS Medicine,* 15 (1)**,** e1002488. doi: 10.1371/journal.pmed.1002488

De Bakker, P. I. W., Yelensky, R., Pe'er, I., *et al.* (2005) Efficiency and power in genetic association studies. *Nature Genetics,* 37 (11)**,** 1217-1223. doi: 10.1038/ng1669

Del Rosario, M. C., Ossowski, V., Knowler, W. C., *et al.* (2014) Potential epigenetic dysregulation of genes associated with MODY and type 2 diabetes in humans exposed to a diabetic intrauterine environment: an analysis of genome-wide DNA methylation. *Metabolism,* 63 (5)**,** 654-660. doi: 10.1016/j.metabol.2014.01.007

Ding, M., Chavarro, J., Olsen, S., *et al.* (2018) Genetic variants of gestational diabetes mellitus: a study of 112 SNPs among 8722 women in two independent populations. *Diabetologia,* 61 (8)**,** 1758-1768. doi: 10.1007/s00125-018-4637-8

Doddabelavangala Mruthyunjaya, M., Chapla, A., Hesarghatta Shyamasunder, A., *et al.* (2017) Comprehensive Maturity Onset Diabetes of the Young (MODY) Gene Screening in Pregnant Women with Diabetes in India. *PLoS ONE,* 12 (1)**,** e0168656. doi: 10.1371/journal.pone.0168656

Ellis, J. A. & Ong, B. (2017) The MassARRAY® System for targeted SNP genotyping. In*:* S. J. White & S. Cantsilieris, (eds.). *Genotyping. Methods in Molecular Biology.* NY: Humana Press, pp. 77-94. doi: 10.1007/978-1-4939-6442-0_5

Fatima, S. S., Chaudhry, B., Khan, T. A., *et al.* (2016) KCNQ1 rs2237895 polymorphism is associated with Gestational Diabetes in Pakistani Women. *Pakistan Journal of Medical Sciences,* 32 (6)**,** 1380-1385. doi: 10.12669/pjms.326.11052

Ferrara, A. (2007) Increasing prevalence of gestational diabetes mellitus. *Diabetes Care,* 30 (Suppl. 2)**,** S141-S146. doi: 10.2337/dc07-s206

Firdous, P., Nissar, K., Ali, S., *et al.* (2018) Genetic Testing of Maturity-Onset Diabetes of the Young Current Status and Future Perspectives. *Frontiers in Endocrinology,* 9**,** 253. doi: 10.3389/fendo.2018.00253

Flannick, J., Johansson, S. & Njolstad, P. R. (2016) Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nature Reviews Endocrinology,* 12 (7)**,** 394-406. doi: 10.1038/nrendo.2016.50

Fuchsberger, C., Flannick, J., Teslovich, T. M., *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature,* 536 (7614)**,** 41-47. doi: 10.1038/nature18642

Fusi, N., Lippert, C., Lawrence, N. D., *et al.* (2014) Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications,* 5 (1)**,** 4890. doi: 10.1038/ncomms5890

Gabriel, S., Ziaugra, L. & Tabbaa, D. (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current Protocols in Human Genetics,* 60 (1)**,** 2.12.1-2.12.18. doi: 10.1002/0471142905.hg0212s60

Gaudet, D., Méthot, J., Déry, S., *et al.* (2013) Efficacy and long-term safety of alipogene tiparvovec (AAV1-LPL S447X) gene therapy for lipoprotein lipase deficiency: an open-label trial. *Gene Therapy,* 20 (4)**,** 361–369. doi: 10.1038/gt.2012.43

Gragnoli, C., Stanojevic, V., Gorini, A., *et al.* (2005a) IPF-1/MODY4 gene missense mutation in an Italian family with type 2 and gestational diabetes. *Metabolism,* 54 (8)**,** 983-8. doi: 10.1016/j.metabol.2005.01.037

Gragnoli, C., Stanojevic, V., Gorini, A., *et al.* (2005b) IPF-1/MODY4 gene missense mutation in an Italian family with type 2 and gestational diabetes. *Metabolism,* 54 (8)**,** 983-988. doi: 10.1016/j.metabol.2005.01.037

Green, M. R. & Sambrook, J. (2018) Isolation and Quantification of DNA. *Cold Spring Harbor Protocols,* 2018 (6)**,** pdb. top093336. doi: 10.1101/pdb.top093336

Green, M. R. & Sambrook, J. (2019) Analysis of DNA by agarose gel electrophoresis. *Cold Spring Harbor Protocols,* 2019 (1)**,** pdb. top100388. doi: 10.1101/pdb.top100388

Greene, C. S., Penrod, N. M., Williams, S. M., *et al.* (2009) Failure to replicate a genetic association may provide important clues about genetic architecture. *PloS ONE,* 4 (6)**,** e5639. doi: 10.1371/journal.pone.0005639

Guariguata, L., Whiting, D. R., Hambleton, I., *et al.* (2014) Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Research and Clinical Practice,* 103 (2)**,** 176-185. doi: 10.1016/j.diabres.2013.11.002

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., *et al.* (2015) The African Genome Variation Project shapes medical genetics in Africa. *Nature,* 517 (7534)**,** 327-332. doi: 10.1038/nature13997

Guz, Y., Montminy, M. R., Stein, R., *et al.* (1995) Expression of murine STF-1, a putative insulin gene transcription factor, in beta cells of pancreas, duodenal epithelium and pancreatic exocrine and endocrine progenitors during ontogeny. *Development,* 121 (1)**,** 11-18.

Hanna, F. W. & Peters, J. R. (2002) Screening for gestational diabetes; past, present and future. *Diabetic Medicine,* 19 (5)**,** 351-358. doi: 10.1046/j.1464-5491.2002.00684.x

Hodgkinson, A. & Eyre-Walker, A. (2010) Human triallelic sites: evidence for a new mutational mechanism? *Genetics,* 184 (1)**,** 233-241. doi: 10.1534/genetics.109.110510

Holland, P. M., Abramson, R. D., Watson, R., *et al.* (1991) Detection of specific polymerase chain reaction product by utilizing the 5'----3'exonuclease activity of Thermus aquaticus DNA polymerase. *Proceedings of the National Academy of Sciences,* 88 (16)**,** 7276-7280. doi: 10.1073/pnas.88.16.7276

Hong, K.-W., Chung, M. & Cho, S. B. (2014) Replication of Interactions between Genome-Wide Genetic Variants and Body Mass Index in Fasting Glucose and Insulin Levels. *Genomics & Informatics,* 12 (4)**,** 236. doi: 10.5808/GI.2014.12.4.236

Hübner, J., Heinzler, R., Arlt, C., *et al.* (2017) An automated and compartmented fluidic reactor device for multi-step sample-to-answer processes using magnetic particles. *Reaction Chemistry & Engineering,* 2 (3)**,** 349-365. doi: 10.1039/C6RE00219F

Huerta-Chagoya, A., Vázquez-Cárdenas, P., Moreno-Macías, H., *et al.* (2015) Genetic determinants for gestational diabetes mellitus and related metabolic traits in Mexican women. *PloS ONE,* 10 (5)**,** e0126408. doi: 10.1371/journal.pone.0126408

Jonsson, J., Carlsson, L., Edlund, T., *et al.* (1994) Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature,* 371 (6498)**,** 606-609. doi: 10.1038/371606a0

Jorgensen, T. J., Ruczinski, I., Kessing, B.*, et al.* (2009) Hypothesis-Driven Candidate Gene Association Studies: Practical Design and Analytical Considerations. *American Journal of Epidemiology,* 170 (8)**,** 986-993. doi: 10.1093/aje/kwp242

Jurinke, C., Van Den Boom, D., Cantor, C. R., *et al.* (2002) The use of MassARRAY technology for high throughput genotyping. In*:* J. Hoheisel, (ed.). *Chip Technology. Advances in Biochemical Engineering/Biotechnology.* Heidelberg, DE: Springer, pp. 57-74. doi: 10.1007/3-540-45713-5_4

Kanthimathi, S., Chidambaram, M., Bodhini, D., *et al.* (2017) Association of recently identified type 2 diabetes gene variants with Gestational Diabetes in Asian Indian population. *Molecular Genetics and Genomics*, 292 (3)**,** 585–591. doi: 10.1007/s00438-017-1292-6

Karki, R., Pandya, D., Elston, R. C., *et al.* (2015) Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Medical Genomics,* 8 (1)**,** 37. doi: 10.1186/s12920-015-0115-z

Kasuga, Y., Hata, K., Tajima, A.*, et al.* (2017) Association of common polymorphisms with gestational diabetes mellitus in Japanese women: A case-control study. *Endocrine Journal*, 64 (4)**,** 463-475. doi: 10.1507/endocrj.EJ16-0431

Kim, C., Berger, D. K. & Chamany, S. (2007) Recurrence of gestational diabetes mellitus. *Diabetes Care*, 30 (5)**,** 1314-1319. doi: 10.2337/dc06-2517

Kim, J. Y., Cheong, H. S., Park, B.-L.*, et al.* (2011) Melatonin receptor 1 B polymorphisms associated with the risk of gestational diabetes mellitus. *BMC Medical Genetics,* 12 (1)**,** 82. doi: 10.1186/1471-2350-12-82

Kim, S. & Misra, A. (2007) SNP genotyping: technologies and biomedical applications. *Annual Review of Biomedical Engineering,* 9**,** 289-320. doi: 10.1146/annurev.bioeng.9.060906.152037

Kircher, M., Witten, D. M., Jain, P.*, et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics,* 46 (3)**,** 310-315. doi: 10.1038/ng.2892

Kitzmiller, J. L., Dang-Kilduff, L. & Taslimi, M. M. (2007) Gestational diabetes after delivery: short-term management and long-term risks. *Diabetes Care,* 30 (Suppl. 2)**,** S225-S235. doi: 10.2337/dc07-s221

Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., *et al.* (2009) Fewer permutations, more accurate P-values. *Bioinformatics,* 25 (12)**,** i161-i168. doi: 10.1093/bioinformatics/btp211

Koch, W. H. (2004) Technology platforms for pharmacogenomic diagnostic assays. *Nature Reviews Drug Discovery,* 3 (9)**,** 749–761 doi: 10.1038/nrd1496

Koetsier, G. & Cantor, E. (2019) A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers. Ipswich, MA: New England BioLabs Inc. Available: https://www.neb.com/-/media/catalog/application-notes/mvs_analysis_of_na_concentration_and_purity.pdf [Accessed: 15 June 2019]

Kraft, P., Zeggini, E. & Ioannidis, J. P. A. (2009) Replication in genome-wide association studies. *Statistical Science,* 24 (4)**,** 561-573. doi: 10.1214/09-STS290

Kropp, P. A., Dunn, J. C., Carboneau, B. A., *et al.* (2018) Cooperative function of Pdx1 and Oc1 in multipotent pancreatic progenitors impacts postnatal islet maturation and adaptability. *American Journal of Physiology-Endocrinology and Metabolism,* 314 (4)**,** E308-E321. doi: 10.1152/ajpendo.00260.2017

Kruger, H. S., Venter, C. S., Vorster, H. H., *et al.* (2002) Physical inactivity is the major determinant of obesity in black women in the North West Province, South Africa: the THUSA study. Transition and Health During Urbanisation of South Africa. *Nutrition,* 18 (5)**,** 422-427. doi: 10.1016/s0899-9007(01)00751-1

Ku, C. S., Loy, E. Y., Salim, A., *et al.* (2010) The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics,* 55 (7)**,** 403-415. doi: 10.1038/jhg.2010.55

Kwak, S. H., Kim, S. H., Cho, Y. M., *et al.* (2012) A genome-wide association study of gestational diabetes mellitus in Korean women. *Diabetes,* 61 (2)**,** 531-541. doi: 10.2337/db11-1034

Li, L.-J., Aris, I. M., Su, L. L., *et al.* (2018) Effect of gestational diabetes and hypertensive disorders of pregnancy on postpartum cardiometabolic risk. *Endocrine Connections,* 7 (3)**,** 433-442. doi: 10.1530/EC-17-0359

Lindsay, R. S., Mackin, S. T. & Nelson, S. M. (2017) Gestational diabetes mellitus—right person, right treatment, right time? *BMC Medicine,* 15 (1)**,** 163. doi: 10.1186/s12916-017-0925-2

Liu, Q., Huang, Z., Li, H., *et al.* (2016) Relationship between melatonin receptor 1B (rs10830963 and rs1387153) with gestational diabetes mellitus: a case-control study and meta-analysis. *Archives of Gynecology and Obstetrics,* 294 (1)**,** 55-61. doi: 10.1007/s00404-015-3948-y

Livak, K. J., Flood, S., Marmaro, J., *et al.* (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *Genome Research,* 4 (6)**,** 357-362. doi: 10.1101/gr.4.6.357

Lowe Jr, W. L., Scholtens, D. M., Sandler, V., *et al.* (2016) Genetics of gestational diabetes mellitus and maternal metabolism. *Current Diabetes Reports,* 16 (2)**,** 15. doi: 10.1007/s11892-015-0709-z

Macaulay, S., Dunger, D. B. & Norris, S. A. (2014) Gestational Diabetes Mellitus in Africa: A Systematic Review. *PLoS ONE,* 9 (6)**,** e97871. doi: 10.1371/journal.pone.0097871

Macaulay, S., Ngobeni, M., Dunger, D. B., *et al.* (2018) The prevalence of gestational diabetes mellitus amongst black South African women is a public health concern. *Diabetes Research and Clinical Practice,* 139**,** 278-287. doi: 10.1016/j.diabres.2018.03.012

Manning, A. K., Hivert, M.-F., Scott, R. A., *et al.* (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genetics,* 44 (6)**,** 659-669. doi: 10.1038/ng.2274

Mao, H., Li, Q. & Gao, S. (2012) Meta-analysis of the relationship between common type 2 diabetes risk gene variants with gestational diabetes mellitus. *PLoS ONE,* 7 (9)**,** e45882. doi: 10.1371/journal.pone.0045882

Marees, A. T., De Kluiver, H., Stringer, S., *et al.* (2018) A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research,* 27 (2)**,** e1608. doi: 10.1002/mpr.1608

May, A., Hazelhurst, S., Li, Y., *et al.* (2013) Genetic diversity in black South Africans from Soweto. *BMC Genomics,* 14 (1)**,** 644. doi: 10.1186/1471-2164-14-644

Mccarthy, M. I., Abecasis, G. R., Cardon, L. R., *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics,* 9**,** 356-369. doi: 10.1038/nrg2344

Mclaren, W., Gil, L., Hunt, S. E., *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome Biology,* 17 (1)**,** 122. doi: 10.1186/s13059-016-0974-4

Metzger, B. E., Gabbe, S. G., Persson, B., *et al.* (2010) International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. *Diabetes Care,* 33 (3)**,** 676-82. doi: 10.2337/dc09-1848

Mfenyana, K., Griffin, M., Yogeswaran, P., *et al.* (2006) Socio-economic inequalities as a predictor of health in South Africa--the Yenza cross-sectional study. *South African Medical Journal,* 96 (4)**,** 323-330

Micklesfield, L. K., Kagura, J., Munthali, R., *et al.* (2018) Demographic, socio-economic and behavioural correlates of BMI in middle-aged black men and women from urban Johannesburg, South Africa. *Global Health Action,* 11 (Suppl. 2)**,** 1448250. doi: 10.1080/16549716.2018.1448250

Miller, H. C. (1946) The effect of diabetic and prediabetic pregnancies on the fetus and newborn infant. *The Journal of Pediatrics,* 29 (4)**,** 455-461. doi: 10.1016/S0022-3476(46)80164-1

Monteiro, L. J., Norman, J. E., Rice, G. E., *et al.* (2016) Fetal programming and gestational diabetes mellitus. *Placenta,* 48 (Suppl. 1)**,** S54-S60. doi: 10.1016/j.placenta.2015.11.015

Myles, S., Davison, D., Barrett, J., *et al.* (2008) Worldwide population differentiation at disease-associated SNPs. *BMC Medical Genomics*, 1 (1)**,** 22. doi: 10.1186/1755-8794-1-22

Nakai, K., Habano, W., Fujita, T., *et al.* (2002) Highly multiplexed genotyping of coronary artery disease-associated SNPs using MALDI-TOF mass spectrometry. *Human Mutation,* 20 (2)**,** 133-138. doi: 10.1002/humu.10099

Namipashaki, A., Razaghi-Moghadam, Z. & Ansari-Pour, N. (2015) The Essentiality of Reporting Hardy-Weinberg Equilibrium Calculations in Population-Based Genetic Association Studies. *Cell Journal,* 17 (2)**,** 187-192. doi: 10.22074/cellj.2016.3711

Newton-Cheh, C. & Hirschhorn, J. N. (2005) Genetic association studies of complex traits: design and analysis issues. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 573 (1-2)**,** 54-69. doi: 10.1016/j.mrfmmm.2005.01.006

Nishizaki, S. S. & Boyle, A. P. (2017) Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends in Genetics,* 33 (1)**,** 34-45. doi: 10.1016/j.tig.2016.10.008

Niu, W. & Qi, Y. (2011) Meta-based association of the lipoprotein lipase gene S447X variant with hypertension and blood pressure variation. *Journal of Human Hypertension,* 25 (6)**,** 383-390. doi: 10.1038/jhh.2010.68

Offield, M. F., Jetton, T. L., Labosky, P. A., *et al.* (1996) PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development,* 122 (3)**,** 983-995

Osbak, K. K., Colclough, K., Saint-Martin, C., *et al.* (2009) Update on mutations in glucokinase (GCK), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. *Human Mutation,* 30 (11)**,** 1512-1526. doi: 10.1002/humu.21110

Ott, J., Kamatani, Y. & Lathrop, M. (2011) Family-based designs for genome-wide association studies. *Nature Reviews Genetics,* 12**,** 465-474. doi: 10.1038/nrg2989

Panagiotou, O. A., Ioannidis, J. P. & Project, G.-W. S. (2011) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology,* 41 (1)**,** 273-286. doi: 10.1093/ije/dyr178

Patnala, R., Clements, J. & Batra, J. (2013) Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genetics*, 14 (1)**,** 39. doi: 10.1186/1471-2156-14-39

Pawlik, A., Teler, J., Maciejewska, A., *et al.* (2017) Adiponectin and leptin gene polymorphisms in women with gestational diabetes mellitus. *Journal of Assisted Reproduction and Genetics,* 34 (4)**,** 511-516. doi: 10.1007/s10815-016-0866-2

Popkin, B. M. & Gordon-Larsen, P. (2004) The nutrition transition: worldwide obesity dynamics and their determinants. *International Journal of Obesity,* 28 (3)**,** S2-S9. doi: 10.1038/sj.ijo.0802804

Poulsen, P., Levin, K., Petersen, I., *et al.* (2005) Heritability of insulin secretion, peripheral and hepatic insulin action, and intracellular glucose partitioning in young and old Danish twins. *Diabetes,* 54 (1)**,** 275-283. doi: 10.2337/diabetes.54.1.275

Pourhoseingholi, M. A., Baghestani, A. R. & Vahedi, M. (2012) How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from Bed to Bench,* 5 (2)**,** 79-83. doi: 10.22037/ghfbb.v5i2.246

Prasad, R. B. & Groop, L. (2015) Genetics of Type 2 Diabetes—Pitfalls and Possibilities. *Genes,* 6 (1)**,** 87-123. doi: 10.3390/genes6010087

Purcell, S. & Chang, C. (2017) PLINK 1.9. Available: http://pngu.mgh.harvard.edu/purcell/plink/ [Accessed: 15 November 2019]

Purcell, S., Neale, B., Todd-Brown, K., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics,* 81 (3)**,** 559-575. doi: 10.1086/519795

Qu, H.-Q., Tien, M. & Polychronakos, C. (2010) Statistical significance in genetic association studies. *Clinical and Investigative Medicine,* 33 (5)**,** E266-E270

Ragoussis, J. (2009) Genotyping technologies for genetic research. *Annual Review of Genomics and Human Genetics,* 10**,** 117-133. doi: 10.1146/annurev-genom-082908-150116

Ranganathan, P., Aggarwal, R. & Pramesh, C. S. (2015) Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research,* 6 (4)**,** 222-224. doi: 10.4103/2229-3485.167092

Ren, J., Xiang, A. H., Trigo, E.*, et al.* (2014) Genetic variation in MTNR1B is associated with gestational diabetes mellitus and contributes only to the absolute level of beta cell compensation in Mexican Americans. *Diabetologia,* 57 (7)**,** 1391-1399. doi: 10.1007/s00125-014-3239-3

Rentzsch, P., Witten, D., Cooper, G. M.*, et al.* (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research,* 47 (D1)**,** D886-D894. doi: 10.1093/nar/gky1016

Retnakaran, R. (2017) Adiponectin and β-Cell Adaptation in Pregnancy. *Diabetes,* 66 (5)**,** 1121-1122. doi: 10.2337/dbi17-0001

Reyes-López, R., Pérez-Luque, E. & Malacara, J. M. (2014) Metabolic, hormonal characteristics and genetic variants of TCF7L2 associated with development of gestational diabetes mellitus in Mexican women. *Diabetes/Metabolism Research and Reviews,* 30 (8)**,** 701-706. doi: 10.1002/dmrr.2538

Robitaille, J. & Grant, A. M. (2008) The genetics of gestational diabetes mellitus: evidence for relationship with type 2 diabetes mellitus. *Genetics in Medicine,* 10 (4)**,** 240-250. doi: 10.1097/GIM.0b013e31816b8710

Rojano, E., Seoane, P., Ranea, J. A.*, et al.* (2018) Regulatory variants: from detection to predicting impact. *Briefings in Bioinformatics,* 20 (5)**,** 1639–1654. doi: 10.1093/bib/bby039

Rosta, K., Al-Aissa, Z., Hadarits, O.*, et al.* (2017) Association Study with 77 SNPs Confirms the Robust Role for the rs10830963/G of MTNR1B Variant and Identifies Two Novel Associations in Gestational Diabetes Mellitus Development. *PLoS ONE,* 12 (1)**,** e0169781. doi: 10.1371/journal.pone.0169781

Sawyer, S. L., Mukherjee, N., Pakstis, A. J.*, et al.* (2005) Linkage disequilibrium patterns vary substantially among populations. *European Journal of Human Genetics,* 13 (5)**,** 677-686. doi: 10.1038/sj.ejhg.5201368

Schmidt, M. I., Duncan, B. B., Reichelt, A. J., *et al.* (2001) Gestational diabetes mellitus diagnosed with a 2-h 75-g oral glucose tolerance test and adverse pregnancy outcomes. *Diabetes Care,* 24 (7)**,** 1151-1155. doi: 10.2337/diacare.24.7.1151

Schork, N. J., Murray, S. S., Frazer, K. A., *et al.* (2009) Common vs. Rare Allele Hypotheses for Complex Diseases. *Current Opinion in Genetics & development,* 19 (3)**,** 212-219. doi: 10.1016/j.gde.2009.04.010

Shaat, N., Karlsson, E., Lernmark, A., *et al.* (2006) Common variants in MODY genes increase the risk of gestational diabetes mellitus. *Diabetologia,* 49 (7)**,** 1545-1551. doi: 10.1007/s00125-006-0258-8

Shameer, K., Tripathi, L. P., Kalari, K. R., *et al.* (2016) Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Briefings in Bioinformatics,* 17 (5)**,** 841-862. doi: 10.1093/bib/bbv084

Shapiro, S. S. & Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika,* 52 (3)**,** 591-611. doi: 10.2307/2333709

Shen, T. H., Tarczy-Hornoch, P., Detwiler, L. T., *et al.* (2010) Evaluation of probabilistic and logical inference for a SNP annotation system. *Journal of biomedical informatics,* 43 (3)**,** 407-418. doi: 10.1016/j.jbi.2009.12.002

Shields, B., Hicks, S., Shepherd, M., *et al.* (2010) Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia,* 53 (12)**,** 2504-2508. doi: 10.1007/s00125-010-1799-4

Shin, H. D., Park, B. L., Shin, H. J., *et al.* (2010) Association of KCNQ1 polymorphisms with the gestational diabetes mellitus in Korean women. *Journal of Clinical Endocrinology & Metabolism,* 95 (1)**,** 445-449. doi: 10.1210/jc.2009-1393

Slatkin, M. (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics,* 9 (6)**,** 477-485. doi: 10.1038/nrg2361

Solomon, C. G., Willett, W. C., Carey, V. J., *et al.* (1997) A prospective study of pregravid determinants of gestational diabetes mellitus. *JAMA,* 278 (13)**,** 1078-1083. doi: 10.1001/jama.1997.03550130052036

Sonagra, A. D., Biradar, S. M., K, D., *et al.* (2014) Normal pregnancy- a state of insulin resistance. *Journal of Clinical and Diagnostic Research,* 8 (11)**,** CC01-CC3. doi: 10.7860/JCDR/2014/10068.5081

Storm, N., Darnhofer-Patel, B., Van Den Boom, D., *et al.* (2003) MALDI-TOF mass spectrometry-based SNP genotyping. In*:* P. Kwok, (ed.). *Nucleotide Polymorphisms.*

Methods in Molecular Biology™. Totowa, NJ: Springer, pp. 241-262. doi: 10.1385/1-59259-327-5:241

Szumilas, M. (2010) Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry,* 19 (3)**,** 227-229

Tam, V., Patel, N., Turcotte, M., *et al.* (2019) Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics,* 20 (8)**,** 467-484. doi: 10.1038/s41576-019-0127-1

Teare, M. D. & Barrett, J. H. (2005) Genetic linkage studies. *Lancet,* 366 (9490)**,** 1036-1044. doi: 10.1016/S0140-6736(05)67382-5

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

Tibazarwa, K., Ntyintyane, L., Sliwa, K., *et al.* (2009) A time bomb of cardiovascular risk factors in South Africa: results from the Heart of Soweto Study "Heart Awareness Days". *International Journal of Cardiology,* 132 (2)**,** 233-239. doi: 10.1016/j.ijcard.2007.11.067

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., *et al.* (2009) The genetic structure and history of Africans and African Americans. *Science,* 324 (5930)**,** 1035-1044. doi: 10.1126/science.1172257

Vaxillaire, M., Bonnefond, A. & Froguel, P. (2012) The lessons of early-onset monogenic diabetes for the understanding of diabetes pathogenesis. *Best Practice & Research Clinical Endocrinology & Metabolism,* 26 (2)**,** 171-187. doi: 10.1016/j.beem.2011.12.001

Vaxillaire, M. & Froguel, P. (2008) Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes. *Endocrine Reviews,* 29 (3)**,** 254-264. doi: 10.1210/er.2007-0024

Voight, B. F., Scott, L. J., Steinthorsdottir, V., *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics,* 42 (7)**,** 579-589. doi: 10.1038/ng.609

Walker, S. H. & Duncan, D. B. (1967) Estimation of the probability of an event as a function of several independent variables. *Biometrika,* 54 (1-2)**,** 167-179. doi: 10.1093/biomet/54.1-2.167

Watanabe, R. M. (2011) Inherited destiny? Genetics and gestational diabetes mellitus. *Genome Medicine,* 3 (3)**,** 18. doi: 10.1186/gm232

Weng, J., Ekelund, M., Lehto, M., *et al.* (2002) Screening for MODY mutations, GAD antibodies, and type 1 diabetes–associated HLA genotypes in women with gestational diabetes mellitus. *Diabetes Care,* 25 (1)**,** 68-71. doi: 10.2337/diacare.25.1.68

Wheeler, B. J., Patterson, N., Love, D. R.*, et al.* (2013) Frequency and genetic spectrum of maturity-onset diabetes of the young (MODY) in southern New Zealand. *Journal of Diabetes & Metabolic Disorders,* 12 (1)**,** 46. doi: 10.1186/2251-6581-12-46

Wilkening, S., Chen, B., Bermejo, J. L., *et al.* (2009) Is there still a need for candidate gene approaches in the era of genome-wide association studies? *Genomics,* 93 (5)**,** 415-419. doi: 10.1016/j.ygeno.2008.12.011

Witte, J. S. (2010) Genome-Wide Association Studies and Beyond. *Annual Review of Public Health,* 31**,** 9-20. doi: 10.1146/annurev.publhealth.012809.103723

World Health Organization (2013) Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy. Geneva, CH: World Health Organization. Available: https://www.who.int/diabetes/publications/Hyperglycaemia_In_Pregnancy/en/ [Accessed: 18 March 2018]

World Health Organization (2018) Obesity and Overweight. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [Accessed: 15 February 2020]

Wu, L., Cui, L., Tam, W. H., *et al.* (2016) Genetic variants associated with gestational diabetes mellitus: a meta-analysis and subgroup analysis. *Scientific Reports,* 6 (1)**,** 30539. doi: 10.1038/srep30539

Yang, S. & Du, Q. (2014) Association of GCK -30G> a polymorphism with gestational diabetes mellitus and type 2 diabetes mellitus risk: a meta-analysis involving 18 case-control studies. *Genet Test Mol Biomarkers,* 18 (5)**,** 289-98. doi: 10.1089/gtmb.2013.0427

Yang, Y. & Chan, L. (2016) Monogenic Diabetes: What It Teaches Us on the Common Forms of Type 1 and Type 2 Diabetes. *Endocrine Reviews,* 37 (3)**,** 190-222. doi: 10.1210/er.2015-1116

Zappala, Z. & Montgomery, S. B. (2016) Non-Coding Loss-of-Function Variation in Human Genomes. *Human Heredity,* 81 (2)**,** 78-87. doi: 10.1159/000447453

Zhang, C., Bao, W., Rong, Y.*, et al.* (2013) Genetic variants and the risk of gestational diabetes mellitus: a systematic review. *Human Reproduction Update,* 19 (4)**,** 376-390. doi: 10.1093/humupd/dmt013

Zhang, Y., Sun, C.-M., Hu, X.-Q., *et al.* (2014) Relationship between melatonin receptor 1B and insulin receptor substrate 1 polymorphisms with gestational diabetes mellitus: a systematic review and meta-analysis. *Scientific Reports,* 4 (1)**,** 6113. doi: 10.1038/srep06113

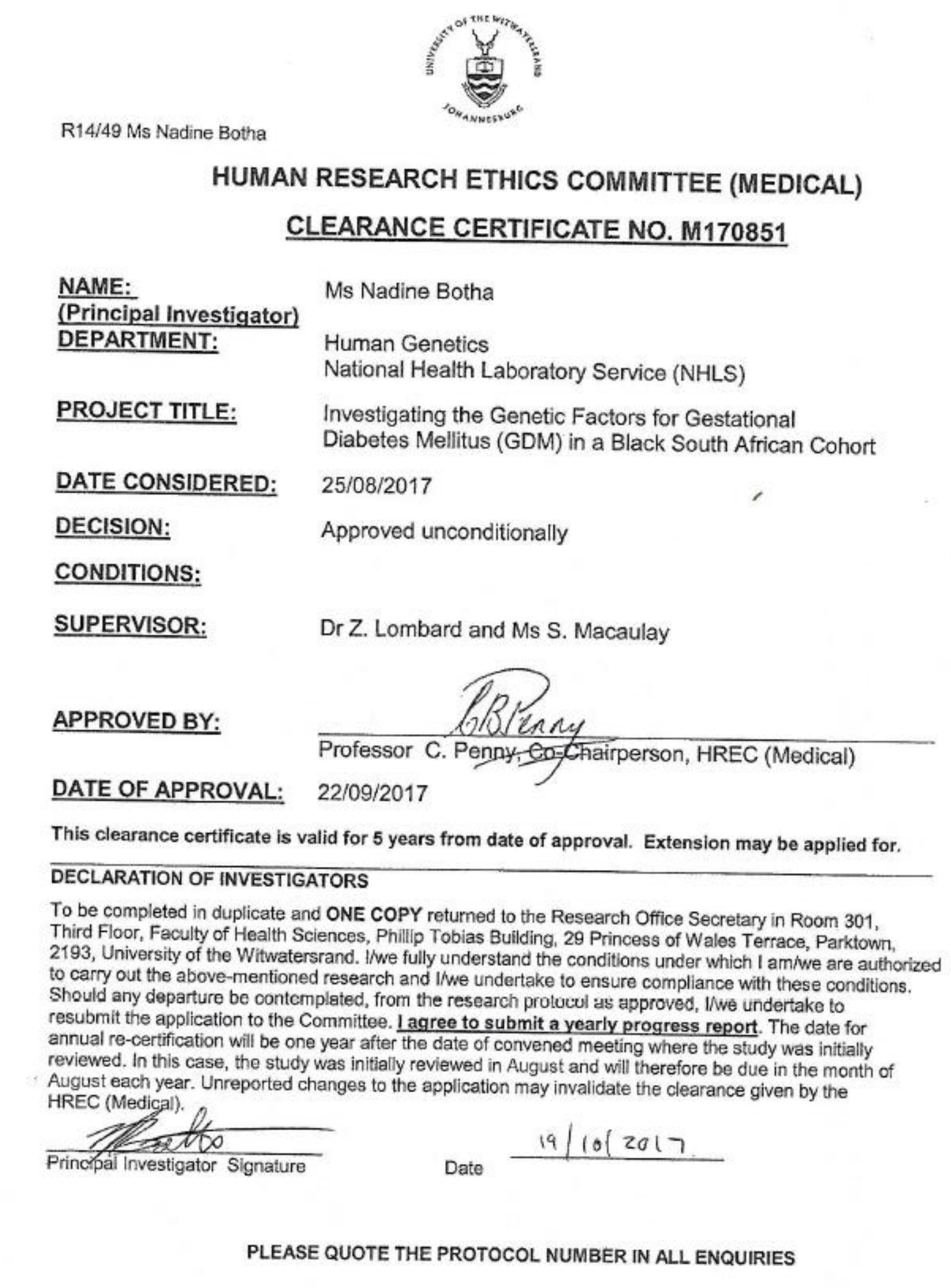
Zhou, Q., Zhang, K., Li, W., *et al.* (2009) Association of KCNQ1 gene polymorphism with gestational diabetes mellitus in a Chinese population. *Diabetologia,* 52 (11)**,** 2466–2468. doi: 10.1007/s00125-009-1500-y

Zhu, Y. & Zhang, C. (2016) Prevalence of Gestational Diabetes and Risk of Progression to Type 2 Diabetes: a Global Perspective. *Current Diabetes Reports,* 16 (1)**,** 7. doi: 10.1007/s11892-015-0699-x

Zondervan, K. T. & Cardon, L. R. (2007) Designing candidate gene and genome-wide case-control association studies. *Nature Protocols,* 2 (10)**,** 2492-2501. doi: 10.1038/nprot.2007.366

7 APPENDICES

7.1 APPENDIX A: ETHICS CLEARANCE CERTIFICATE



R14/49 Ms Nadine Botha

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170851

NAME: (Principal Investigator) Ms Nadine Botha

DEPARTMENT: Human Genetics
National Health Laboratory Service (NHLS)

PROJECT TITLE: Investigating the Genetic Factors for Gestational Diabetes Mellitus (GDM) in a Black South African Cohort

DATE CONSIDERED: 25/08/2017

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR: Dr Z. Lombard and Ms S. Macaulay

APPROVED BY: CBPenny

Professor C. Penny, Co-Chairperson, HREC (Medical)

DATE OF APPROVAL: 22/09/2017

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

DECLARATION OF INVESTIGATORS

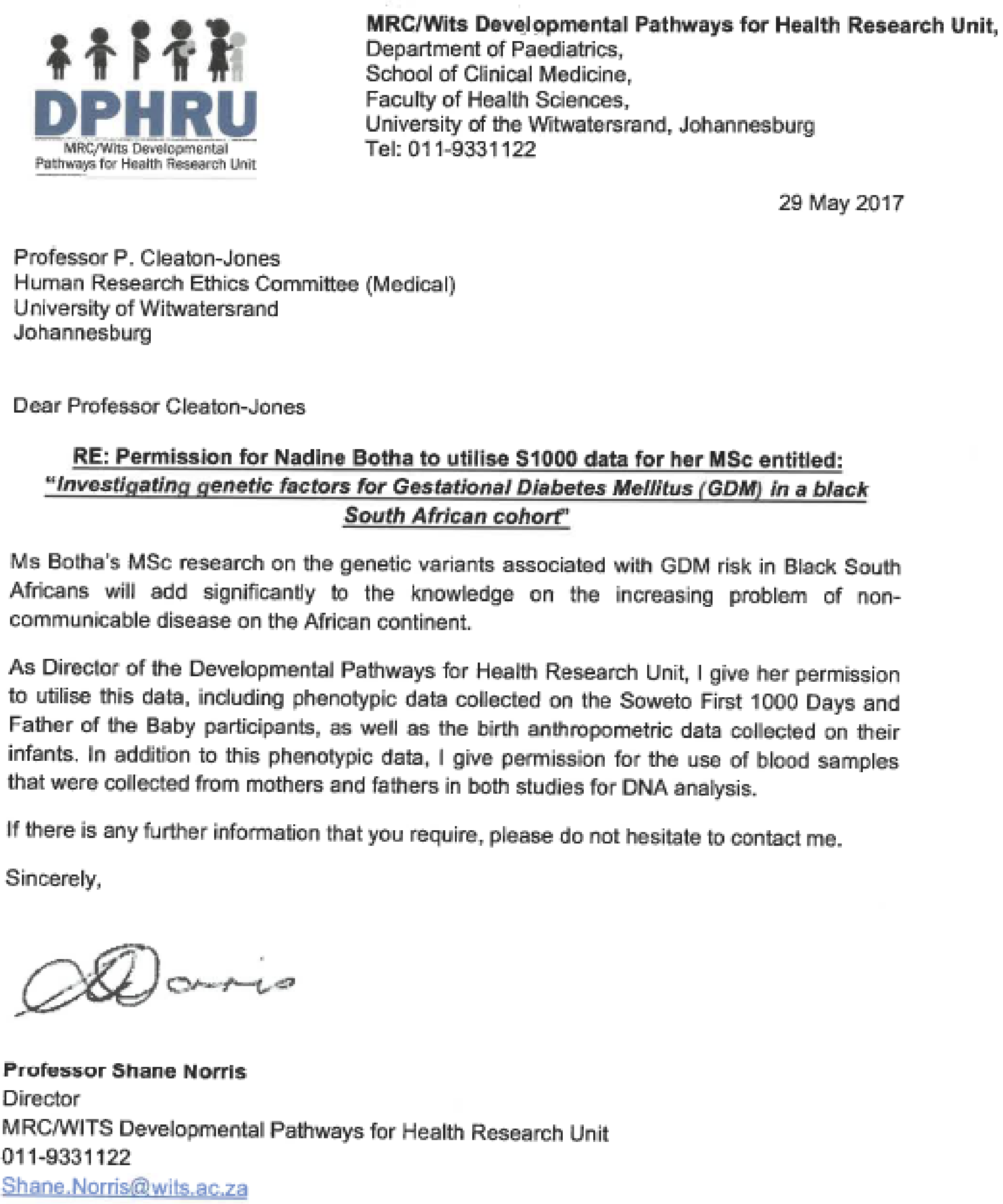
To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in August and will therefore be due in the month of August each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 19/10/2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

7.2 APPENDIX B: PERMISSION FOR THE USE OF DATA AND SPECIMENS



DPHRU
MRC/Wits Developmental
Pathways for Health Research Unit

MRC/Wits Developmental Pathways for Health Research Unit,
Department of Paediatrics,
School of Clinical Medicine,
Faculty of Health Sciences,
University of the Witwatersrand, Johannesburg
Tel: 011-9331122

29 May 2017

Professor P. Cleaton-Jones
Human Research Ethics Committee (Medical)
University of Witwatersrand
Johannesburg

Dear Professor Cleaton-Jones

RE: Permission for Nadine Botha to utilise S1000 data for her MSc entitled: "*Investigating genetic factors for Gestational Diabetes Mellitus (GDM) in a black South African cohort*"

Ms Botha's MSc research on the genetic variants associated with GDM risk in Black South Africans will add significantly to the knowledge on the increasing problem of non-communicable disease on the African continent.

As Director of the Developmental Pathways for Health Research Unit, I give her permission to utilise this data, including phenotypic data collected on the Soweto First 1000 Days and Father of the Baby participants, as well as the birth anthropometric data collected on their infants. In addition to this phenotypic data, I give permission for the use of blood samples that were collected from mothers and fathers in both studies for DNA analysis.

If there is any further information that you require, please do not hesitate to contact me.

Sincerely,

Professor Shane Norris
Director
MRC/WITS Developmental Pathways for Health Research Unit
011-9331122
Shane.Norris@wits.ac.za

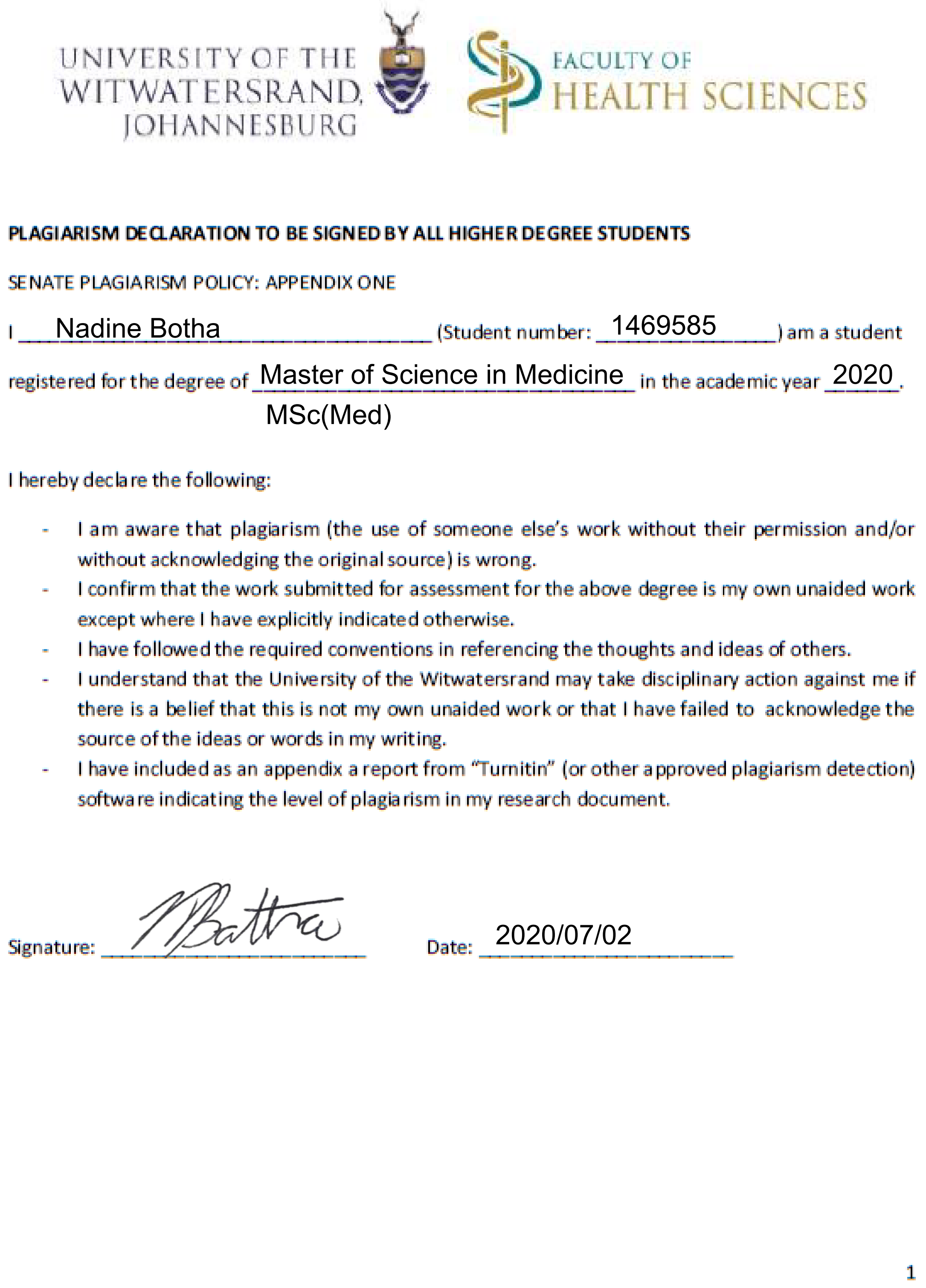
7.3 APPEDIX C: ALL GENOTYPE RESULTS

Summary statistics of the genotype results obtained from the 22 SNPs genotyped

Gene	rs number	Failure Rate	No. of failed samples	Minor Allele (A1)	HWE	MAF	GDM Positive					GDM Negative Controls				
							A1/A1 (n)	A1/A2 (n)	A2/A2 (n)	MAF	HWE	A1/A1 (n)	A1/A2 (n)	A2/A2 (n)	MAF	HWE
GCK	rs1799884[b]	0.03	7	T	1.00	0.20	2	28	48	0.21	0.50	7	48	100	0.20	0.62
	rs112257899	0.00	0	T	1.00	0.08	1	15	64	0.11	1.00	0	22	138	0.07	1.00
	rs758983[a]	0.06	14	T	0.53	0.20	1	29	47	0.20	0.28	6	47	96	0.20	1.00
	rs4607517[b]	0.03	8	A	1.00	0.07	0	6	71	0.04	1.00	1	22	132	0.08	1.00
TCF7L2	rs34347733[a]	0.06	14	T	0.79	0.15	0	16	60	0.10	1.00	4	43	103	0.17	1.00
	rs34872471	0.00	1	C	0.79	0.41	12	39	29	0.39	1.00	27	79	53	0.42	0.87
	rs7901695	0.00	0	C	0.20	0.50	15	49	16	0.49	0.07	39	82	39	0.50	0.87
	rs7903146[a b]	0.03	8	T	0.00	0.10	7	0	70	0.09	0.00	16	0	139	0.10	0.00
	rs115626858	0.05	12	T	0.70	0.09	1	14	63	0.10	0.58	1	23	126	0.08	1.00
	rs115758892	0.03	7	A	0.06	0.06	0	9	70	0.06	1.00	3	15	136	0.07	0.02
	rs12255372[a]	0.03	7	T	0.31	0.26	6	28	45	0.25	0.56	13	56	85	0.27	0.41
KCNJ11	rs5210	0.05	11	G	0.67	0.37	11	33	33	0.36	0.62	21	70	61	0.37	0.86
	rs5214	0.00	1	C	1.00	0.07	0	6	74	0.04	1.00	1	24	134	0.08	1.00
	rs5215	0.03	6	C	1.00	0.06	0	10	68	0.06	1.00	0	16	140	0.05	1.00
HNF1A	rs2244608	0.05	12	G	0.61	0.06	0	6	70	0.04	1.00	0	23	129	0.08	1.00
PDX1	rs61944006	0.02	4	C	0.06	0.26	4	40	36	0.30	0.12	6	62	88	0.24	0.27
	rs4581569	0.02	5	T	0.42	0.28	3	27	48	0.21	1.00	18	63	76	0.32	0.36
	rs73169687[a]	0.06	15	A	1.00	0.06	0	8	68	0.05	1.00	0	19	130	0.06	1.00
	rs4415872[a]	0.87	209	C	0.62	0.27				0.10	1.00				0.25	0.63
	rs7981781[a]	0.06	14	A	0.00	0.18	0	31	43	0.21	0.03	0	48	104	0.16	0.03
HNF 4A	rs80276513	0.01	3	A	0.00	0.07	4	6	70	0.09	0.00	2	15	140	0.06	0.10
	rs6031551	0.03	7	C	0.64	0.16	2	21	55	0.16	1.00	5	41	109	0.165	0.57

SNP = single nucleotide polymorphism, SNP ID = universal SNP identification tag (rs number) assigned by National Center for Biotechnology Information (NCBI), A1 = minor allele, A2 = major allele, A1/A1, A1/A2, A2/A2 (n) = genotype counts, MAF = Minor Allele Frequency, HWE = Hardy-Weinberg Equilibrium, [a] previously reported GDM-associated SNP

7.4 APPENDIX D: PLAGIARISM DECLARATION AND REPORT



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

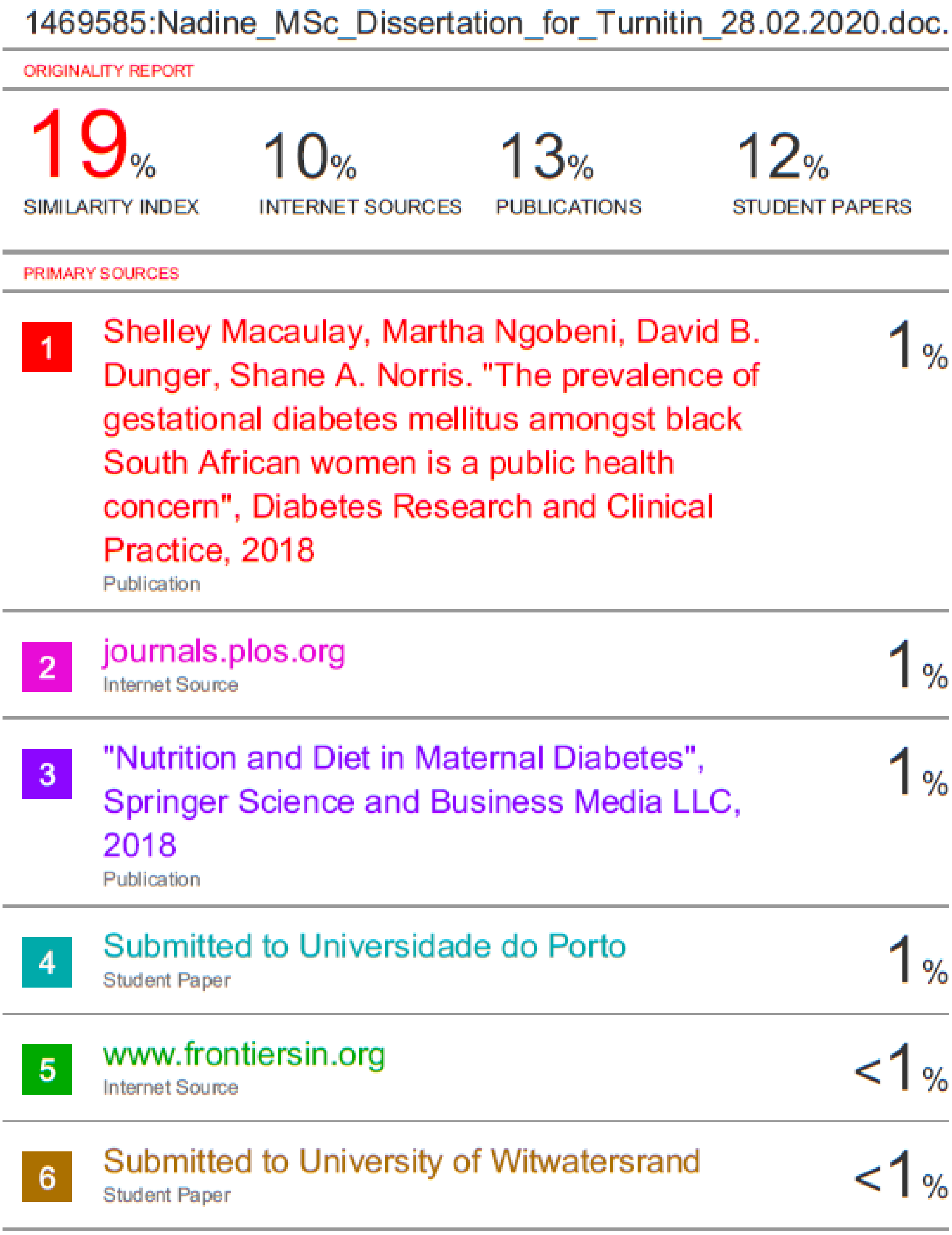
I Nadine Botha (Student number: 1469585) am a student registered for the degree of Master of Science in Medicine MSc(Med) in the academic year 2020.

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: NBotha Date: 2020/07/02

1



1469585:Nadine_MSc_Dissertation_for_Turnitin_28.02.2020.doc.

ORIGINALITY REPORT

19%	10%	13%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shelley Macaulay, Martha Ngobeni, David B. Dunger, Shane A. Norris. "The prevalence of gestational diabetes mellitus amongst black South African women is a public health concern", Diabetes Research and Clinical Practice, 2018 Publication	1%
2	journals.plos.org Internet Source	1%
3	"Nutrition and Diet in Maternal Diabetes", Springer Science and Business Media LLC, 2018 Publication	1%
4	Submitted to Universidade do Porto Student Paper	1%
5	www.frontiersin.org Internet Source	<1%
6	Submitted to University of Witwatersrand Student Paper	<1%

7	Submitted to Otago Polytechnic Student Paper	<1 %
8	currentprotocols.onlinelibrary.wiley.com Internet Source	<1 %
9	Submitted to University of Hull Student Paper	<1 %
10	online.boneandjoint.org.uk Internet Source	<1 %
11	www.nature.com Internet Source	<1 %
12	"Abstracts", Diabetologia, 1997 Publication	<1 %
13	www.eje-online.org Internet Source	<1 %
14	onlinelibrary.wiley.com Internet Source	<1 %
15	Methods in Molecular Biology, 2016. Publication	<1 %
16	Tasneem Khan, Shelley Macaulay, Shane A. Norris, Lisa K. Micklesfield, Estelle D. Watson. "Physical activity and the risk for gestational diabetes mellitus amongst pregnant women living in Soweto: a study protocol", BMC Women's Health, 2016	<1 %

NATIONAL HEALTH LABORATORY SERVICE

School of Pathology, University of the Witwatersrand

NATIONAL HEALTH LABORATORY SERVICE

DIVISION OF HUMAN GENETICS

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

Hospital Street, Johannesburg, 2001 I PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 I [M] +27 78 080 8841 I [F] +27 11 489 9226 I [E] human.genetics@nhls.ac.za

Faculty of Health Sciences, University of the Witwatersrand
Phillip V Tobias Health Sciences Building
29 Princess of Wales Terrace
Parktown
Johannesburg
2193

05 March 2020

To Whom It May Concern:

We have reviewed the Turnitin report for Ms Nadina Botha's MSc (Med) dissertation, and concur that there are no signs of overt plagiarism. On review of this report it shows that all highlighted citations only contain 1% or less overlap with the dissertation.

Please do not hesitate to contact us if you need any further input.

Kind Regards

SMacaulay

Zané Lombard, PhD

Principal Medical Scientist | Associate Professor
Division of Human Genetics
Tel: 011 489 9208 | Email: zane.lombard@wits.ac.za

Shelley Macaulay, PhD

Principal Medical Scientist | Senior Lecturer
Division of Human Genetics
Tel: 011 489 9243 | Email: shelley.macaulay@wits.ac.za

This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report