Sequencing for high-risk type 1 diabetes genotypes in the South African black population using AmpliSeq Nanopore next generation sequencing

by

Nomfundo Mathabela (1060104)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science in Medicine

in

Chemical Pathology

2023

In the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Supervisors: Dr Stuart Ali (60%), Dr Carolyn Padoa (20%), and Dr Eleanor Cave (20%)

Declaration

I, Nomfundo Ntombikayise Mathabela, declare that this research report, submitted in partial fulfilment for the degree of Master of Science in Medicine in Chemical Pathology at the University of the Witwatersrand, is my own work and has not been submitted before for any degree or examination at any other University.

_____

Nomfundo Mathabela

Signed on the ____31_____of August 2023 at ____The University of the Witwatersrand___

Plagiarism Declaration

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I, Nomfundo Mathabela, (Student number: 1060104) am a student registered for the degree of MSc (Med) Chemical Pathology in the academic year 2020-2023.

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:  Nomfundo Mathabela

Signed on the 31 of August 2023 at The University of the Witwatersrand

Abstract

Introduction: Type 1 diabetes (T1D) is a chronic autoimmune disorder characterized by the destruction of the β-cells of the pancreas, resulting in the inability to produce/maintain insulin. This results in an inability to maintain the blood glucose at homeostasis. Prolonged hyperglycaemia leads to micro- and macro-vascular complications. Thus, it is vital to diagnose and treat patients in a timely manner. It is important to identify individuals at increased risk of developing T1D allowing for appropriate follow ups. Numerous mutations/variants in specific genes confer an increased risk of T1D with the HLA gene accounting for approximately 40-50 % of the risk. Therefore, it is possible that by looking at genetic variation in T1D associated genes we can develop a tool that can determine the likelihood of an individual developing T1D.

Study aims and objectives: The development, implementation, and validation of a Nanopore NGS method to sequence and genotype polymorphisms associated with T1D in a cohort of black South Africans which could be used to develop a genetic risk score (GRS) for T1D in this population. In addition, we aimed to compare sequencing results to two HLA genotypes obtained using PCR-RFLP.

Method: Participants with T1D (n=19) and control participants (n=5) were genotyped for 12 T1D associated polymorphisms through Ampliseq Nanopore sequencing. In addition, Ion Torrent was used to confirm the Ampliseq Nanopore results. A bioinformatics pipeline that involves reference sequence generation using an in-house script, alignment of sequencing data with the reference sequence, filtering and variant calling was developed. A genetic risk score (GRS) was calculated for the participants. Participants (n=73) were genotyped by PCR-RFLP for the HLA rs2040410 and rs7454108 polymorphisms.

Results: Sequencing samples individually was found to have a slightly higher Qscore than samples sequenced in multiplex (9.73 vs. 9.5). Samples sequenced individually had higher average reads (187.60 vs. 151.14 Mb), passed reads (41.47 vs. 25.99 Mb), and estimated bases (54.72 vs. 49.57 Mb) than those sequenced in multiplex. In addition, samples sequenced in a multiplex had higher average failed reads (475.58 Mb) in comparison to those sequenced individually (13.58 Mb).

The average percentage difference in sequencing data generated using Ion Torrent compared to Nanopore was 5.67%. Variant calling produced average Phred-scale quality scores of 73.89 for standards (The Coriell Trio) and 89.77 for participants. The GRS calculator was not able to accurately predict which participants had T1D.

Discussion and Conclusion: A next generation sequencing method and bioinformatics pipeline for the screening of participants for 645 T1D associated polymorphisms was investigated. The method combined two sequencing techniques i.e., Ion AmpliSeq and Nanopore sequencing to achieve this. The data can then be processed by in-house variant callers. With a larger sample group, this method will be useful for the investigation of genetic variants linked to T1D.

Acknowledgements

I would like to thank the following individuals and organisations:

To my supervisors Dr Stuart Ali, Dr Carolyn Padoa and Dr Eleanor Cave I would like to thank you for helping me with my research project as my supervisors and providing support since the commencement of the research project.

To Dr Jonathan Featherston I would like to thank you for your help and guidance for the bioinformatics analysis. My limited knowledge in comparison to yours never made you view me as inadequate or a nuisance when I came to you for advice and input. You always encouraged and reassured me that I was on the right track and that I am doing better than I thought I was. You taught me to be patient and to believe in myself and my abilities and for that I am eternally grateful.

To The National Health Laboratory Research Trust and the Medical Research Council for funding and Akili Labs (Pty) Ltd for providing a laboratory facility for Nanopore sequencing work that was completed to contribute towards the research project.

To my family and friends for their support and prayers, especially my mother Busisiwe Mathabela, Nonhlanhla Segololo, Gabriel Laranjeira, Avani Baruthram, Daniesha Govender and Palesa Seemi. You were always there when I needed to cry when things were not going well and particularly when I was experiencing problems that were outside my control. You shared in all the frustration and encouraged me to hang in there when I wanted to give up. I appreciate all the patients and time that you invested.

And finally God, with whom without, all this would not be possible.

Table of contents

List of abbreviations

A	Adenine
AMX	Adapter mix
APC	Antigen presenting cells
ASIC	Application-specific integration circuit
AUC	Area under the curve
C	Cytosine
CDC	Complement-dependent cytotoxicity
CTL	Cytotoxic T cells
DR3	DRB1*03:01-DQA1*05:01-DQB1*02:01
DR4	DRB1*04:01/02/04/05/08-DQA1*03:01-DQB1*03:02/04
EB	Elution buffer
EDTA	Ethylenediaminetetraacetic acid
FB	Flush buffer
FLT	Flush tether
G	Guanine
GAD65	65 kilo Dalton isoform of glutamic acid decarboxylase
Gb	Gigabyte
GLUT	Glucose transporter
GRS	Genetic risk score
GWAS	Genome Wide Association Studies
HbA1C	Glycated haemoglobin A1c
HLA	Human Leukocyte Antigen
IA-2	Protein tyrosine phosphatase
IAA	Antibodies to insulin
ICA	Islet cell autoantibodies

IFN	Interferon
IGV	Integrated genomic viewer
IL	Interleukin
L	Litre
LADA	Latent autoimmune diabetes in adults
LB	Loading beads
LFB	Long fragment buffer
LNB	Ligation buffer
LRS	Long read sequencing
MHC	Major histocompatibility complex
MODY	Maturity-onset diabetes of the young
NCBI	National centre for biotechnology information
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
ROC	Receiver operating characteristics
RTG	Real genomic analysis tools
SD	Standard deviation
SFB	Short fragment buffer
SNP	Single Nucleotide Polymorphism
SQB	Sequencing buffer
SSP	Sequence Specific Primers
T	Thymine
T1D	Type 1 diabetes
T2D	Type 2 diabetes
Tb	Terabyte
TBE	Tris-Borate-EDTA
TCR	T cell receptor

Th0	T-helper cells
Th1	Pro-inflammatory cells
TNF	Tumour necrosis factor
U	Uracil
UCSC	University of California Santa Cruz
USA	United States of America
vp1	Enteroviral protein 1
WES	Whole exome sequencing
WGS	Whole genome sequencing
Zn8	Zinc transport 8
α	Alpha
β	Beta
γ	Gamma
±	Plus or minus
µ	Micro
∞	Infinite hold

List of Figures

List of Tables

List of Equations

1 Chapter 1: Introduction

1.1 The classification of T1D

Type 1 diabetes (T1D) is a chronic autoimmune disorder characterised by hyperglycaemia which results from the destruction of the insulin secreting beta cells (β-cells) of the pancreas (1). T1D can be classified as either type 1A (autoimmune) or type 1B (idiopathic) diabetes. A disease is considered an autoimmune disorder if it meets three criteria, namely: the presence of defined autoantigens and autoantibodies, passive transfer of T-lymphocytes, which leads to disease development, and successful immunomodulation of disease (2). Seventy to ninety percent of people with T1D have type 1A diabetes as they have immunological, self-reactive autoantibodies (3) whereas the remainder of people with T1D have type 1B since their specific pathogenesis remains unclear (4). T1D is one of the most prevalent chronic diseases of childhood (5,6). In the past, T1D was thought to only present in early childhood and adolescence however, it is now known it can present at any age (7).

1.2 Symptoms, diagnosis, and treatment of T1D

Symptoms of T1D include polyphagia (excessive hunger), polydipsia (excessive thirst), polyuria (excessive urination), blurred vision, fatigue, and weight loss (1) and appear when approximately 80 % of the β-cells have been destroyed.

The diagnosis of T1D can be challenging especially amongst adults as it can be misclassified as other forms of diabetes such as type 2 diabetes (T2D), latent autoimmune diabetes in adults (LADA) and maturity-onset diabetes of the young (MODY). The criteria for diagnosing T1D include a fasting blood glucose level ≥7 mmol/L, a random blood glucose level ≥11.0 mmol/L, an abnormal 2 hour oral glucose-tolerance test (≥11.1 mmol/L), or a glycated haemoglobin (HbA1C) ≥6.5% (8). Greater than 90 % of individuals newly diagnosed with T1D have one or more autoantibodies associated with T1D. These autoantibodies are targeted to glutamic acid decarboxylase 65, insulin, zinc transporter 8, and insulinoma-associated autoantigen 2 (9). These autoantibodies can be present months or years before disease onset. The number of

autoantibodies and age at which they first appear are predictive of age at diagnosis of T1D (10). Accurate diagnosis of T1D is crucial to ensure the correct treatment and possibly survival of an affected person.

Lack of treatment and/or poor glycaemic control can lead to microvascular (retinopathy, neuropathy, and nephropathy), macro-vascular (cardiovascular disease and stroke), and other (diabetic coma, hypoglycaemia) complications (11). Insulin pumps, multiple daily insulin injections and continuous glucose monitoring are being used currently for the treatment of T1D. However, treatments that imitate the normal functioning of the pancreas (sensor-augmented pump therapy) are desirable as they ensure an automated form of treatment that is less reliant on patient monitoring and administration of insulin (12). These methods have been found to decrease the HbA1c level substantially (12). Future treatments such as the use of an artificial pancreas (closed-loop system), insulin analogues and pluripotent stem cells (13) are currently being investigated.

1.3 The incidence and prevalence of T1D

T1D accounts for 5-10 % of individuals diagnosed with diabetes and is said to be more prevalent in Caucasian populations (14). T1D was initially said to peak at puberty (10-14 years) in the European population but in recent years the incidence of T1D has increased with the largest increase seen in young children (0-4 years) especially in developed countries (15,16). These findings differ from the South African black population where two peak ages of onset have been reported: an early peak at 14-17 years and a later peak at 22-23 years of age (17,18).

A recent systematic review and meta-analysis (19) of the global prevalence and incidence of T1D (Table 1.1) was conducted using 193 articles published between 1990 and 2019. The paper grouped all articles geographically into Asia, Africa, Europe, and America.

Table 1.1: Incidence and prevalence of T1D in Asia, Africa, Europe, America and the world (19)

Continent	Incidence	Prevalence
Asia	15 per 100 000 population	6.9 per 10 000 people
Africa	8 per 100 000 population	3.5 per 10 000 people
Europe	15 per 100 000 population	12.2 per 10 000 people
America	20 per 100 000 population	12.2 per 10 000 people
World	15 per 100 000 population	9.5 per 10 000 people

Furthermore, they showed that the incidence of T1D has increased over time (Figure 1.1) (19). The reason for this increase over time is not fully understood but is thought to be as a result of changes in environmental factors such as viral influence, early life feeding patterns, gut microbiome, childhood growth and increased obesity, perinatal risk factors, and vitamin D deficiency (20).

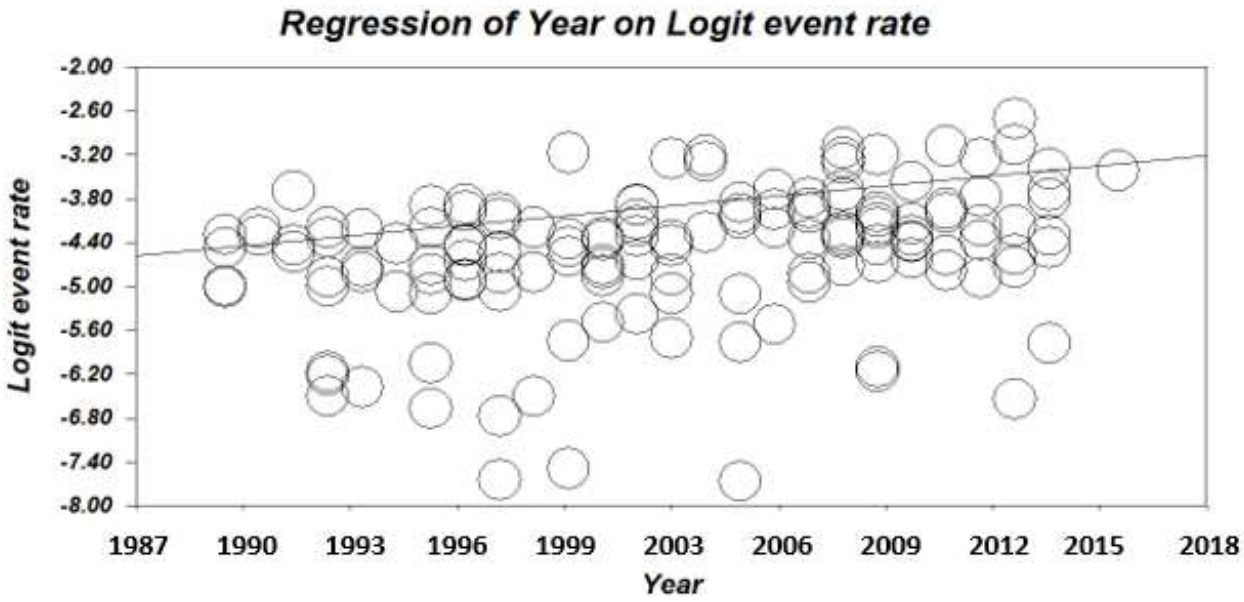


Figure 1.1: A meta-regression plot of the incidence of T1D over time. Data shown includes seven African (3.7 %), 111 European (58.7 %), 42 American (22.2 %) and 29 Asian (15.3 %) studies. This

further emphasizes the need for more research and data from African populations especially Southern African populations (19).

1.4 T1D aetiology

The exact aetiology and pathogenesis of T1D is still unknown but is thought to involve autoantigens released by the β-cells of the pancreas as shown in Figure 1.2. The autoantigens are presented to naïve T-helper (Th0) cells via the major histocompatibility complex class II molecules (MHC II) on antigen presenting cells (APCs); triggering the secretion of interleukin (IL)-12. This results in the differentiation of Th0 cells into pro-inflammatory Th1 cells. The Th1 cells secrete interferon gamma (IFN-γ) which leads to activation of macrophages into cytotoxic macrophages that are mobilized to the β-cells where they release chemokines such as free radicals, tumour necrosis factor alpha (TNF-α), and IL-1β that aid in the destruction of the β-cells. Concurrently, Th1 secretes IL-2 which leads to the differentiation of pre-cytotoxic T-cells to cytotoxic T-cells and mobilisation of these cells to the β-cells resulting in destruction of the β-cells by secretion of granzymes and perforins as well as via Fas-mediated apoptosis. This results in hyperglycaemia and subsequent development of T1D (21).

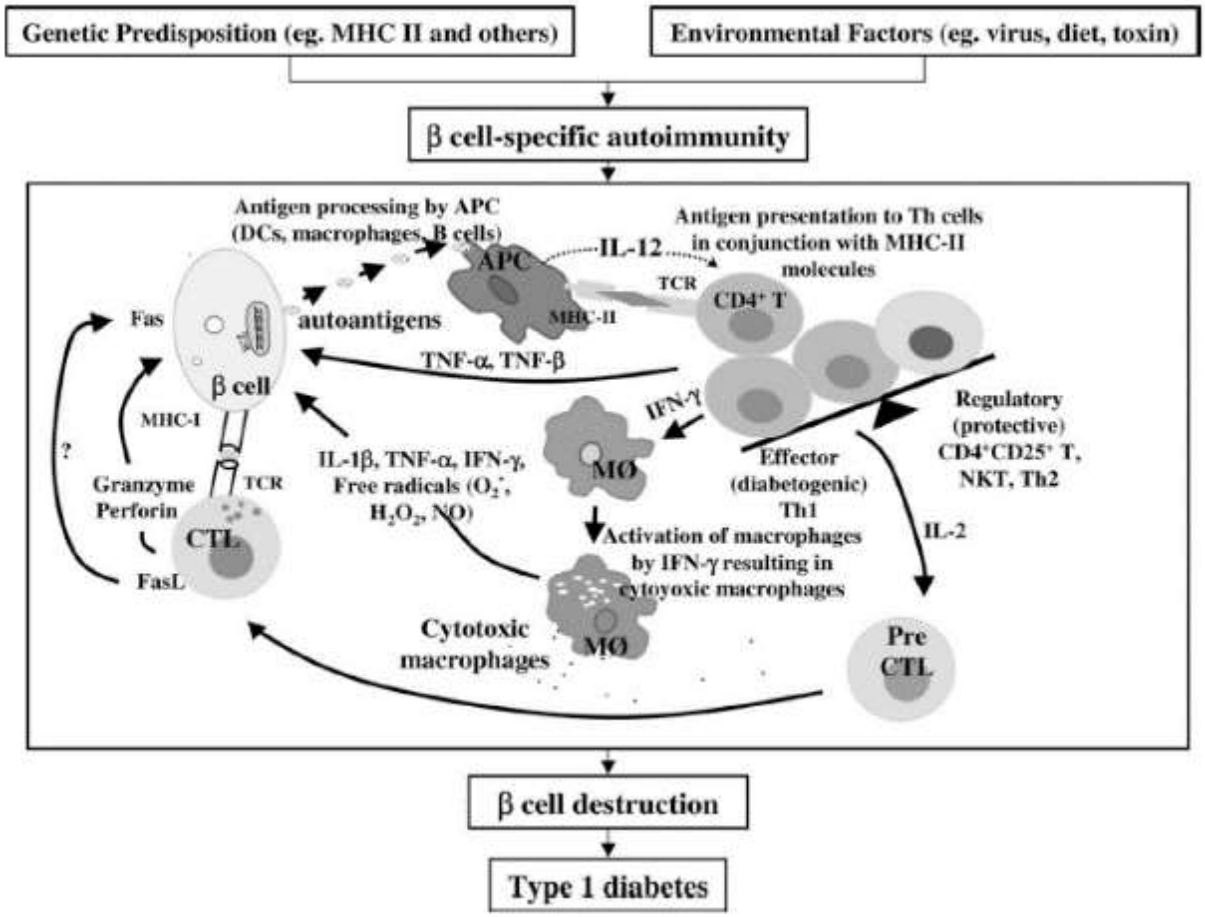


Figure 1.2: The mechanism by which autoimmune destruction of the insulin secreting β-cells of the pancreas occurs. An environmental trigger in genetically susceptible individuals leads to the differentiation of CD4+ Th0 cells into Th1 cells. This occurs when APC present β-cell specific autoantigens via MHC class II molecules. Th1 cells secrete cytokines that activate macrophages and cytotoxic T-cells that release perforin, granzymes and cytokines that are toxic to β-cells. Together with Fas-mediated apoptosis this leads to the destruction of the insulin secreting β-cells. The destruction of the β-cells thus prevents the production of insulin and uptake of glucose by the cells of the body. This essentially leads to the high glucose levels found in the body and thus a diabetic pathology (21).

1.5 Genetic susceptibility of T1D

Several variations in specific genes have been shown to contribute to the aetiology and susceptibility of T1D, as evidenced by Genome Wide Association Studies (GWAS) (22–25). However, these associations may differ based on the population group studied.

1.5.1 The human leukocyte antigen (HLA) genes

The Human Leukocyte Antigen (*HLA*) genes are a cluster of about 128 genes located on the short arm of chromosome 6 (26) that participate in the adaptive immune system in the human body. These genes code for the HLA class I (HLA I) or class II (HLA II) antigens that are displayed on the surface of APCs such as macrophages, dendritic cells, B cells, Langerhans cells, and Kupffer cells. These antigens are crucial as they are recognized by the T lymphocytes through their T-cell receptors (TCR) thus facilitating immune homeostasis (27). There are two major subtypes of T lymphocytes, CD8+ cytotoxic T-cells (CTLs), which recognize HLA I molecules and CD4+ T helper cells, which recognize HLA II molecules. The HLA alleles have been found to confer the highest risk of developing T1D, with the HLA class II alleles accounting for 40-50 % of the genetic risk of T1D (28).

The HLA class I and class II antigens are divided into three sub-classes each, namely: HLA-A, -B and -C and HLA-DR, HLA-DQ, and HLA-DP, respectively. The respective positions of the subclasses on chromosome 6 are shown in Figure 1.3 (26).

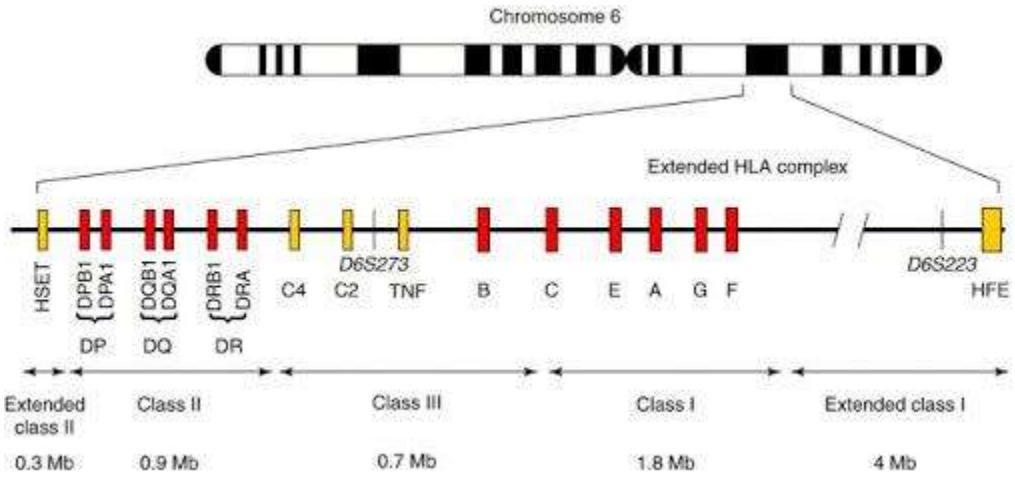


Figure 1.3: Positions of HLA classes (Class I, Class II, and Class III), subclasses (DP, DQ, and DR) and subtypes (DPB1, DPA1, DQB1, DQA1, DRB1, DRA, B, C, E, A, G, and F) on chromosome 6 and their respective sizes (29)

The HLA alleles are further grouped within the classes into supertypes consisting of molecules with similar peptide binding features (30). Thus, peptides that can bind to one allele within the supertype can also bind to the other alleles within the same supertype. HLA I is grouped into 12 supertypes, namely: HLA-A1, -A2, -A3, -A24, -B7, -B27, -B44, -B58, -B62, -A26, -B8, and -B39 (31). Class II molecules are more polymorphic and structurally complex than class I molecules, and were initially grouped into twelve supertypes, namely: five DRs (DR1, DR3, DR4, DR5, and DR9), three DQs (DQ1, DQ2, and DQ3), and four DPs (DPw1, DPw2, DPw4, and DPw6) (32). Subsequently, seven HLA II supertypes were defined based on protein-structural supertype classification. All the molecules belonging to the DP, DQ and DRB1 genetic locus are grouped according to supertypes as outline in Table 1.2 (33).

Table 1.2: Supertype grouping of HLA loci (33)

HLA locus	No. of supertypes	Supertype
DP	1	• DPB1*0101, DPB1*0201, DPB1*0401, DPB1*0401, DPB1*0402, DPB1*0501, DPB1*1401
DQ	2	• DQB1*0301, DQB1*0302, DQB1*0401 • DQB1*0201, DQB1*0501, DQB1*0602
DR	4	• DRB1*0401, DRB1*0405, DRB1*0802, DRB1*1101 • DRB3*0101, DRB3*0202 • DRB1*0301, DRB1*1302 • Containing the remaining DR proteins

The HLA region is known to have a high degree of linkage disequilibrium thus many of the alleles are inherited together, and a specific combination of these alleles, termed the haplotype, can either cause susceptibility, protection or be neutral with respect to T1D thus facilitating screening for these susceptible genotypes (34,35). The DR-DQ haplotypes most known to confer susceptibility and protection to T1D across different populations have been summarised in Table 1.3 (36). The DR3 and DR4 haplotypes have been found to be the most important contributors to T1D risk in the European and African populations (32); with DR3/DR4 heterozygosity conferring a greater risk than DR3 or DR4 homozygosity (34).

Table 1.3: The most susceptible and protective haplotypes for T1D (36)

Susceptible haplotypes	Protective haplotypes
• DRB1*0301-DQA1*0501-DQB1*0201 (DR3) • DRB1*0405-DQA1*0301-DQB1*0302 (DR4)	• DRB1*1501-DQA1*0102-DQB1*0602 (DR2) • DRB1*1401-DQA1*0101-DQB1*0503 (DR14) • DRB1*0701-DQA1*0201-DQB1*0303 (DR7)

1.5.2 HLA typing

There are various methods that have been used for typing HLA alleles in South Africa and internationally including serology-based methods (complement-dependant cytotoxicity test) (37), and molecular based approaches such as phototyping of known HLA variants using PCR based sequence specific primers (PCR-SSP) (23). While PCR based methods are suitable for screening large numbers of samples, they are problematic as reactions are known to fail and can result in having to repeat screening. PCR reaction failures could be due to primer dimer formation, the presence of weak bands due to DNA impurities, failure of control primer amplification, false negative and false positive detection of amplicons, multiple PCR reactions for the amplification of each region, as well as non-specific amplification (38). False negatives could possibly result from sequence variation in the HLA gene at the site where the primer binds, which is not unlikely given the high sequence variability within the *HLA* gene. To overcome some of these issues, a PCR restriction fragment length polymorphism (PCR-RFLP) method was developed that is both time and cost effective. Two HLA SNPs (rs2040410 and rs7454108) are used to predict whether individuals have the high risk DR3 and DR4 alleles (34,35).

1.5.3 PCR-RFLP HLA typing using rs2040410 and rs7454108

The rs2040410 (G>A) and rs7454108 (T>C) SNPs were investigated in four different population groups (consisting of 90 individuals of African, 274 of European, 44 of Japanese, and 45 of Chinese descent) where it was found that the rs2040410 A allele was associated with the HLA DR3 allele while the rs7454108 C allele was associated with the HLA DQB3 (DR4) allele (35). A second study conducted using data from European and American participants from six different cohorts, confirmed the findings of de Bakker and colleagues. They found that the rs2040410 A allele was present in 96.8 % (2291 of 3018) of participants with the DR3 allele and that the rs7454108 C allele was present in 98.9 % (3315 of 3353) of participants with the DQB3 allele (34). They further found that the two polymorphisms could be used to accurately (>99 %) identify individuals with the heterozygous DR3/DR4 genotype. They found that the majority (94.1 %; 1121 of 1191) of individuals with the DR3/DR4 genotype had the GA/TC genotype for rs2040410 and rs7454108,

respectively. A further 4.6 % (52 of 1191) of participants heterozygous for the DR3/DR4 alleles had the AA/TC genotype for the two polymorphisms. In contrast, of 3828 individuals that were not DR3/DR4, only 12 (0.3 %) had the GA/TC genotype and 1 (0.03 %) individual had the AA/TC genotype. Therefore, they concluded that the GA/TC and AA/TC genotypes can be used as markers to identify individuals heterozygous for DR3/DR4 alleles with a 98.5 % sensitivity and 99.7% specificity (34). The relative positions of these SNPs as well as the distance between them is outlined in Figure 1.4.

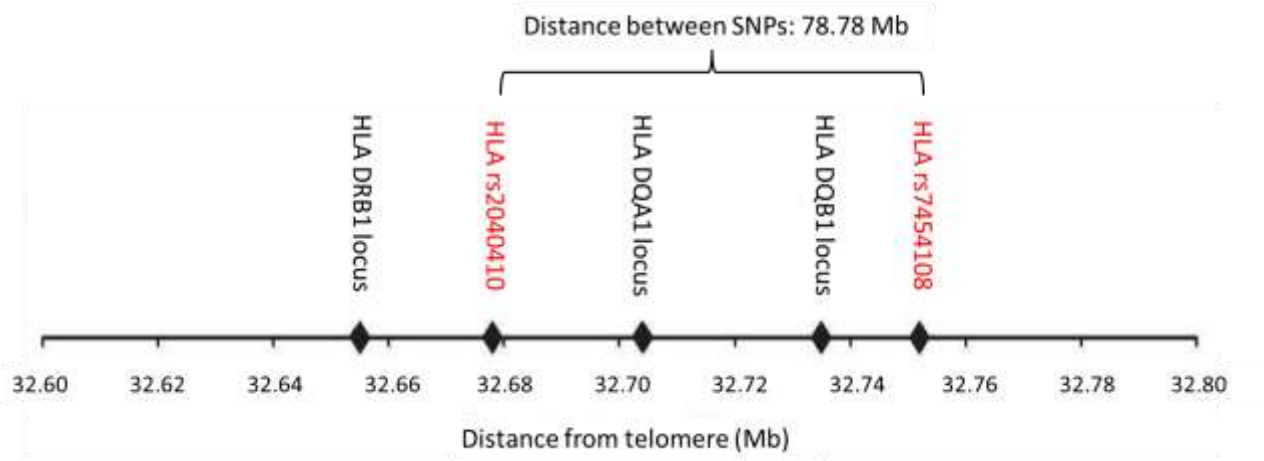


Figure 1.4: The position of rs2040410 and rs7454108 (shown in red) on chromosome 6 in relation to the HLA-DRB1, DQA1 and DQB1 genes, their positions on the chromosome (32.68 and 32,75 Mb, respectively) and the distance between them (78.78 Mb) Adapted from de Bakker and colleagues (35).

This method focuses on two SNPs predicted to confer risk through their association with the high-risk heterozygous (DR3/DR4) genotype. This may not be appropriate in the South African population as there might be other SNPs that give a higher genetic risk. There is a recent study that highlights the association between SNP rs7990 with HLA-DQA1 in an East Indian population (39). This study shows that it is important to develop screening tools which are population specific to accurately identify people who are at risk for development of T1D, in this case, screening tools that are specific to the South African population. A sequencing-based approach is a more accurate method for screening samples for known and novel variants related to T1D.

1.6 Next generation sequencing as a screening method for T1D

Next generation sequencing (NGS), which is otherwise referred to as second generation sequencing or massive parallel sequencing is a high-throughput sequencing technique that uses the concept of massively parallel processing (40). Massively parallel processing uses sequencing by synthesis technology by recording which labelled nucleotide is added as the chain is elongated. (41). This technology is used to determine the sequence of nucleotides in DNA and RNA. NGS can be used for whole exome sequencing (WES), whole genome sequencing (WGS), targeted regions (SNPs or small number of genes) and most recently long read sequencing (LRS) (41). Technologies such as Illumina and Ion Torrent are considered sequencing gold standards as they offer high read depth because of their high throughput, they also offer high quality scores (Qscore); a Qscore of 20 represents an error rate of 1 in a 100 bases sequenced and a 99 % call accuracy (40,41). These methods, however, are capital intensive, and require specialist equipment and technical expertise. What is needed is a simple, cost-effective method that uses the power of NGS. Nanopore sequencing will be used in this research as an NGS technique. Nanopore sequencing was chosen specifically for this study as it allows for flexibility in terms of the facilities and "man-power" needed. The sequencing machinery is compact and easily transportable and thus can be used in remote South African locations allowing easy access to NGS. A comparison of the Illumina, Ion Torrent, and Oxford Nanopore Technologies (ONT) sequencing platforms is outlined in Table 1.4 (42–44).

Table 1.4: Read length, data output, maximum run time, quality, and average score for different sequencing platforms as well as a cost per sample comparison (42–44)

Platform	Read length	Data output	Max run time (hours)	Quality	Approximant cost of instrument ($ US)	Cost per sample ($ US)
Illumina (NovaSeq 6000)	300 (paired end)	6 Tb	44	Q34	1 000 000	125.00
Ion Torrent (Ion GeneStudio S5 Prime)	600 (single end)	50 Gb	12	>Q30	100 000	116.00
ONT (PromethION)	4 Mb	14 Tb	72	Q34	230 000	100.00
ONT (MinION)	150 k	50 Gb	72	Q31	4 900	

*Paired end refers to sequencing that is done on both directions of the DNA strand whereas single end refers to sequencing that is done in one direction. Gb- Gigabyte, Tb- terabyte.

1.7 Oxford Nanopore as a screening and sequencing tool for T1D

Oxford Nanopore sequencing is a fourth-generation real-time sequencing technology that relies on charges generated during DNA/RNA translocation through a membrane bound protein nanopore. The translocation of bases through the nanopore results in a disruption of the current across the membrane resulting in a charge spike being generated; these charges are used for basecalling, which is a process that identifies nucleotides at specific locations throughout the input sequence (45).

The Oxford Nanopore sequencing has been used to determine the HLA type of tissues used for organ transplantation (46–48). ONT uses genomic DNA which is then sequenced using an Oxford Nanopore MinION sequencer. Genomic DNA can be labelled with a unique DNA barcode, and then pooled for sequencing to reduce costs to less than $100 per sample (46). Nanopore sequencing is ideal for long read data (up to 50 Gigabytes of data) with the MinION (49).

1.7.1 Mechanism of Nanopore sequencing

High molecular weight genomic DNA is needed to initialise library preparation. This DNA can then be fragmented (optional) into desired sizes. PCR amplification is not needed for sequencing, this ensures higher accuracies and confidence in basecalling when sequencing (50). The DNA is then prepared for adapter ligation through nick repair by blunting the ends. The adapters are then bound to the 5' end of the DNA, this allows for binding of the DNA to an enzyme to facilitate unidirectional translocation into the membrane-bound nanopore. The adapters also facilitate the aggregation of DNA to the membrane thus ensuring that the DNA can be captured into the nanopore (42). Barcodes, which allow for the identification of the individual DNA sequence, can be added allowing for multiplexing. This results in a library ready for loading onto a flow cell. This process is summarised in Figure 1.5.

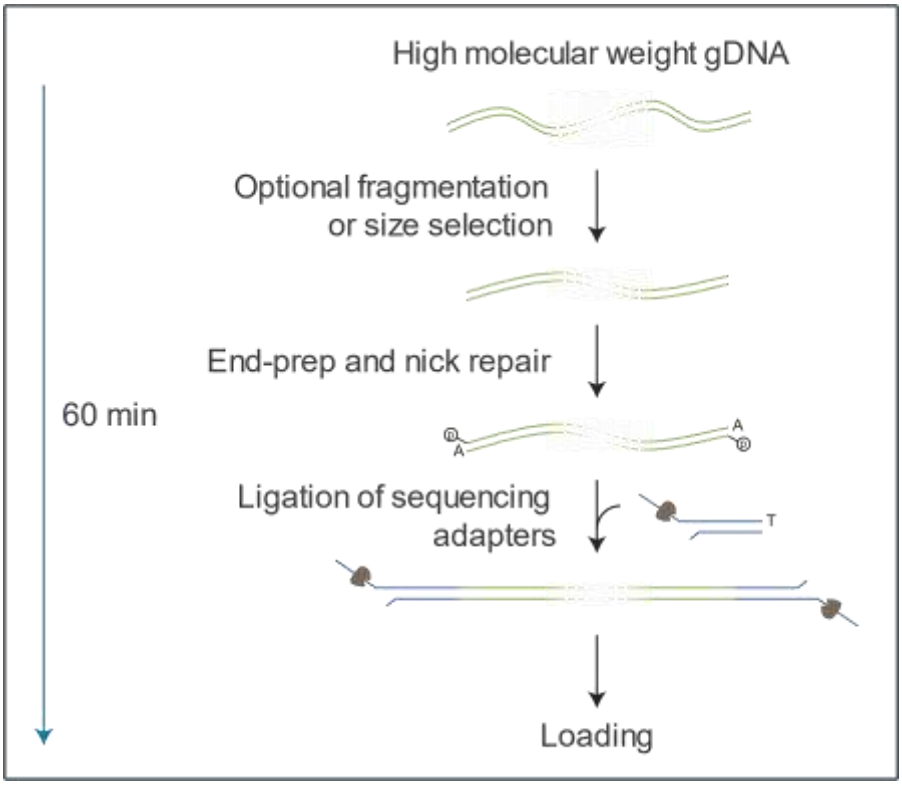


Figure 1.5: Genomic DNA is made into a library by fragmentation or size selection, nick repairing through end-preparation which involves blunting the ends which also prepares the DNA for adapter ligation. Once the adapters have been ligated to the DNA the library is ready for sequencing (51).

A flow cell is made of nanopores that are embedded in an electrically resistant polymer membrane. Each nanopore is only large enough for one strand of DNA to pass through at a time. The nanopores contain a motor protein that is responsible for nucleic acid translocation and unwinding by means of a helicase enzyme. The double stranded DNA is unwound and the negatively charged single stranded DNA is then drawn through the nanopore to the positively charged trans side of the membrane. This process is driven by voltage. The pores on the membrane are connected to a sensor chip and an application-specific integration circuit (ASIC), which is used to control electrodes responsible for monitoring the individual channels. Once a sample is loaded onto a flow cell, a constant voltage is applied across a membrane resulting in

an ionic current passing through the nanopore. The translocation of the DNA through the pore disrupts the current resulting in “squiggles” that are used for basecalling (Figure 1.6) (45).

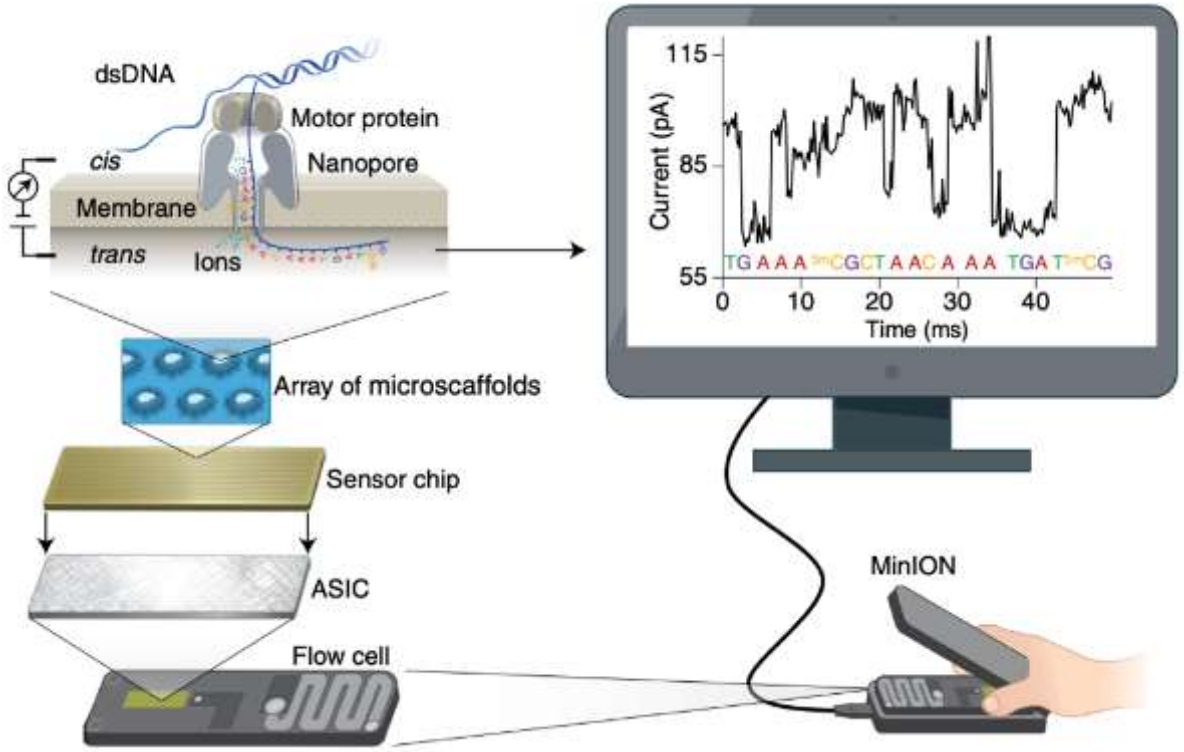


Figure 1.6: Schematic representation of Oxford Nanopore Sequencing. Genomic DNA is translocated through a membrane bound protein pore disrupting the current allowing for basecalling (45).

To read both strands of the DNA molecule, a hairpin loop is added to one end of the double stranded genomic DNA during sample preparation. Reading both strands of the DNA molecule (2D read) improves basecalling accuracy. The basecalling is done in real time (45,51,52). ONT offers different sequencing instruments and flow cells based on sequencing needs and desired data yield (Table 1.5).

Table 1.5: Comparison of ONT sequencers

Instrument	Flow cell	No. pores	Data generated
MinION	512 channels	2048	50 Gb
GridION	2560	10 240	250 Gb
PromethION	3 000	288 000/576 000	7/14 Tb
Flongle	126	504	2.5 Gb

1.7.2 Advantages and disadvantages of Nanopore sequencing

Nanopore has various advantages and disadvantages as can be seen in Table 1.6. Studies have shown that the Illumina and Ion Torrent platforms had up to 99 % accuracy whereas Nanopore sequencing has 89-90 % accuracy which is the biggest disadvantage (53–55).

Table 1.6: Advantages and disadvantages of ONT sequencing

Advantages	Disadvantages
• Low cost, rapid turnaround time, long-read sequencing generation, as well as user-friendly bioinformatics pipelines • Allows for developing bioinformatics pipelines specifically suited for data generated during sequencing • Allows for third-party analysis pipelines • The bioinformatics analysis does not require highly trained staff and/or tools to analyse the data. • The MinION, allows you to view the sequencing run in real time	• Low basecalling accuracy in comparison (52). • Installing third-party bioinformatics pipelines may be challenging and require specialized graphics processing chips

The third-party software can be used for functions such as basecalling (Albacore, Scrappie, etc), DNA modification detection (Nanopolish, Megalodon, etc), Genome alignment (Minimap2, STAR, etc) and SNP detection (Longshot, Deep Variant, etc). Because of its real-time reporting, screening and results can be viewed and reported in a shorter time.

In this research we outline a novel approach with AmpliSeq technology being implemented concurrently with ONT sequencing. We adopted a targeted approach in which we examined a panel of SNPs that are known to be associated with T1D. The combination of these technologies allows us to have a hybrid of desirable characteristics for screening of T1D susceptibility in the South African population.

1.8 Ion AmpliSeq Sequencing

Ion AmpliSeq technology is a common NGS technique that involves a targeted sequencing approach that only requires 1 ng of DNA or RNA and can be used to detect single nucleotide variants, indels and copy number variants. This technology can be classified as targeted as it has an enrichment step that uses PCR to enrich DNA regions of interest (56).

1.8.1 Mechanism of Ion AmpliSeq sequencing

Ion AmpliSeq utilises a pool of up to 24 000 oligonucleotide primer pairs in a single PCR reaction, to amplify a specific genomic region. This method can be summarised in three steps: PCR amplification of target regions using oligonucleotide primer pools, partial digestion of the amplicons, library preparation (ligation of adapters/barcodes onto the amplicons), and sequencing. (56).

1.9 Genetic risk score for prediction of T1D

GWAS have been used to identify risk loci associated with multifactorial diseases. These loci can contribute to disease development in varying degrees (24). Genetic risk scores (GRS) are predictive measures of disease susceptibility obtained from genetic data. GRS calculates the relative genetic predisposition for a trait based on the number of risk/susceptibility alleles present and different alleles will have different weighting based on the risk conferred. The individual scores are then aggregated to produce the GRS (57). These scores are generated from combining environmental, demographic as well as genetic data to predict the likelihood that an individual will develop a disease (57). A GRS gives us a probability that a specified outcome is likely to be observed, in this case disease susceptibility. Published GRS data may not be applicable as GRS are often population dependent.

The accuracy of a GRS calculator is based on the area under the receiver operating characteristic (ROC) curve (AUC). Determining the AUC measures the overall performance of the predictor, the sensitivity (true positives) as well as the specificity (false positives) (25). Very high AUC (as high as 0.99) are desirable, but all curves are expected to have an AUC > 0.75 for confident screening of individuals at risk of developing the disease (25). Factors such as whether the markers (SNPs, phenotypic or environmental traits) used can be considered strong or weak classifiers and could be the deciding factor of whether a person can be classified as susceptible or not (57,58).

A few GRS' to determine T1D susceptibly have been developed, such as that developed by Oram and colleagues (22), based on as few as 9 SNPs to predict susceptibility (AUC 0.87; $P < 0.0001$). The study was conducted on data from 3 887 individuals to distinguish between T1D and T2D. Similarly, a study conducted by Sharp and colleagues (59) aimed to determine GRS for T1D using 6 481 T1D participants, 9 247 controls and 67 SNPs. They found that through their GRS they could predict T1D susceptibility with high accuracy (AUC 0.92; $P < 0.0001$).

These studies, have however, been carried out in European populations. The development of a GRS calculator for T1D risk that is specific to the South African black population as opposed to

extrapolating upon data widely generated from European populations would allow for more accurate T1D risk determination. Individuals at increased risk can then be monitored for early identification of disease symptoms and subsequently treated early, thus delaying/preventing onset of diabetic complications, especially diabetic ketoacidosis. The incidence of T1D is increasing annually (15,16,20) and the disease places a large burden on the healthcare system (hospitalisations and medication), society (productivity) and the individual/family (20). Current diagnostic (glucose and HbA1c measurements) and treatment (insulin) methods occur too late in the disease process, thus, the development of a method that can accurately predict an individual's risk of future T1D development would be able to reduce some of these burdens.

1.10 Study aims and objectives

This study aimed to develop an accurate and cost-effective method for Nanopore sequencing that can be used to sequence target regions associated with T1D. Using the sequencing data generated, we further aimed to establish a GRS for estimation of T1D risk in the black South African population.

Study objectives:

1. To establish and validate a method for Nanopore sequencing of T1D associated HLA genes
2. To establish a method for sequencing T1D associated target regions using AmpliSeq Nanopore technology
3. To sequence target regions of T1D associated genes using AmpliSeq Nanopore sequencing in black South African participants with T1D and healthy controls.
4. To validate AmpliSeq Nanopore sequencing using Ion Torrent Ampliseq technology
5. To compare data generated by AmpliSeq Nanopore Sequencing to that obtained from PCR-RFLP.
6. To assess the utility of a Genetic Risk Scores in identifying individuals at risk of developing T1D in the South African black population.

2 Chapter 2: Materials and methods

2.1 Genomic reference DNA

DNA samples of known sequence (the Coriell trio) were selected to develop the sequencing method, this was done since one of the recommendations of the NGS validation guideline is for the utilization of reference cell lines and reference materials for evaluation of assay performance (74). The Coriell trio (NA12891, NA12892, NA12878) are standardized genomic reference samples that were developed by the Genome in a Bottle Consortium and were selected from cell lines. The DNA for these samples were extracted from the homogenized growth retrieved from the B lymphoblastoid cells from the NIGMS Human Genetic Cell Repository (https://www.coriell.org/1/NIGMS/Collections/NIST-Reference-Materials). The Trio consists of a mother (NA12892), father (NA12891) and daughter (NA12878) of white ethnicity. The trio's genomic DNA was sequenced using the Ampliseq Nanopore method developed and the resultant data compared to the known published sequences to confirm the accuracy of the method.

2.2 Study participants

Seventy-three participants were used for this study; 39 black participants with T1D were previously recruited from diabetes clinics at Charlotte Maxeke Johannesburg Academic Hospital and Steve Biko Academic Hospital and 34 black control participants recruited from staff and students at the University of the Witwatersrand. The age range of participants with T1D was 15 to 50 years and for the control participants 18 to 50 years.

Participants with clinical evidence of chronic pancreatitis, pregnancy related diabetes or type 2 diabetes were excluded from the study. In addition, control participants were excluded from the study if they had a random blood glucose concentration >11.1 mmol/L to exclude participants with undiagnosed diabetes.

All participants were required to sign a consent form (Appendix A) and complete a questionnaire (Appendix B); parents or legal guardians were asked to fill in the consent form for all participants under 18 years of age. DNA was available for all participants, Ethical clearance was obtained from the Wits Human Research Ethics Committee (HREC) (Certificate no. M200174; Appendix C).

A flow diagram that illustrates the different techniques used in the study, the number of participants used and the purpose of each technique is shown in Figure 2.1.

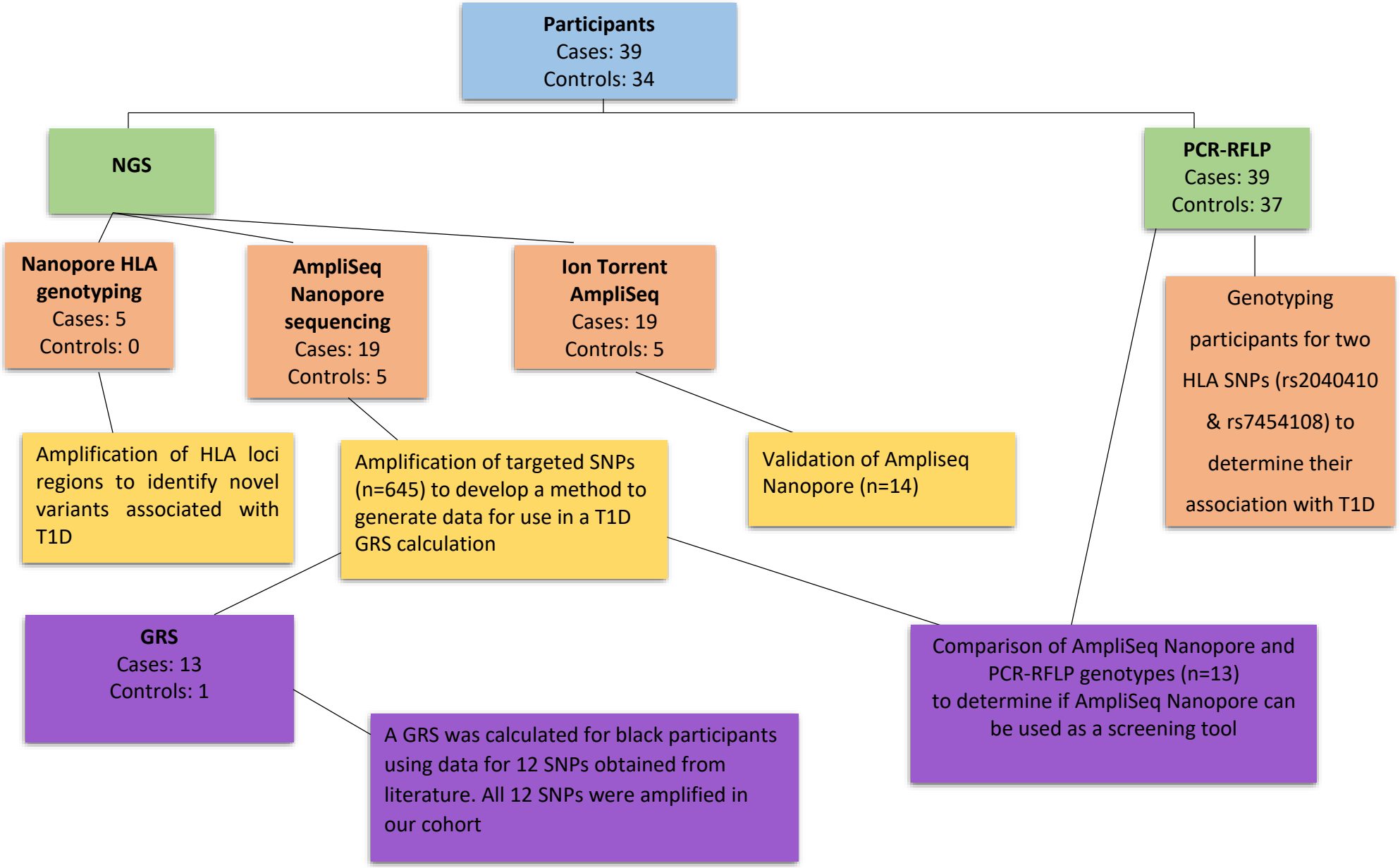


Figure 2.1: The different techniques used in the study, the number of participants used and the purpose of each technique

2.3 Nanopore sequencing of HLA regions

An ONT method that involved amplifying targeted HLA regions (HLA-A, B, C, DRB1, DQA&B and DPA&B) was investigated. Previously published primer pairs (60) (Table A.5.1; Appendix D) were used for amplification of the target HLA regions. The primers were designed to overlap to ensure that the entire region was amplified. Multiple primers were used to amplify the same HLA region to account for possible variation in the primer binding site.

2.3.1 PCR amplification

PCR reactions were set up on a MiniAmp Plus Thermo Cycler (Thermo Fisher scientific, Massachusetts, USA) using DNA from the Coriell Trio then subsequently from 5 selected T1D participants. All PCR reactions contained 10 µl 2x GoTaq (Promega, Wisconsin, USA), 50 ng DNA, PCR primers (Integrated DNA technologies, California, USA; Appendix E: Table A.5.2) and were made up to a final volume of 20 µl with PCR grade water (New England Biolabs, Massachusetts, USA). All primers had a final concentration of 10 µM.

PCR reactions were run according to the conditions outlined below:

95 ˚C for 2 min
94 ˚C for 30 secs
65 ˚C for 10 min
x30
65 ˚C for 10 min
4 ˚C ∞

2.3.2 Agarose gel electrophoresis

A 0.6 % agarose gel was prepared by adding one 0.5 g agarose tablet (Thermo Scientific, Massachusetts, USA) to 83.3 ml of 1x TBE buffer (Thermo Scientific, Massachusetts, USA) and, 5 µl of Gelgreen (Merck, New Jersey, USA). PCR product (5 µl) and 1 k ladder (4 µl; New England biolabs, Massachusetts, USA) were mixed with 5 µl of loading dye (New England biolabs, Massachusetts, USA) and then loaded onto the gel. The gel was run for 45 minutes on a gel electrophoresis machine with a built in transilluminator (bluGel™, Alabama, USA; MyGel™ mini, New Jersey, USA). Amplicons were visualised using blue LED light to ensure amplification of the correct size fragment. PCR products were quantified using the Qubit Fluorometer (Thermo Fisher scientific, Massachusetts, USA).

2.3.3 Quantification of PCR products using the Qubit fluorometer

Qubit working solution (190 µl) was added to 10 µl of each Qubit standard to make the standards for curve calibration. A volume of 199 µl of Qubit working solution was added to 1 µl of each amplified PCR product in individual tubes. The tubes were then vortexed for 5 seconds and left at room temperature for 2 minutes before being quantified. The Qubit fluorometer was set to read double stranded DNA and the prepared standards were read to generate a calibration curve. The fluorometer was then set to read 1 ng of DNA per µl and each PCR amplicon read, and the concentration recorded.

The PCR products were purified using the ProNex® DNA purification kit (Promega, Wisconsin, USA) to remove any impurities and reagents not utilised in the PCR reaction.

2.3.4 Amplicon purification using ProNex® Chemistry

The ProNex® Chemistry reagent was allowed to equilibrate to room temperature for 1 hour. The ProNex® Chemistry resin was resuspended by vigorous vortexing for 10 seconds. A 15 µl volume of ProNex® Chemistry was added into a 1.5 ml microcentrifuge tube containing 15 µl PCR product (i.e., a 1:1 v/v ratio of ProNex® Chemistry to sample) and mixed by pipetting 10 times. The sample was incubated at room temperature for 10 minutes and placed on a magnetic stand for 2

minutes; the supernatant was carefully removed and discarded. The sample was placed on the magnetic stand and 20 µl of Wash Buffer added to the sample and allowed to incubate for 30–60 seconds. The Wash Buffer was then removed and discarded. This wash step was repeated. The sample was allowed to air-dry for 5 minutes and then removed from the magnetic stand. A 15 µl volume of Elution Buffer was added to the sample and the resin resuspended by shaking on a Hula mixer™ Sample mixer (Thermo Scientific, Massachusetts, USA) for 5 minutes and the samples incubated at room temperature for 5 minutes to elute the DNA. The sample was returned to the magnetic stand for 1 minute and the eluted DNA carefully transferred to a clean tube or well. An additional Qubit reading was taken to ensure that there was sufficient DNA to proceed with the Nanopore library preparation.

2.3.5 Library preparation

A library preparation step that involved DNA end-prep, adapter ligation, and DNA purification and clean-up was performed using the Oxford Nanopore sequencing ligation kit according to the manufacturer's instructions (Oxford Nanopore Technologies, Oxford, England).

2.3.5.1 DNA repair and end-prep

DNA CS (DCS) was thawed at room temperature, spun down, mixed by pipetting and placed on ice. The DNA was prepared in nuclease-free water by transferring 1 µg of genomic DNA into a 1.5 ml microcentrifuge tube. The volume was adjusted to 47 µl with nuclease-free water and the contents mixed thoroughly by flicking the tube and spinning down briefly in a microfuge (Benchmark Scientific, New Jersey, USA). The end-prep volumes and reagents were prepared in a 0.2 ml thin-walled PCR tube as per Table 2.1.

Table 2.1: Reagent volumes used for PCR amplified DNA repair and end-prep

Reagent	Volume (µl)
DNA CS	1
Prepared DNA	47
NEBNext FFPE DNA Repair Buffer	3.5
NEBNext FFPE DNA Repair Mix	2
Ultra II End-prep reaction buffer	3.5
Ultra II End-prep enzyme mix	3
Total	**60**

The components were thoroughly mixed by flicking the tube and spinning down in a microfuge. The reactions were incubated in a thermocycler at 20 °C for 5 minutes followed by 65 °C for 5 minutes.

2.3.5.2 AMPure XP bead clean-up

The AMPure XP beads (Beckman Coulter, California, USA) were resuspended by vortexing. The end-prepped DNA was transferred to a clean 1.5 ml microcentrifuge tube. A volume of 60 µl of resuspended AMPure XP beads was added to the end-prep reaction and the contents mixed by flicking the tube. The tube was incubated on a Hula mixer™ Sample mixer for 5 minutes at room temperature. The sample was spun for 5 minutes at 6 000 rpm and pelleted on a magnet until the eluate was clear and colourless (1-2 minutes). The tube was kept on the magnet, and the supernatant removed. The beads were washed with 200 µl of freshly prepared 70 % ethanol without disturbing the pellet. The ethanol was removed using a pipette and discarded. The wash step was repeated. The tube was centrifuged and placed back on the magnet. The residual ethanol was pipetted off and the sample dried for approximately 30 seconds (not allowing the pellet to dry to the point of cracking). The tube was removed from the magnetic rack and the pellet resuspended in 61 µl nuclease-free water and incubated for 2 minutes at room temperature. The beads were pelleted on a magnet until the eluate was clear and colourless (1-

2 minutes). The eluate was removed and pipetted into a clean 1.5 ml microcentrifuge tube. The DNA was quantified using a Qubit fluorimeter (section 2.3.3)

2.3.5.3 Adapter ligation and clean-up

Adapter Mix (AMX) and NEBNext Quick T4 DNA Ligase were spun down and placed on ice. Ligation Buffer (LNB), Elution Buffer (EB), and one tube of Long Fragment Buffer (LFB) were thawed at room temperature, spun down, mixed by vortexing and immediately placed on ice. LFB was chosen to enrich for DNA fragments of 3 k or longer. The adapter ligation reaction was set up in a final volume of 100 µl using the reagents listed in Table 2.2. The ligation reaction was thoroughly mixed by flicking the tube, spun down and incubated at room temperature for 10 minutes.

Table 2.2: Reagents used for the binding of sequencing adapters to repair and end-prepped DNA

Reagent	Volume (µl)
End-prepped DNA	60
Ligation Buffer (LNB)	25
NEBNext Quick T4 DNA Ligase	10
Adapter Mix (AMX)	5
Total	**100**

2.3.5.4 AMPure XP bead clean-up

The AMPure XP beads were mixed by vortexing. A volume of 40 µl of resuspended AMPure XP beads was added to the ligation reaction and mixed by flicking the tube. The tube was incubated on a Hula mixer™ Sample mixer for 5 minutes at room temperature. The sample was spun down and pelleted on a magnet, the tube kept on the magnet, and the supernatant pipetted off. The beads were washed by adding 250 µl Long Fragment Buffer (LFB). The beads were flicked to resuspend, spun down, then the tube returned to the magnetic rack and the beads allowed to pellet. The supernatant was removed using a pipette and discarded. The wash step was repeated. The tube was spun down and placed back on the magnet. Any residual supernatant was pipetted

off and the pellet was allowed to dry for approximately 30 seconds (pellet was not allowed to dry to the point of cracking). The tube was removed from the magnetic rack and the pellet resuspended in 15 µl of Elution Buffer (EB). The tube was then spun down and incubated for 10 minutes at room temperature. The beads were pelleted on a magnet until the eluate was clear and colourless (5 minutes). Fifteen microliters of eluate containing the DNA library was removed and transferred into a clean 1.5 ml microcentrifuge tube. One microliter of the DNA was quantified using a Qubit fluorometer (section 2.3.3).

2.3.6 Priming and loading of the flow cell

A Flongle flow cell (Oxford Nanopore Technologies, Oxford, England) was primed and loaded as outlined below.

The Sequencing Buffer (SQB), Loading Beads (LB), Flush Tether (FLT) and one tube of Flush Buffer (FB) were thawed at room temperature, mixed by vortexing and spun down. The MinION Mk1C (Oxford Nanopore Technologies, Oxford, England) lid was opened and the flow cell slid under the clip. In a fresh 1.5 ml microcentrifuge tube, 117 µl of FB was mixed by pipetting with 3 µl of FLT. The seal tab was peeled back from the Flongle flow cell, up to the point where the sample port was exposed. The flow cell was primed with the mix of FB and FLT, ensuring that there was no air gap in the sample port or the pipette tip. A P200 pipette tip was placed inside the sample port and the priming fluid slowly dispensed into the Flongle flow cell by turning the pipette dial to zero to avoid flushing the flow cell too vigorously. The vial of LB was vortexed. The Sequencing Mix (30 µl) was prepared in a fresh 1.5 ml microcentrifuge tube by mixing 15 µl SQB with 5 µl DNA library and 10 µl LB (mixed immediately before use).

A P100 tip was placed inside the sample port and the Sequencing Mix slowly dispensed into the flow cell by twisting the pipette dial to zero ensuring no air gaps. The Flongle flow cell was sealed using the adhesive on the seal tab and the sequencing machine lid closed. The sequencing run was set up on the MinION Mk1C instrument and the sequencing run carried out overnight. Sequencing data was then stored on the interface for further analysis.

The samples were run individually with sequencing parameters set as follows: Basecalling enabled, run length set for 24 hours, mux scan set for 30 minutes, Basecall model set to High accuracy basecalling and read filtering set to a minimum Phred-scaled quality score (Qscore) of 9.

Not enough sequencing data was generated for HLA typing via sequencing thus analysis was not carried out, data coverage of 100x is desirable to ensure that the results obtained are significant. None of the sequenced samples had a coverage of 100x. Therefore, it was decided to combine the preparation steps of Ion Torrent AmpliSeq with the portable Nanopore sequencing platform to target multiple T1D associated SNPs.

2.4 AmpliSeq Nanopore Sequencing for SNP genotyping

AmpliSeq Nanopore sequencing (the name given to the hybrid Ion Torrent AmpliSeq and Nanopore sequencing technique) combines PCR amplification using oligonucleotide primer pairs designed from Ion AmpliSeq for targeted genomic sequencing with the library preparation and sequencing technique provided by Oxford Nanopore Technologies. The modified method combines Ion AmpliSeq targeted screening through PCR enrichment with the ONTs user friendly sequencing technology. The method allows for genotyping participants for high risk T1D polymorphism. The method is cheap, and generates shorter reads with shorter runtimes, therefore it requires minimal computational resources and bioinformatics expertise.

2.4.1 SNP selection and primer design

The Ion AmpliSeq Primer design tool (56) was used to generate a pool of 611 primers that covered 645 T1D associated SNPs. The SNPs were selected by Evans Mathebula after reviewing the literature. However, for the current study, twelve of the 645 SNPs (Table 2.3) were selected for further analysis because of their reported high GRS from GWAS studies (22,61–63).

Table 2.3: The SNP number, chromosome, gene, nucleotide change and the type of each selected SNP

Chromosome	Gene	rs number	Nucleotide change	Type of SNP
6	LOC124901301	rs2040410	G>A, T	500 bp downstream variant,
6	-	rs7454108	T>C	-
1	PTPN22	rs2476601	A>G, T	Missense variant
6	NCR3	rs2857595	G>A, C, T	-
6	TSBP1-AS1	rs1980493	T>C	Intron variant, genetic downstream transcript variant
6	HLA-DQA1	rs9272346	G>A, T	Non-coding transcript variant, 2 K upstream variant
6	HLA-DQB1	rs2647044	G>A	-
10	IL2RA	rs12722495	T>C	Intron variant
10	LINC00993	rs7100025	G>A	-
11	Ins-IGF2	rs689	A>G, T	Intron variant, 5’ UTR variant
15	SMAD3	rs72743477	A>G, T	Intron variant
16	CLEC16A	rs12708716	A>G	Intron variant, genic downstream transcript variant

2.4.2 PCR amplification

PCR reactions were set up using the reagents listed in Table 2.4. This was a multiplex reaction where all SNPs were amplified at the same time.

Table 2.4: AmpliSeq Nanopore reagents for PCR amplification of T1D susceptibility genes

Reagent	Manufacturer	Volume (µl)	Final concentration
Ion AmpliSeq Primer pool mix	Thermo Fisher scientific, Massachusetts, USA	25	20 µM
LongAmp® Taq	New England Biolabs, Massachusetts, USA	25	2x
DNA template	-	2	50 ng
Final volume		**52**	

The PCR reactions were run in a MiniAmp Plus Thermo Cycler according to the PCR conditions shown below:

95 ˚C for 2 min

98 ˚C for 13 secs
60 ˚C for 8 mins } x30

10 ˚C ∞

2.4.3 Agarose gel electrophoresis

PCR products were loaded into a precast 4 % E-Gel EX agarose gel (Thermo Fisher, Massachusetts, USA) pre-stained with SYBR Gold II DNA gel stain and run in the Invitrogen E-gel™ Power Snap system (with built in blue-light trans-illuminator; Thermo Fisher scientific, Massachusetts, USA) for 20 mins at 48 V to confirm that the correctly sized amplicons were generated. Thereafter, the

concentration of the PCR products was determined using a Qubit fluorometer (section 2.3.3) to check there was enough DNA for purification and library preparation.

2.4.4 Purification of PCR products

The PCR products were purified using the Zymogen DNA Clean and Concentrator 5 purification kit (Zymo Research, California, USA). Amplified DNA (45 μl) was added to 225 μl of Binding buffer in a 1.5 ml microcentrifuge tube. The mixture was then transferred to a Zymo-Spin™ column in a collection tube and centrifuged at 12 000 rpm for 30 seconds. The flow-through was discarded and 200 μl of Wash buffer added to the column. The column was then centrifuged for 30 seconds. This wash step was repeated. A volume of 15 μl of water was added directly to the column matrix and incubated at room temperature for one minute. The column was then transferred to a labelled 1.5 ml microcentrifuge tube and centrifuged for 30 seconds to elute the DNA. The concentration of DNA was determined using a Qubit fluorometer (section 2.3.3).

2.4.5 Chemical modification of AmpliSeq PCR fragments

Prior to the library preparation an additional chemical "modification" step was added to the protocol to allow binding of the Nanopore sequencing adapters. The reagents used in the chemical modification step contain a cocktail of enzymes including an endonuclease responsible for creating an abasic site on the DNA strand by cleaving an uracil on the DNA sequence while leaving the backbone intact; the reagents also include enzymes responsible for repairing and preparing DNA for PCR. This is carried out by repairing nicks in the DNA as well as ensuring that PCR inhibitors (apurinic/apyrimidinic sites, thymine dimers, gaps) are removed. The chemical modification ensured that an active hydroxyl group is available on the 3' end of the DNA to ensure binding of the Nanopore adapter. The reaction was set up in a final volume of 50 μl as outlined in Table 2.5. Please note that a description of the reaction buffer and cocktail of reagents 1-4 can be obtained from Dr Ali. The details are not provided here as they are the subject of a patent filing.

Table 2.5: AmpliSeq Nanopore reagents used for chemical modification of PCR amplicons

Reagent	1x reaction volume (µl)	Final concentration
PCR Product	30	-
Reaction buffer	5	10x
Reagent 1	0.5	-
Reagent 2	0.5	-
Reagent 3	1	-
Reagent 4	1	-
Water	12	-
Total volume	**50**	

These reagents were incubated in a MiniPCR® machine (MiniPCR® bio, Massachusetts, USA) at 37 °C for 15 minutes followed by 65 °C for 20 minutes. The DNA was purified using the Zymogen DNA Purification Protocol as outlined in section 2.4.4. Recovery of DNA was quantified using a Qubit fluorometer (section 2.3.3).

2.4.6 Library preparation

2.4.6.1 End prep

The library preparation step was then initiated using a modified Oxford Nanopore Sequencing by Ligation Protocol. All reagents were obtained from New England Biolabs (Massachusetts, USA). The first step involved an end-prep step as outlined in Table 2.6. This step repairs nicks on DNA and blunts the end for adapter binding.

Table 2.6: AmpliSeq Nanopore reagents for End prep

Reagent	1 x reaction (µl)
Purified DNA	10
NEBNext Ultra II End Prep Reaction Buffer	1.75
NEBNext Ultra II End Prep Enzyme Mix	0.75
Water	2.5
Total volume	**15**

The DNA was then incubated on a MiniAmp Plus Thermo Cycler for 15 minutes at 25 °C and 65°C for 15 minutes. Water (35 µl) was added to the DNA which was then purified using the AMPure bead purification protocol (section 2.3.5.2).

2.4.6.2 Native barcoding

We wanted to compare samples sequenced individually and in a multiplex sequencing run for data yield, sequencing time, sequencing quality and an overall comparison of the two sequencing methods. Thus, some samples had barcodes added to them to allow for multiplex sequencing. A barcode was added to PCR amplicons to identify individual samples. Native barcode (2.5 µl) and 25 µl 2x Blunt/TA Ligase Master Mix (Thermo Fisher, Massachusetts, USA) were sequentially added to the end prepped DNA (22.5 µl). The reaction was incubated at room temperature for 10 minutes. The reaction then underwent an AMPure bead purification as outlined in section 2.3.5.2. This barcoding step was only performed when sequencing multiple samples at a time. If a single sample was sequenced, this step was omitted.

2.4.6.3 Adapter Binding

Reagents for the Adapter ligation step (Table 2.7) were then mixed with 15 µl end prepped/barcoded DNA and the reaction incubated at room temperature for 20 minutes.

Table 2.7: AmpliSeq Nanopore reagents for Adapter binding

Reagent	1x reaction volume (µl)
LNB Ligation buffer	12.5
Quick T4 DNA Ligase	5
AMX Adapter mix	5
Water	12.5
Final volume	**50**

The adapter bound amplicons were purified using AMPure XP beads (section 2.3.5.4). However, Short Fragment Buffer (SFB) was used instead of LFB. A final concentration of 15-20 Femto moles (Fmol) was required to continue with the sequencing reaction.

2.4.7 Flow cell priming and Sequencing

2.4.7.1 Loading of a Flongle flow cell for single sample sequencing

Flow cells were primed and loaded according to section 2.3.6 above for non-multiplexed samples.

2.4.7.2 MinIon flow cell priming and loading for sequencing multiple barcoded samples

A MinION flow cell (Oxford Nanopore Technologies, Oxford, England) was primed and loaded as per the protocol described below. All buffers were thawed and mixed before use.

The priming port cover was slid clockwise to open the priming port and the flow cell checked for a small air bubble under the cover. A small volume (20 µl) of flow cell sequencing storage buffer was drawn back to remove any bubbles. To prepare the flow cell priming mix, 30 µl of FLT was added directly to the tube of FB and mixed by vortexing. Priming mix (800 µl) was added into the flow cell via the priming port, avoiding the introduction of air bubbles and left to stand for 5 minutes. During this time, the library was prepared for loading by combining 37.5 µl SBQ with 12 µl DNA library and 25.5 µl LB (thoroughly mixed by pipetting immediately before use). An additional 200 µl Priming mix was loaded into the flow cell via the priming port, avoiding the introduction of air bubbles. The prepared library was mixed by gently pipetting up and down and

75 µl added immediately to the flowcell via the SpotON sample port in a drop wise fashion (ensuring each drop flowed into the port before adding the next). The SpotON sample port cover was gently replaced, making sure the bung enters the SpotON port and sequencing initialised.

The sequencing run was set up on the MinION Mk1C platform for the sequencing run to be carried out overnight. The samples were run in multiplex with sequencing parameters set as follows: Basecalling enabled, run length set for 72 hours, mux scan set for 30 minutes, Basecall model set to High accuracy basecalling and read filtering set to a minimum Qscore of 9; Qscores of greater 20 are desirable but for nanopore data Qscores of greater than 9 are desirable due to the reported 89-90 % accuracy for sequencing method. Sequencing data was then stored on the interface for further analysis.

2.5 Bioinformatics analysis

A bioinformatics pipeline that aligned sequence data to a reference sequence, to allow for variant calling, was established. The overall aim of the pipeline was to determine whether an individual is at risk for developing T1D based on the overall GRS derived from the SNP sequence data. Standard bioinformatics pipelines were applied wherever possible such that analysis could be performed on any Unix platform without the need for complex dependencies or hardware. In the pipeline, the reference file used for alignment is generated using an in-house script written using Python. The Fastq sequencing files are aligned to the reference file using Samtools generating a .bam file. The alignments are indexed, sorted and converted into a .sam file. The aligned file then undergoes variant calling and filtering using Varfilter and bcftools generating a .bcf file which is then converted to a readable .vcf file. The .vcf files are analysed manually and variants identified for each sample.

2.5.1 Reference sequence design

Initially the Coriell trio (NA12891, NA12892, NA12878) were used to develop the bioinformatics pipeline as standards where NA12878 was used as a reference sequence for alignment. Standards were used as they have known published sequences for comparison. Initially each chromosome for NA12878 on NCBI (https://www.ncbi.nlm.nih.gov) was downloaded to produce a test reference sequence. This reference sequence was generated to reduce computation time during analysis as it is significantly shorter when compared to the chromosome sequence. The reference sequence produced was uploaded on Galaxy (https://usegalaxy.org) and extractseq with coordinates in the bed file used to obtain the sequence of the SNP regions of interest. These sequences were checked using SnapGene (https://www.snapgene.com) to see if the sequences matched the reference sequence (NA12878/GRCh38). The University of California Santa Cruz (UCSC) genome browser (https://genome.ucsc.edu) and National Centre for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov) were used for risk allele identification.

This method required the use of several different programs and websites for data analysis. Therefore, a method that bypassed the steps outlined above and allowed for the retrieval of desired sequences and reference file generation in one step was developed. This method was specific to the 12 SNPs and is described below.

2.5.1.1 In-house reference script

A reference file was created from the Human genome (hg38) chromosome sequences retrieved from NCBI (https://www.ncbi.nlm.nih.gov/genome/?term=human+genome), hg38 was used in this case because the Trio only served to validate the bioinformatics pipeline. The sequences flanking the 12 SNPs were retrieved from the human genome chromosome sequences using an in-house script written on Python (Figure 2.2). This in-house script is designed to retrieve regions of interest from hg38. The downloaded file is saved as a variable first which will then be used for sequence retrieval. The desired coordinates of the region of interest are then plugged into the code and retrieved from the created variable using a *For* and *Print* function. The sequences of

the 12 DNA regions enriched using AmpliSeq, were then concatenated to derive a "virtual" reference sequence that was 11.6 k using the *cat* function.

```
import sys                                                          #Saves
output in the desired file

stdoutOrigin = sys.stdout
sys.stdout        =        open("Save        file        name",        "w")
#Input file name that you want it saved as

with open('file retrieving from', 'rt') as myfile:                          #
Open desired file for reading (you want to retried from)
    for myletter in myfile:                                                 #
For each letter, read to a string,
        print(myletter["start  coordinate  +  4":"end  coordinate  +  4"])
# and print the string or the range.
```

Figure 2.2: In-house python script for the retrieval of target region sequences to construct a reference genome.

2.5.2 Sequence alignment

The pipeline uses Minimap2 (https://github.com/lh3/minimap2) for the indexing and alignment of the fastq sequencing files to the generated fasta reference file.

2.5.3 Alignment file sorting and conversion

Samtools (https://www.htslib.org/download/) was used for alignment file manipulation and sorting. Samtools was used to convert the generated binary .bam file from Minimap2 to an indexed and sorted alignment file saved as a readable .sam file.

2.5.4 Variant calling and filtering

The pipeline used bcftools (https://www.htslib.org/download/) for variant calling and filtering. Variant calling parameters were set at -mv -Ob -o. These parameters specify the input file inserted and output file that needs to be generated and how the filtering should be carried out. In this case the input file is a .bam file and the output file is a .bcf file and default filtering parameters are applied. vcfutils.pil and VarFilter were used for filtering. Bcftools was also used to convert the generated file after variant calling from a .bcf file to a readable .vcf file. The scripts used for 2.5.2-2.5.4 are summarised in Figure 2.3.

Index reference file

```
[ ] minimap2 -x map-ont -d ref.mmi ref.fa
```

Align Nanopore data

```
[ ] minimap2 -ax map-ont ref.mmi sample.fq.gz > sample.sam
```

convert from sam to bam

```
[ ] samtools view -S -b sample.sam > sample.bam
```

Sorting bam file

```
[ ] Samtools sort sample.bam > sample_sorted.bam
```

Calling variants

```
[ ] bcftools mpileup -f ref.fa sample_sorted.bam | bcftools call -mv -Ob -o sample.bcf
```

```
[ ] samtools mpileup -f ref.fa my-sorted-
    1.bam my-sorted-2.bam my-sorted-n.bam > my- raw.bcf
```

Converting bcf to vcf and filtering

```
[ ] bcftools view sample.bcf | vcfutils.pl varFilter - > sample_filtered.vcf
```

Figure 2.3: Script for the analysis of sequencing data generated using minimap2, Samtools and bcftools on Google colab for reference file and sequencing file alignment, sorting indexing and variant calling.

2.5.5 Variant analysis identification

The variant files (.vcf) were then opened in Microsoft Excel and the wild-type alleles that were called identified and recorded if present for each sample.

2.5.6 Validation of Ampliseq Nanopore sequencing by ion torrent sequencing

Samples (n=20) were sequenced using the newly developed Ampliseq Nanopore technique and on the NGS gold standard Ion Torrent sequencing machine to validate the accuracy of the sequencing results obtained from AmpliSeq Nanopore sequencing. The samples for Ion Torrent were prepared and sequenced according to manufacturer's instructions (61). The PCR amplification and amplicon purification steps were carried out by myself and the sequencing by employees at Akili Labs.

2.6 Genetic risk score determination for this population

For this research project we used odds ratios that were obtained from the largest meta-analysis study (https://bio.tools/t?topicID=%22topic_3517%22) and weightings developed by Oram and team (22) to generate a GRS for the population. They used 30 SNPs (Table 2.8) known to have a high association with T1D obtained from GWAS studies. GWAS are used to determine odds ratios as they investigate large populations thus ensuring that results obtained are significant. GWAS studies develop a beta coefficient, which is an expected result due to an investigated variable. This beta coefficient is used to calculate odds ratios to determine GRS in GWAS studies. For this study a beta coefficient of 0.4287 obtained from the study by Oram and colleagues was used (22).

Table 2.8: SNPs used in the GRS calculator, the genes they are found in, their odds ratios and their weighting in the calculator (22)

SNP	Gene	Odds ratio	Weight	Ancestral allele
	DR3/DR4-DQ8	48.18	3.87	
	DR3/DR3	21.12	3.05	
Rs2187668, rs7454108	DR4-DQ8/DR4-DQ8	21.98	3.09	
	DR4-DQ8/X	7.03	1.95	
	DR3/X	4.53	1.51	
RS1264813	HLA-A	1.54	0.43	T
RS2395029	HLA-B	2.5	0.92	T
RS3129889	HLA-DRB1	14.88	2.70	A
RS2476601	PTPN22	1.96	0.67	A
RS689	INS	1.75	0.56	T
RS12722495	IL2RA	1.58	0.46	T
RS2292239	ERBB3	1.35	0.30	T
RS10509540	C10orf59	1.33	0.29	T
RS4948088	COBL	1.3	0.26	C
RS7202877	-	1.28	0.25	G
RS12708716	CLEC16A	1.23	0.21	A
RS3087243	CTLA4	1.22	0.22	G
RS1893217	PTPN2	1.2	0.18	G
RS11594656	IL2RA	1.19	0.17	T
RS3024505	IL10	1.19	0.17	G
RS9388489	C6orf173	1.17	0.16	G
RS1465788	-	1.16	0.15	C
RS1990760	IFIH1	1.16	0.15	T
RS3825932	CTSH	1.16	0.15	C

RS425105	-	1.16	0.15	T
RS763361	CD226	1.16	0.15	T
RS4788084	IL27	1.16	0.15	C
RS17574546	-	1.14	0.13	C
RS11755527	BACH2	1.13	0.12	G
RS3788013	UBASH3A	1.13	0.12	A
RS2069762	IL2	1.12	0.11	A
RS2281808	-	1.11	0.10	C
RS5753037	-	1.1	0.10	T

The standard GRS calculation equation was employed (Equation 1) to calculate the GRS for this population. In the study by Oram et. al. a GRS >0.280*30 (number of SNPs used in the GRS calculator) = 8.40 was indicative of T1D, with 95 % specificity and 50 % sensitivity, thus the same criteria were applied to this population.

Equation 1: GRS calculation equation; where N is the number of SNPs in the score, βi is the effect size (or beta) of variant I and $dosage_{ij}$ is the number of copies of the allele in the genotype of individual j.

$$PRS_j = \sum^{N}{}_i \beta_i * dosage_{ij}$$

2.7 Genotyping participants for two HLA gene SNPs using PCR-RFLP

All participants (n=73) were genotyped for two *HLA* gene polymorphisms (rs2040410 and rs7545108) by PCR based RFLP

2.7.1 PCR amplification of regions flanking rs2040410 and rs7454108

Primers (Appendix F: Table A.5.3) were designed using Primer3 software (64) to amplify a 484 bp and 326 bp fragment containing the rs2040410 and rs7454108 polymorphism, respectively.

A mastermix (Table 2.9) was prepared in a 1.5 ml microcentrifuge tube to amplify the regions of interest. Mastermix (21 µl) was pipetted into each of the labelled PCR tubes. DNA (4 µl) was added to each PCR tube and 4 µl of water was added to the negative control (blank). The PCR tubes were centrifuged to bring all the reagents to the bottom of the tube and placed in the T100^{TM} Thermal Cycler (Bio Rad, California, USA).

Table 2.9: PCR reagents for the amplification of the two HLA polymorphisms

Reagent	1x Reaction volume (µl)	Final concentration
Kapa Taq buffer A[1] (10x)	2	1x
dNTPs[2] (10 mM)	0.5	0.4 mM
Forward primer[3] (10 pmol/µl)	1	0.35 pmol/µl
Reverse primer[3] (10 pmol/µl)	1	0.35 pmol/µl
Kapa Taq[1] (5U/µl)	0.1	1U/µl
Water	16.4	-
DNA	4	±36 ng/µl
Final volume	**25**	

[1] Kapa Biosystems, Massachusetts, USA, [2] Bioline, Massachusetts, USA, [3] Inqaba biotech, Pretoria, SA

The PCR reactions for both SNPs were run using the PCR conditions shown below:

95 ˚C for 3 min
95 ˚C for 30 secs
58 ˚C for 30 secs
72 ˚C for 30 secs
72 ˚C for 1 min
4 ˚C

x30

∞

2.7.2 Agarose gel electrophoresis

A 2 µl aliquot of the PCR product mixed with 2 µl Ficol loading dye was run on a 1 % agarose for 30 minutes at 100 V and viewed on a ChemiDoc™ MP Imaging System (Bio Rad Laboratories, (PTY) Ltd, Hercules, California, USA) to confirm amplification of the correct size fragments.

2.7.3 Restriction digestion of the PCR products

To genotype participants for the rs2040410 and rs7454108 *HLA* polymorphisms, the PCR products were digested with BsrGI (New England Biolabs, Massachusetts, USA) and SspI (New England Biolabs, Massachusetts, USA), respectively. PCR products (10 µl) were digested using BsrGI/SspI (0.5 µl) and 10 x restriction enzyme buffer (2 µl) in a final volume of 20 µl (Table 2.10).

Table 2.10: Reagents required for the digestion of the PCR products

Reagents	rs2040410	rs7454108
Restriction enzyme	BsrGI 20 000 U/ml	SspI 10 000 U/ml
10x Buffer	CutSmart buffer	NEB buffer SspI

Samples for the rs2040410 polymorphism were digested in a T100™ Thermal Cycler for two hours at 37 °C and 20 minutes at 80 °C. Samples for the rs7454108 polymorphism were incubated in a T100™ Thermal Cycler for one hour at 37°C followed by 30 minutes at 65°C to deactivate the enzyme.

2.7.4 Gel electrophoresis

Digested samples were loaded onto a 2 % agarose gel and run for 30 minutes at 100 V and visualised. Participants were scored for the presence or absence of the BsrGI and SspI restriction sites based on the fragment sizes present. The three possible genotypes for each polymorphism and their corresponding fragment sizes are listed in Table 2.11.

Table 2.11: The three possible genotypes for each polymorphism and their corresponding fragment sizes

HLA polymorphism	Genotype	Fragment size (bp)
rs2040410	GG	299 and 185
	GA	185, 299 and 484
	AA	484
rs7454108	TT	129, 197
	TC	129, 197 and 326
	CC	326

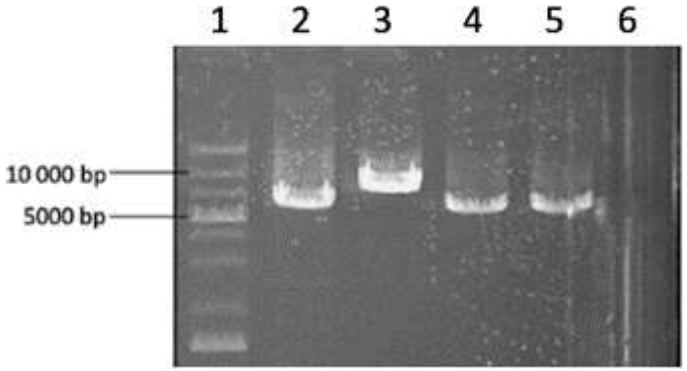
2.8 Statistical analysis

Genotypic frequencies and allelic frequencies between patients and controls were compared using the Chi-squared (χ^2) test (https://www.socscistatistics.com/tests/chisquare2/default2.aspx) to determine statistical significance and values presented as frequency . Hardy-Weinberg equilibrium was investigated and the probability of the observed results being due to chance and to observe how well the data obtained fits with the expected data for genotypic data. For all analyses, $p < 0.05$ was considered statistically significant, and would mean the allele observed is not due to chance but association with T1D susceptibility.

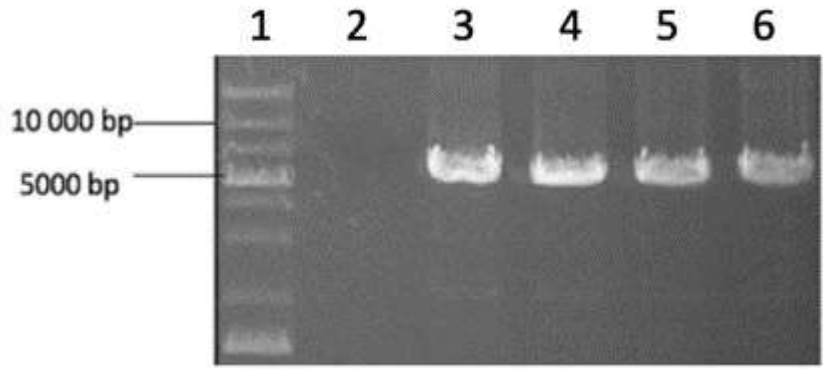
3 Chapter 3: Results

3.1 HLA typing by Nanopore sequencing

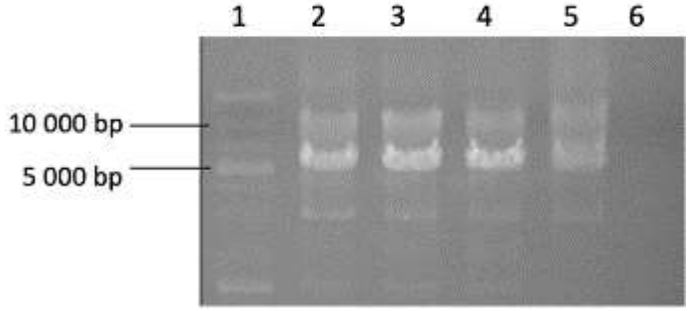
The method was developed using the Coriell Trio (NA12878, NA12891 and NA12892) and Promega human DNA (https://worldwide.promega.com/products/biochemicals-and-labware/nucleic-acids/human-genomic-dna/?catNum=G1521) to optimize PCR amplification. HLA genes (HLA-A, -B, and -C, DQA1 &2, DPA1, DPB1, DRB1) were individually amplified using PCR and some of the resultant amplicons are illustrated in Figure 3.1.


1 2 3 4 5 6
10 000 bp
5000 bp

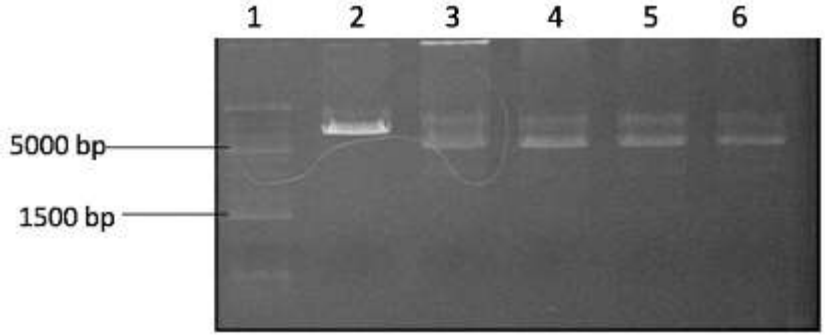
A


1 2 3 4 5 6
10 000 bp
5000 bp

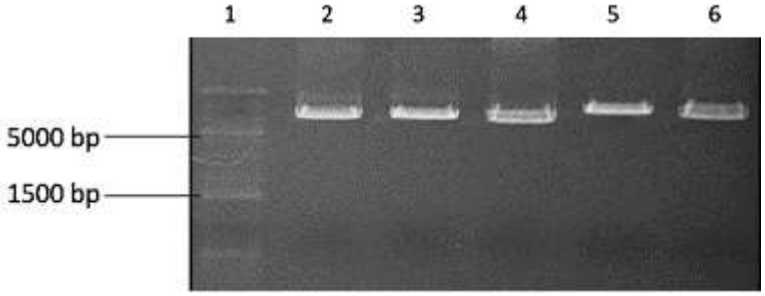
B


1 2 3 4 5 6
10 000 bp
5 000 bp

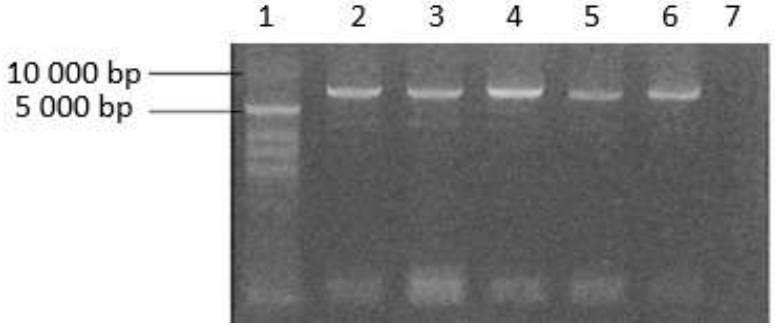
C


1 2 3 4 5 6
5000 bp
1500 bp

D


1 2 3 4 5 6
5000 bp
1500 bp

E


1 2 3 4 5 6 7
10 000 bp
5 000 bp

F

Figure 3.1: A. PCR products (5000 and 9 000 bp) from the amplification of the HLA-A region run on a 0.6 % agarose gel.

Lane 1: 1 kb Molecular ladder, Lane 2: NA12891, Lane 3: NA12892, Lane 4: NA12878, Lane 5: Promega DNA, Lane 6: Negative control.

B. PCR products (5 000 bp) from amplification of HLA-B run on a 0.6 % gel.

Lane 1: Molecular ladder, Lane 2: Negative control, Lane 3: NA12891, Lane 4: NA12892, Lane 5: NA12878, Lane 6: Promega DNA

C. PCR products (3 000 - 11 000 bp) from the amplification of the HLA-C region run on a 0.6 % agarose gel.

Lane 1: 1 kb Molecular ladder, Lane 2: NA12891, Lane 3: NA12892, Lane 4: NA12878, Lane 5: Promega DNA, Lane 6: Negative control.

D. PCR products (6 000 - 9 000 bp) from the amplification of the DQA&B1 region run on a 0.6 % agarose gel.

Lane 1: 1 kb Molecular ladder, Lane 2: Promega DNA DQA1, Lane 3: DQB1 NA128921, Lane 4: DQB1 NA12872, Lane 5: DQB1 NA12878, Lane 6: DQB1 Promega DNA.

E. PCR products (6 000 - 9 000 bp) from the amplification of the DPA1&B1 region run on a 0.6 % agarose gel.

Lane 1: 1 kb Molecular ladder, Lane 2: DPA1 NA12891, Lane 3: DPA1 NA12892, Lane 4: DPA1 NA12878, Lane 5: DPB1 NA12891, Lane 6: DPB1 NA12898.

F. PCR products (9 000 bp) from the amplification of the DRB1 region run on a 0.6 % agarose gel.

Lane 1: 1 kb Molecular ladder, Lane 2: PR39, Lane 3: PR48, Lane 4: PR53, Lane 5: PR69, Lane 6: PR73, Lane 7: Negative control,

3.2 AmpliSeq Nanopore Sequencing

3.2.1 PCR amplification of T1D associated genes

DNA from participants with T1D and control participants were amplified using the Ion AmpliSeq primer pool. Samples (n=24; 19 T1D participants and five controls) were randomly selected for NGS sequencing for method development and validation. Samples that amplified successfully (n=14) during the multiplex PCR and had sequencing data of ± 120 000 reads were analysed further. A representative gel of the resultant PCR product is shown in Figure 3.2.

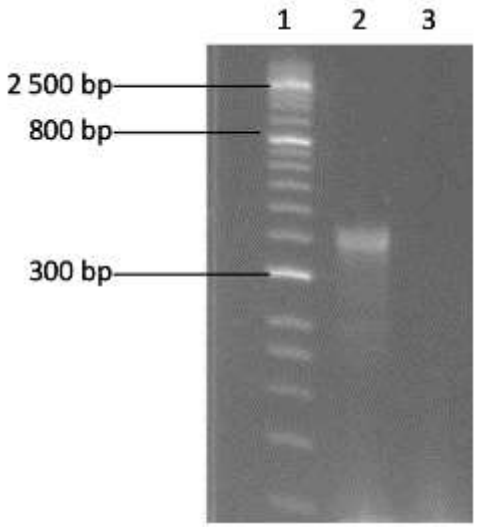


Figure 3.2: A 4 % agarose gel of amplicons after PCR amplification of the regions flanking the 645 SNPs in a T1D patient (PR30). Lane 1: 50 bp Molecular ladder Lane 2: Amplicons (Sizes 100-350). Lane 3: Negative control.

3.2.2 Quality control prior to sequencing

Samples were purified after every step prior to sequencing and DNA concentration measured using a Qubit Fluorimeter to ensure that each sample had sufficient DNA to continue with preparation for sequencing. The readings taken for the samples are listed in Table 3.1. Not all samples passed the quality control (n=4) step thus were not included in the analysis step.

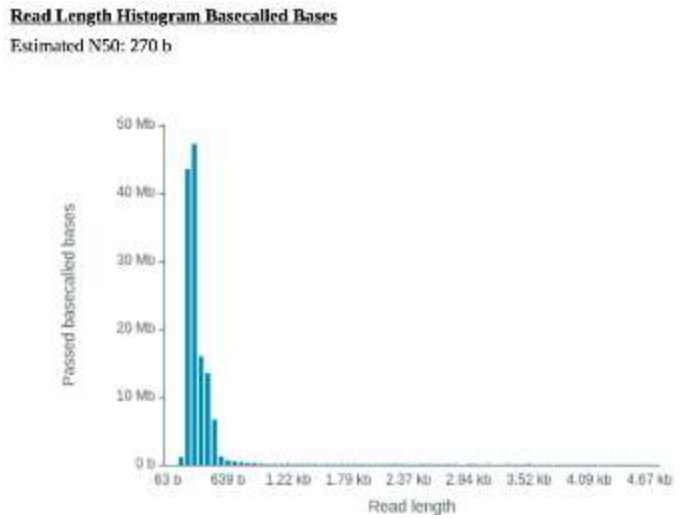
Table 3.1: DNA concentrations for after purification as well as the volume used for sequencing

Sample	DNA concentration				Volume for sequencing (µl)
	After chemical modification (ng/µl)	After End prep (ng/µl)	After adapter binding (ng/µl)	After adapter binding (Fmol)	
PR2(I)	2.46	0.878	0.792	5.772	3
PR7	5.52	0.258	1.43	12.74	1.5
PR10	0.356	0.160	0.202	4.172	4
PR11	2.98	1.71	1.62	11.81	1.5
PR15	1.03	0.492	0.748	5.451	3
PR30	0.364	0.340	0.340	5.55	3
PR31	0.782	0.434	0.474	3.454	5
PR35	3.62	1.57	0.554	4.037	4
PR36	10.30	4.38	2.08	16.20	1
PR37	17.6	8.42	5.50	40.08	0.4
PR38	3.22	1.87	2.10	15.30	1
PR39	3.10	1.17	1.63	11.88	1.5
PR40	4.14	4.44	1.36	10.42	1.5
PR48	14.30	4.64	2.24	15.05	1
PR53	2.38	4.44	1.74	10.73	1.5
PR54	0.394	0.667	0.398	5.55	3
PR69	1.35	2.50	0.656	7.23	2
PR73	2.72	1.96	0.798	8.47	2
VDR160	0.438	0.174	0.340	3.478	5
VDR247	0.302	0.136	0.104	0.758	7
VDR258	<0.005	<0.005	0.110	0.802l	7
VDR262	<0.005	<0.005	<0.005	<0.05	7
VDR266	0.588	0.198	0.402	2.930	3

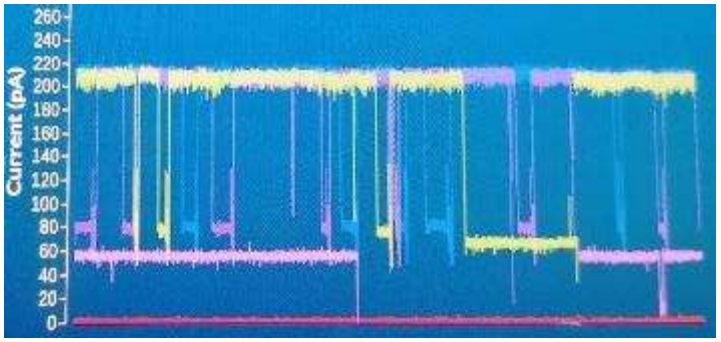
3.2.3 Nanopore sequencing

A Nanopore sequencing run for T1D patient sample PR48 is illustrated in Figure 3.3 as an indication of the typical sequencing run. This figure shows the read length histogram of the fragments that were being sequenced, the current that was produced for each nanopore, the voltage that was passed across the flow cell at any given time and the state of each nanopore (i.e., which nanopores were being used for sequencing (green), those which were inactive (light blue/white) and those that were recovering (blue)). If an air bubble is introduced during library loading the pores would appear as white “dead” pores. The software would show if a sequencing run was in progress, which sequencer the sample(s) is being sequenced on, the flow cell ID, the capacity of the flow cell as well as which process (Mux scanning, sequencing, and flow cell check) during sequencing is underway.

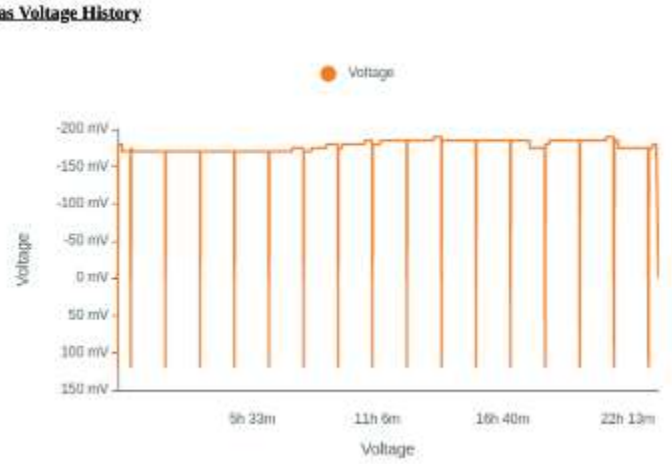
A.


Read Length Histogram Basecalled Bases
Estimated N50: 270 b
50 Mb
40 Mb
30 Mb
20 Mb
10 Mb
0 b
Passed basecalled bases
63 b
639 b
1.22 kb
1.79 kb
2.37 kb
2.94 kb
3.52 kb
4.09 kb
4.67 kb
Read length

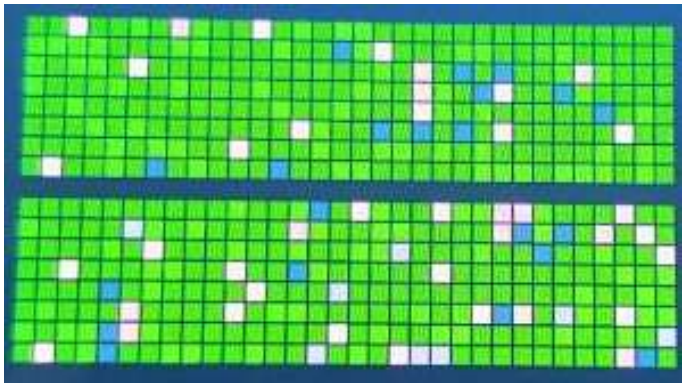
B.


260
240
220
200
180
160
140
120
100
80
60
40
20
0
Current (pA)

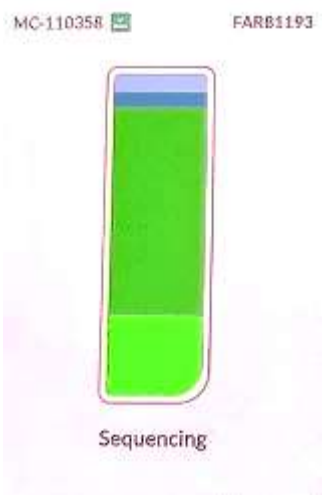
C.


Bias Voltage History
Voltage
-200 mV
-150 mV
-100 mV
-50 mV
0 mV
50 mV
100 mV
150 mV
Voltage
5h 33m
11h 6m
16h 40m
22h 13m
Voltage

D.



E.



F.

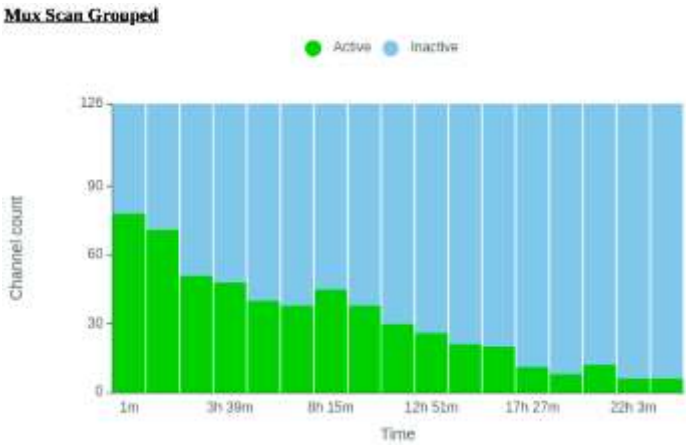
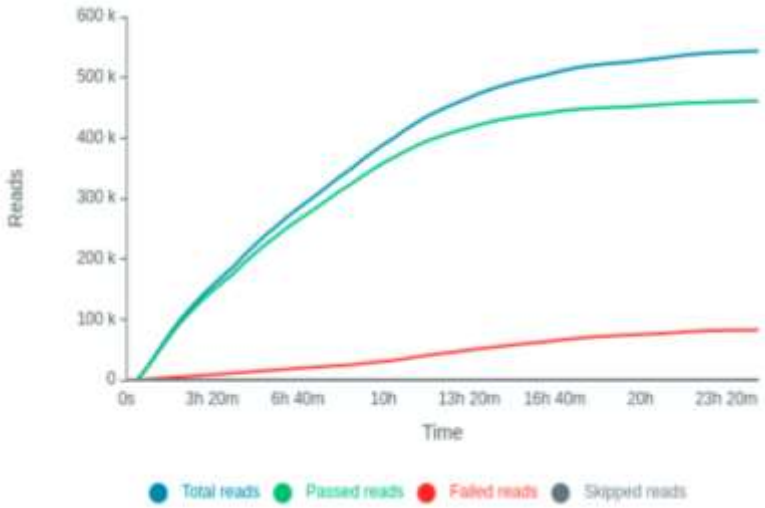


Figure 3.3: Nanopore sequencing run for T1D patient PR48. A. Average read length identified during the sequencing run. B. Current produced in each nanopore during the sequencing run. C. Mux scan and voltage history during sequencing run. D. Real-time active, recovering, and inactive pores. E. Progress of sequencing run, which sequencing machine and flow cell is running and capacity of flow cell. F. Summarised sequencing times and pore count over sequencing duration.

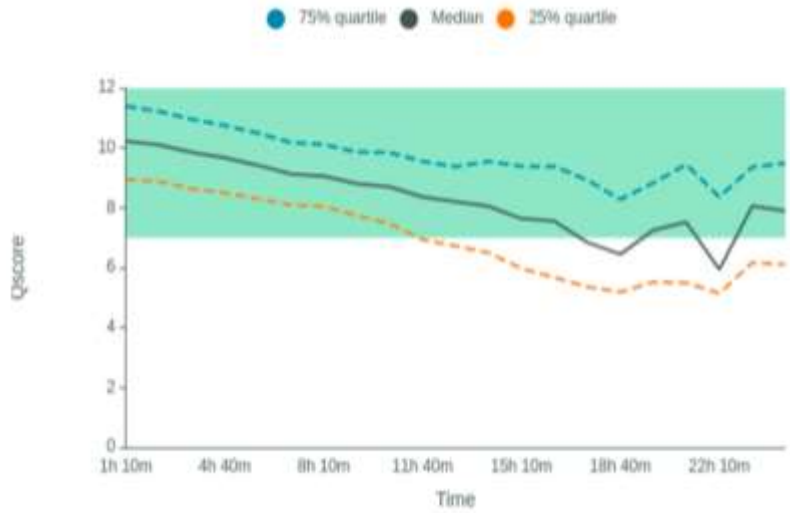
3.2.4 Sequencing summary information

Figure 3.4 shows a data file generated after a sequencing run for PR48. Graph A shows the reads that were obtained from the sequencing run, we obtained close to 500 kilobases (kb) reads in a 24-hour sequencing run on a Flongle flow cell. Graph B indicates the average quality score obtained for this sequencing run, the average quality score exceeded 9. A Qscore reflects the probability that the sequence generated is correct. The higher the Qscore, the higher the probability that the sequence is correct. A Qscore of 20 represents an error rate of 1 in a 100 bases sequenced and a 99 % call accuracy (40,41). Due to the low accuracy of the Nanopore flow cell Qscores of greater than 9 were expected but those >20 would still be ideal. Lastly, Graph C shows the estimated sequence yield in relation to the estimated read length which helped to verify that amplicons of the anticipated sizes (± 270 bp) were being sequenced.

A


600 k
500 k
400 k
300 k
200 k
100 k
0
Reads
0s
3h 20m
6h 40m
10h
13h 20m
16h 40m
20h
23h 20m
Time
Total reads
Passed reads
Failed reads
Skipped reads

B


75% quartile
Median
25% quartile
12
10
8
6
4
2
0
Qscore
1h 10m
4h 40m
8h 10m
11h 40m
15h 10m
18h 40m
22h 10m
Time

C

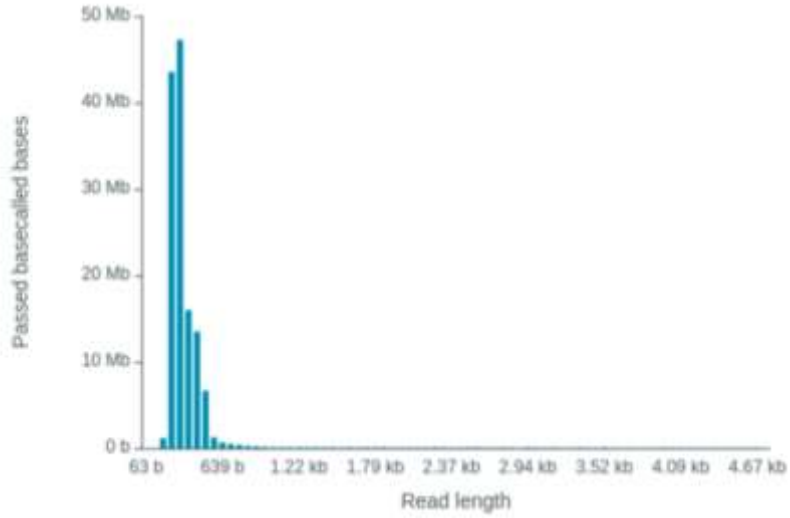


Figure 3.4: A. The total reads generated for sample PR48. B. The average Qscore generated. C. The average read length and the passed basecalled bases generated during this sequencing run.

Table 3.2 outlines the average Qscore, reads generated, passed bases, failed bases, estimated bases and run time for the sequencing run for individually sequenced samples. Each sample was run until the end of the programmed sequencing time (24 hours) or until the library was depleted (<24 hours). This table includes participants who obtained adequate data for analysis (± 120 000 reads) and those that had inadequate data. Participants that are not included in the table had failed sequencing runs. The average Qscores for these runs was 9.73.

Table 3.2: Average Qscore, reads generated, passed bases, failed bases, estimated bases and run time for T1D participants and controls that were sequenced individually

	Sample	Average Qscore	Reads generated (kb)	Passed bases (Mb)	Failed bases (Mb)	Estimated bases (Mb)	Run time
Standards	NA12878	10	199.18	12.23	47.16	56.56	21h27m
	NA12891	9.5	72.22	15.11	10.33	22.09	18h7m
	NA12892	11	35.24	7.91	2.34	9.8	51m
T1D	PR10	10	15.33	3.78	2.02	5.88	21h46m
participants	PR11	10	73.64	17.04	5.2	23.31	23h36m
	PR15	8.5	171.8	38.5	13.97	55.33	24h
	PR27	9	314.75	79.21	10.86	91.43	20h26m
	PR30	9.5	300.52	82.26	5.74	85.38	24h1m
	PR37	9	261.3	45.85	24.78	75.04	24h
	PR48	9	543.61	136.34	22.94	153.38	24h1m
Controls	VDR160	10	75.97	17.92	4.03	23.97	24h
	VDR258	12	0.0039	0.00749	0.0073	0.01435	4h10m
	VDR266	9	0.0322	0.00754	0.0039	0.01197	24h1m

*kb-kilobases, Mb-Megabase

We wanted to compare data yield, read depth, read quality as well as run time with samples sequenced individually vs samples run with barcodes in a “multiplex” sequencing run. The same samples were run the first time, then upon library depletion the sequencing run was stopped, the flow cell was flushed and reloaded with the same library and data from both runs recorded. The data is summarised in Table 3.3. The average Qscores generated during these runs was 9.5.

Table 3.3: Average Qscore, reads generated, passed bases, failed bases, estimated bases and run time for T1D participants and controls that were in a multiplex sequencing run using barcoding

	Sample	Average Qscore	Reads generated (kb)	Passed bases (Mb)	Failed bases (Mb)	Estimated bases (Mb)	Run time
First run	**T1D participants** PR7, PR36, PR40, PR53, PR54, PR69, PR73, PR83 **Controls** VDR159, VDR160, VDR207, PR247	8	286.65	47.93	42.35	94.04	24h21m
Second run	**T1D participants** PR7, PR36, PR40, PR53, PR54, PR69, PR73, PR83 **Controls** VDR159, VDR160, VDR207, PR247	11	15.63	4.05	908.8	5.09	2h15m

Data generated from the sequencing runs with samples run individually as well as data generated with sequencing runs run in multiplex were compared as outline in Table 3.4. Average Qscores, reads generated, passed bases, failed bases, estimated bases and run time for T1D participants and controls were compared.

Table 3.4: Comparison of data generated for individual samples vs multiplex sequencing run

Run type	Average Qscore	Average reads generated (kb)	Average passed bases (Mb)	Average failed bases (Mb)	Average estimated bases (Mb)	Average run time
Samples run individually	9.73	187.60	41.47	13.58	54.72	19.46
Samples run in multiplex	9.5	151.14	25.99	475.58	49.57	13.5

3.3 Bioinformatics analysis

3.3.1 Pipeline validation using standards

Variant calling of standard samples (The Coriell Trio) as determined by the bioinformatics pipeline that was developed is outlined in Table 3.5. The table summarises the ancestral allele for each SNP as well as the allele observed after Ion AmpliSeq Nanopore sequencing. If the participant had an allele that is different from the ancestral allele, the allele would be identified by the variant caller and the Qscore for how much the variant called can be trusted produced. A high Qscore (≥ 20) is desired as it indicates the confidence in the called variant. Qscores were only generated if the wild-type allele was present i.e. If the participant had an alternate allele and not the ancestral allele. The average Qscore for called variants was 73.89.

Table 3.5: Called alleles for each standard (NA12878, NA12891 and NA12892) and their Qscores.

SNP	Ancestral	NA12878 (Qscore)	NA12891 (Qscore)	NA12892 (Qscore)
rs2040410	C	CC (-)	CC (-)	CC (-)
rs7454108	T	TT (-)	TT (-)	TT (-)
rs9272346	G	GG (-)	GG (-)	GG (-)
rs2647044	G	GG (-)	GG (-)	GG (-)
rs7100025	G	AG (116.16)	GG (-)	AG (26.42)
rs12708716	A	AA (-)	AA (-)	AA (-)
rs72743477	A	GA (187.38)	GG (68.37)	GA (28.42)
rs689	A	AA (-)	TA (80.41)	AA (-)
rs2857595	G	GG (-)	AG (79.36)	AG (19.13)
rs2476601	A	GA (84.33)	GG (79.36)	GA (69.02)
rs12722495	T	TT (-)	TT (-)	TT (-)
rs1980493	T	TT (-)	CT (48.41)	TT (-)

3.3.2 Genotyping participant samples using Ion AmpliSeq analysis

The pipeline was then tested on sample data. All of the 12 chosen SNPs were mapped successfully to the reference genome. Four out of five of the control samples produced less data than what was required (≥ 120 000 sequencing reads) due to poor DNA quality and concentration, therefore only one control sample (VDR160) was genotyped. The data collected after alignment, filtering and variant calling is outlined in Table 3.6. The table outlines the ancestral allele for each SNP as well as the allele observed after Nanopore sequencing. The average quality score for called variant was 89.77. The table includes samples that generated enough data for analysis (13), samples that did not have enough data were omitted (11).

Table 3.6: Called alleles for T1D participants and a control

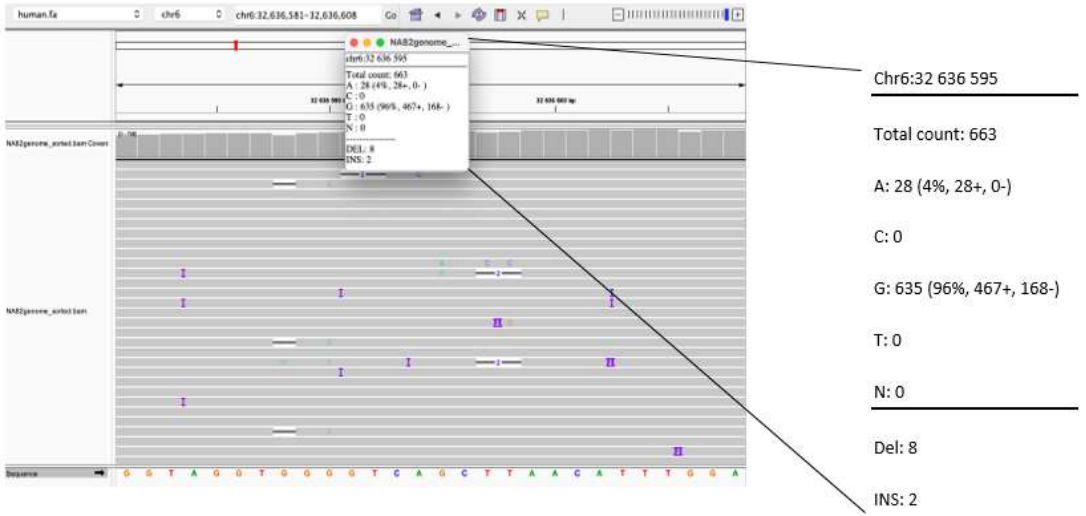
		rs2040410	rs7454108	rs9272346	rs2647044	rs7100025	rs12708716
	Ancestral	**G**	**T**	**G**	**G**	**G**	**A**
T1D Participant	PR11	GG	TT	GG	GG	AG	AA
	PR27	GG	TT	AG	AG	GG	AA
	PR30	GG	TT	GG	GG	GG	AA
	PR36	GG	TT	GG	GG	GG	GA
	PR37	GG	TT	AG	AA	AG	GA
	PR40	GG	TT	GG	GG	GG	GA
	PR48	GG	TT	GG	AG	AG	GA
	PR53	GG	TC	GG	GG	AG	GG
	PR54	GG	TT	GG	GG	AG	AA
	PR69	GA	TT	AG	AG	AG	GG
	PR73	GG	TT	AG	AG	GG	GA
	PR83	GG	TT	GG	GG	AA	GA
Control	VDR160	GG	TT	GG	GG	GG	AA

		rs72743477	rs689	rs2857595	rs2476601	rs12722495	rs1980493
	Ancestral	**A**	**A**	**G**	**A**	**T**	**T**
	PR11	AA	AA	AG	GA	TT	CT
	PR27	AA	AA	AG	GA	TT	TT
	PR30	AA	AA	AG	GG	TT	TT
	PR36	AA	AA	AA	GA	TT	TT
	PR37	AA	AA	AG	GA	TT	TT
	PR40	AA	AA	GG	GG	TT	TT
	PR48	AA	AA	GG	GG	CT	TT
	PR53	AA	AA	AA	GG	TT	TT
	PR54	AA	AA	AA	GG	TT	CT
	PR69	AA	AA	AA	GA	TT	TT
	PR73	GA	AA	GG	GA	TT	CT
	PR83	AA	AA	AG	GG	TT	TT
Control	VDR160	AA	AA	AG	GG	TT	TT

3.4 Validation of Ampliseq Nanopore Sequencing by Ion Torrent Sequencing

Samples used for the research were sequenced using the Ion GeneStudio S5 system. Initially, the sequencing run data analysis was carried out on the Coriell Trio to confirm accurate basecalling as these sequences for these standards is known. Real Genomic Analysis tools (RTG-tools) were used for comparison between data generated during Nanopore and Ion Torrent sequencing. These RTG-tools help accurately analyse VCF files generated after bioinformatics analysis. The RTG tools identified accuracy and precision rates of 0.388 and 0.433 respective (precision rates as close to 1 as possible are desirable). Figure 3.5 shows data obtained from variant calling using these RTG-tools using integrated genomic viewer (IGV), Figure 3.5A shows a variant called for SNP rs9272346, 4 % of the bases were the disease allele and 96 % of the base were the wild-type allele. Figure 3.5B shows a SNP that was called correctly to be heterozygous. Based on the data obtained from Ion Torrent sequencing and the use of RTG-tools the sequencing data obtained from Nanopore sequencing was found to be comparable to Ion Torrent data.

A



B

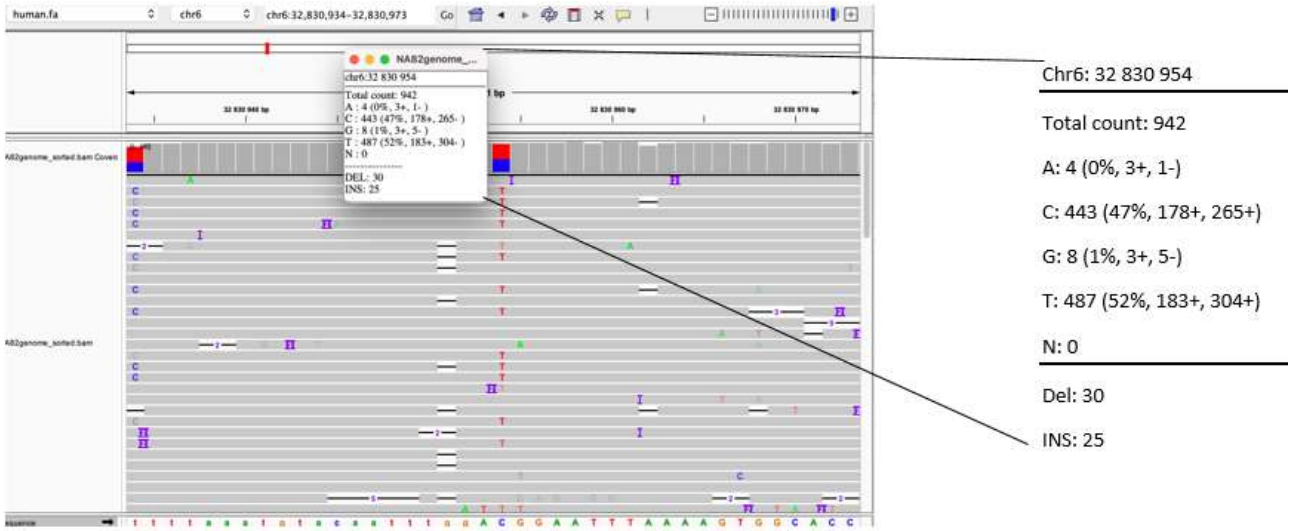


Figure 3.5: A. shows a variant called for SNP rs9272346, 4 % of the bases were the disease allele (A) and 96 % (G) of the base were the wild-type allele. B. Shows a SNP that was called correctly to be heterozygous CT.

The comparison of T1D participants and controls was then investigated. Due to the poor DNA yield obtained for the samples only a few samples could be sequenced and compared successfully using Ion Torrent sequencing. Based on the data generated, Ion Torrent data was more accurate but ONT data was found to be similar to that generated using Ion Torrent. For the comparable samples the average percentage difference for sample that could be compared was 5.67 %. More high quality samples are needed to validate these results.

3.5 Genetic risk scores calculated for the South African black population

The GRS' were calculated for each participant in this population as outlined in Table 3.7 with the help of Dr Jonathan Featherston. A GRS of >0.280*30 (number of SNPs used in the GRS calculator) = 8.40 was indicative of T1D, with 95 % specificity and 50 % sensitivity. The GRS was also trained by Dr Featherston on data obtained from the Teddy cohort for validation of the GRS. An AUC graph was generated and can be found in Appendix I.

Table 3.7: GRS' obtained for this participants using odds ratios obtained from GWAS from literature

	Participant	GRS
T1D Participant	PR11	7.25
	PR15	7.65
	PR27	6.99
	PR30	6.83
	PR36	6.62
	PR37	7.02
	PR40	7.14
	PR48	7.15
	PR53	7.64
	PR54	6.64
	PR69	7.98
	PR73	7.31
	PR83	7.23
Control	VDR160	7.51

3.6 Participants were genotyped for the two HLA gene polymorphisms (rs2040410 and rs7454108) using PCR-RFLP.

Figure 3.6 illustrates PCR products obtained following amplification of DNA with primers flanking the rs2040410 and rs7454108 HLA polymorphisms. Figure 3.7 illustrates results obtained following restriction digest of the PCR products.

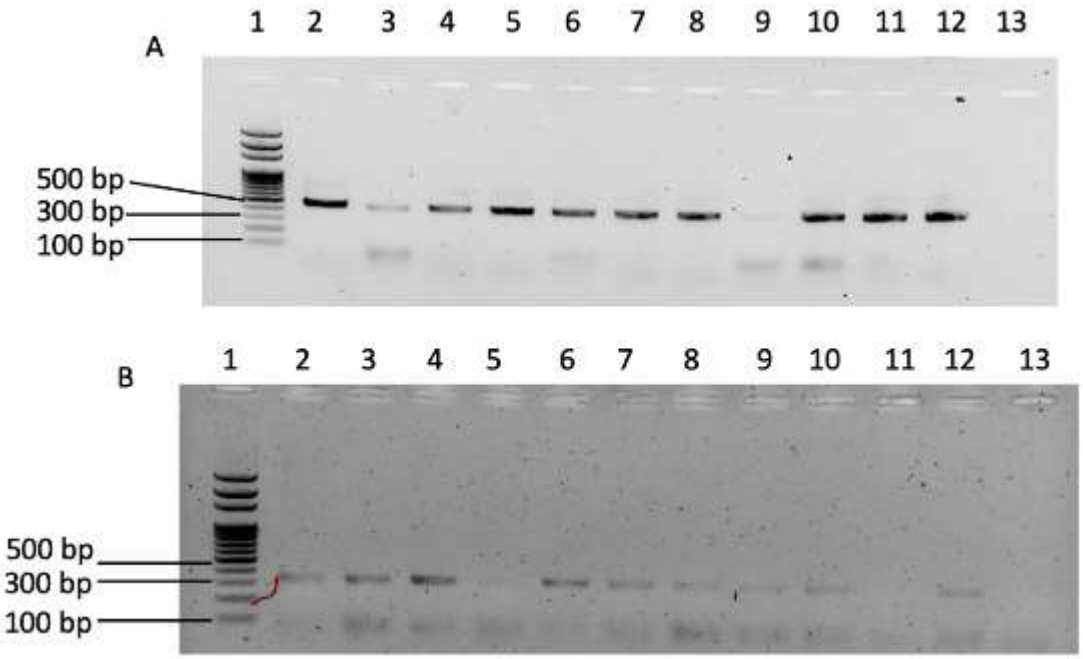


Figure 3.6: A. PCR products (484) from the amplification of the rs2040410 SNP region run on a 1 % agarose gel.

Lane 1: 100 bp Molecular ladder, Lane 2-12: samples patient samples, Lane 13: Negative control.

B. PCR products (326) from the amplification of the rs2040410 SNP region run on a 1 % agarose gel.

Lane 1: 100 bp Molecular ladder, Lane 2-12: Control samples, Lane 13: Negative control.

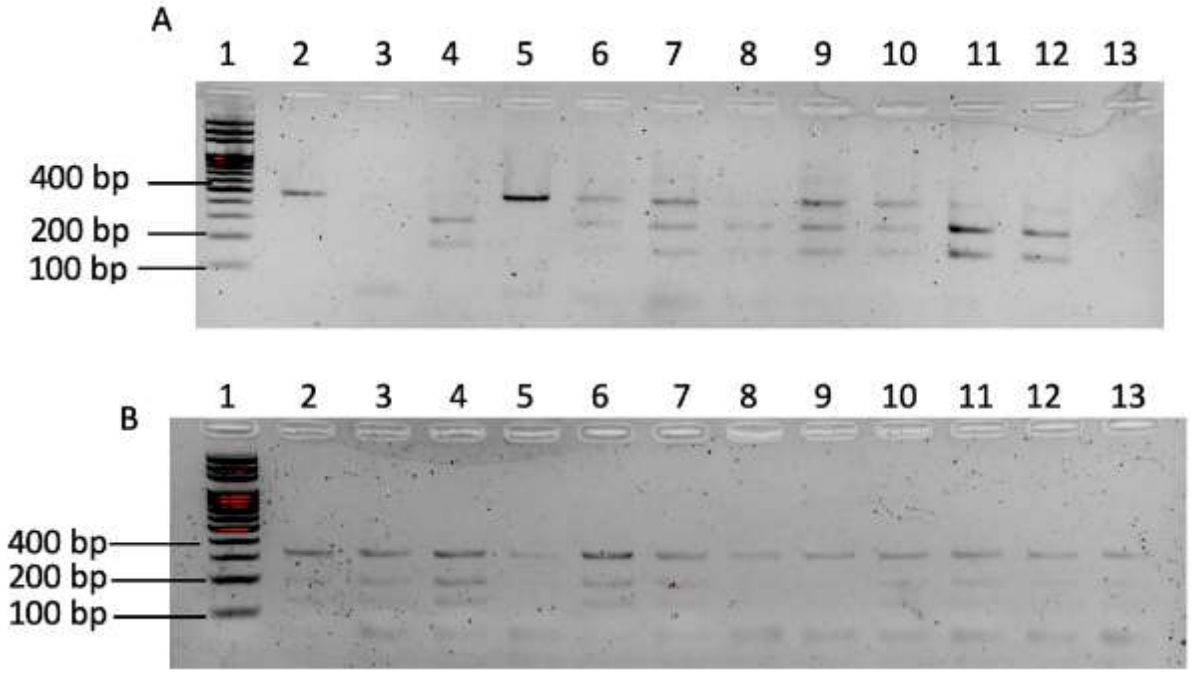


Figure 3.7: A. PCR-RFLP digest for the rs2040410 SNP on a 2 % agarose gel.

Lane 1: 100 bp Molecular ladder, Lane 1 and 5: Genotype AA (484 bp) Lane 4,8, 11, and 12: Genotype GG (299 bp and 185 bp), Lane 6, 7, 9, and 10: Genotype GA (299 bp, 185 bp, and 484 bp).

B. PCR-RFLP digest for the rs7454108 SNP on a 2 % agarose gel.

Lane 1: 100 bp Molecular ladder, Lane 5, and 9: Genotype CC (326 bp) Lane 2-4, 6-8, and 10-13: Genotype TC (299 bp, 185 bp, and 326 bp).

3.6.1 Allelic and genotypic frequencies for rs2040410 and rs74542108

The genotypes obtained for each T1D participant and control for rs2040410 and rs7454108 are outlined in Table 3.8. Table 3.9 illustrates a summary of the allelic and genotypic data obtained for the T1D and control participants. Only samples that had genotypes for both SNPs (61 samples) were included in the statistical analysis.

Table 3.8: Genotypes obtained from PCR-RFLP of black T1D participants and black controls for rs2040410 and rs7454108

Sample type	Sample ID	Rs2040410 genotype	Rs7454108 genotype
Patients	PR2	AA	TT
	PR 7	AA	TT
	PR 10	GG	TT
	PR 11	GG	TT
	PR 15	GA	TT
	PR27	AA	TT
	PR30	AA	TT
	PR 35	GG	TT
	PR36	AA	TT
	PR 37	GA	TT
	PR 38	GG	CC
	PR 39	GA	TT
	PR48	AA	TT
	PR 40	GG	TT
	PR 53	GG	TC
	PR54	AA	TT
	PR 69	AA	TT
	PR 83	AA	TT
	PR 97	GG	TC
	PR 102	AA	TC
	PR 103	GA	TC
	PR 104	GA	TC
	PR 107	GG	TC
	DBN 15	GA	TC
	DBN 33	GA	TC
	DBN 37	GG	TC
	DBN 41	GG	TC
	DBN 57	-	TC
	DBN 58	AA	TC
	T1D 15	GG	TC
	T1D 50	GG	TC
	T1D 85	GA	TC

	T1D 119	GG	TC
	VDR 150	GA	TC
	VDR 199	GG	TC
	VDR 200	GG	TC
	VDR 230	GA	TC
	VDR 281	GG	TC
	VDR 300	-	TC
Controls	VDR 29	-	TC
	VDR 34	GG	TC
	VDR 70	GA	TC
	VDR 95	GG	CC
	VDR 123	GG	TC
	VDR 126	GA	TC
	VDR159	AA	CC
	VDR 160	GA	CC
	VDR207	AA	CC
	VDR 241	-	CC
	VDR247	AA	CC
	VDR 258	-	TC
	VDR 263	AA	TC
	VDR 304	GG	TC
	VDR 324	GA	-
	JWM 1	GG	TC
	JWM 2	-	TC
	JWM 3	-	TC
	JWM 4	-	TC
	JWM 5	-	TC
	JWM 6	GG	TC
	JWM 7	GG	TC
	JWM 15	GG	TC
	JWM 17	GG	TC
	JWM 25	GA	TC
	T1D 300	GG	TC
	T1D 303	GG	TC
	T1D 304	GG	TC

	T1D 305	GG	TC
	T1D 306	GG	TC
	T1D 311	GA	TC
	T1D 368	GA	TC
	T1D 384	-	TC

Table 3.9: Genotypic and allelic frequencies for the rs2040410 and rs7454108 polymorphisms in the black south African population

SNP	Frequency (n)		p-value
	T1D participants (n=38)	Control participants (n=23)	
rs2040104			
Genotype model			
GG	0.42 (16)	0.61 (14)	
GA	0.26 (10)	0.26 (6)	0.220
AA	0.32 (12)	0.13 (3)	
Allele model			
G	0.68 (42)	0.85 (34)	0.039*
A	0.45 (34)	0.26 (12)	
rs7454108			
Genotype model			
TT	0.42 (16)	0.0 (0)	
TC	0.55 (21)	0.78 (18)	0.0012*
CC	0.03 (1)	0.22 (5)	
Allele model			
T	0.66 (53)	0.45 (18)	0.000895*
C	0.37 (23)	0.55 (28)	

*Statistically significant

3.6.2 Comparison of HLA PCR-RFLP genotyping with Ion AmpliSeq Nanopore sequencing

Table 3.10 illustrates the genotypes observed for all participants for rs2040410 and rs7454108 obtained from PCR-RFLP and Nanopore sequencing. Only participants that had sequencing data as well as RFLP data were compared

Table 3.10: Comparison of SNP genotypes for rs2040410 and rs7454108 for T1D/control participants with PCR-RFLP and Nanopore sequencing

Sample type	Sample ID	rs2040410		rs7454108	
		PCR-RFLP	ONT	PCR-RFLP	ONT
T1D participants	PR11	GG	GG	TT	TT
	PR27	AA	GG	TT	TT
	PR30	AA	GG	TT	TT
	PR36	AA	GG	TT	TT
	PR37	GA	GG	TT	TT
	PR40	GG	GG	TT	TT
	PR48	AA	GG	TT	TT
	PR53	GG	GG	TC	TC
	PR54	AA	GG	TT	TT
	PR69	AA	GA	TT	TT
	PR73	AA	GG	TT	TT
	PR83	AA	GG	TT	TT
Control	VDR160	GA	GG	CC	TT

3.6.3 Comparison of Ampliseq Nanopore, Ion Torrent AmpliSeq and PCR-RFLP

Genotypes obtained from AmpliSeq Nanopore and PCR-RFLP (n=13) were compared to those obtained from Ion Torrent sequencing (Table 3.11) which was used as a gold standard (all Qscores >20). Eleven (84.61 %) and 12 (92.30 %) genotypes agreed with Ion Torrent for rs2040410 and rs7454108, respectively, for Ampliseq Nanopore sequencing. Four (30.76 %) and 13 (100 %) genotypes agreed with Ion Torrent for rs2040410 and rs7454108, respectively, for PCR-RFLP.

Table 3.11: Comparison of genotypes generated using AmpliSeq Nanopore sequencing, Ion Torrent sequencing and PCR-RFLP

	AmpliSeq Nanopore		**Ion Torrent**		**PCR-RFLP**	
	rs2040410	**rs7454108**	**rs2040410**	**rs7454108**	**rs2040410**	**rs7454108**
PR11	GG	TT	GG	TT	GG	TT
PR27	GG	TT	GG	TT	AA	TT
PR30	GG	TT	GG	TT	AA	TT
PR36	GG	TT	GG	TT	AA	TT
PR37	GA	TT	GG	TT	GA	TT
PR40	GG	TT	GG	TT	GG	TT
PR48	GG	TT	GG	TT	AA	TT
PR53	GG	TC	GG	TC	GG	TC
PR54	GG	TT	GG	TT	AA	TT
PR69	GA	TT	AA	TT	AA	TT
PR73	GG	TT	GG	TT	AA	TT
PR83	GG	TT	GG	TT	AA	TT
VDR160	GG	CC	GG	TT	GA	TT

4 Discussion

T1D is a chronic autoimmune disease that is characterised by the destruction of the beta cells of the pancreas, this can cause irreparable damage to the normal functioning of the immune system, the homeostasis of the body and overall quality of life of the affected individual. Diagnosis of T1D in black population can be challenging because of the late age of onset and thus may be confused with other types of diabetes e.g., MODY, T2D, and LADA which can lead to incorrect treatment. The *HLA genes* confer the greatest risk of T1D development, however other genes (e.g., Insulin) also confer risk. Investigating a larger proportion of these susceptibility genes will allow for a more accurate genetic risk prediction of T1D for each individual. Screening for these at-risk individuals will allow for earlier detection and hence treatment, therefore reduced chance of complications especially ketoacidosis. Methods for typing HLA alleles including serology-based methods (complement-dependant cytotoxicity test) (37), and molecular based approaches such as phototyping of known HLA variants using PCR based sequence specific primers (PCR-SSP) have been developed for screening for T1D amongst individuals, but these methods can present challenges. These methods can be resource, time and cost inefficient and up for interpretation as some results are dependent on observation.

4.1 AmpliSeq Nanopore sequencing vs other methods

Nanopore sequencing is a long-read sequencing method, this method does not need PCR amplification as genomic DNA can be used to generate libraries for sequencing. Ion AmpliSeq is a method known for short read sequencing and genotyping by sequencing. Combining these two methods ensured that we get the best of both worlds at a fraction of the cost (~ $5 000 vs $1 000 000) (42,54). Since this method is a targeted approach focusing on the enrichment of target regions that contain SNPs of interest, Ion AmpliSeq was identified as the ideal method to do this. The primer pools that were used are a combination of 611 primer pairs which allowed for us to do a multiplex PCR for the amplification of all the target regions (56). Haplotype sequencing, which was initially investigated uses 144 primer pairs each (60), with each region needing to be

amplified in a separate PCR reaction. PCR optimization worked comparably better for DNA standards but when attempted on participant samples it did not work as well. This might be due to PCR concentrations and quality for the participant DNA, as DNA purity and concentration are dependent on the DNA extraction method used. The results obtained varied with some regions amplifying while others did not amplify enough for sequencing; thus, optimization had to be carried out for each participant and each HLA region. The haplotype sequencing method also only focused on the HLA genes; they have been found to confer a high risk of T1D susceptibility but other genes outside the HLA region have also been found to confer risk for T1D (22,59,65). The AmpliSeq Nanopore method is inclusive of regions within and outside the HLA region. As shown in section 2.4.7.1 and 2.4.7.2 samples can be multiplexed during sequencing runs, since we are sequencing short reads using sequencing technology that is ideal for long reads this ensures no data loss overall and cost and time efficiency. This is because multiple samples (up to 24 for this study) that are run on a MinION flow cell only use one flow cell, one sequencing machine, one sequencing run, and library preparation is carried out as you would when preparing one sample because each sample has a unique identifier in the form of a barcode.

The overall goal of this method is not to diagnose T1D but for the prediction of T1D susceptibility. This will ensure that individuals found to have susceptibility can be monitored and screened regularly, thus preventing misdiagnosis and micro- and macro- complications that can arise if the disease goes undiagnosed. We wanted to develop a screening tool that is cheap, user friendly, that has good predictive ability and does not require complex bioinformatics. Based on the methodology that has been developed, the process from DNA extraction, PCR amplification, and library preparation could be automated using the VolTRAX (66).

4.1.1 Sequencing samples individually in comparison to sequencing in a multiplex sequencing run

Samples sequenced individually where compared to those sequenced in a multiplex sequencing run. We found that samples sequenced individually had a higher Qscore (9.73) than those sequenced in multiplex (9.5). Samples sequenced individually had higher average reads generated (187.60 Mb), passed reads (41.47 Mb), and estimated bases (54.72 Mb) than those sequenced in multiplex (151.14 Mb, 25.99 Mb, 49.57 Mb), respectively. Samples sequenced in a multiplex sequencing run had higher average failed reads (475.58 Mb) in comparison to those sequenced individually (13.58 Mb). This might be due to the fact that significantly higher DNA concentrations were used when sequencing samples individually (20 Fmol) compared to in multiplex (2 Fmol). Thus there was a greater probability of obtaining more reads with a higher Qscore for samples sequenced individually. Although sequencing in a multiplex sequencing run saves time and cost as each library is treated as though only one samples is being run during the preparation process, we found that based on the above, for this population sequencing individually was more ideal than sequencing in a multiplex sequencing run.

4.1.2 Advantages of using Nanopore sequencing as a sequencing method

Nanopore sequencing allows an individual to view sequencing run information in real time. This is advantageous as errors such as poor library loading (thus killing off pores) can be detected at the beginning of the sequencing run. The user can therefore cancel the run, flush the flowcell and reuse the flowcell to load the same library correctly, thus saving time and ensuring that each sequencing run yields adequate data. The instrument provides a map of sequencing pores that shows the presence of adapters that were not properly cleaned out in the purification steps during library preparation that got carried into the sequencing procedure. This affects sequencing negatively as unbound adapters can saturate the pores preventing translocation of DNA to be sequenced (42) thus this informs the user to be more meticulous when preparing the library so that only sequencing of the target sequences is achieved. The read length histograms act to ensure that the desired regions are being sequenced. This is done by comparing the expected

fragment lengths and the sequencing read lengths, taking into account the addition of primers, adapters and possibly barcodes.

The data file produced after sequencing gives valuable information about Qscore and data generation as a function of time. This allows us to determine what the ideal sequencing time for the samples would be. As shown in Figure 3.4, the Qscore dropped below the ideal range after a period of approximately 17 hours this is due to the depleting ATP concentrations on the flow cell as well as depleting DNA to be sequenced (41). The data generated after this time is unusable as the sequences generated cannot be trusted. Thus, sequencing runs can be set for only the duration of time that allows for desired Qscore which saves sequencing and analysis time. The read graph also allows us to make this choice, as if we set the desired reads that we want we can set the time of the sequencing run to align with this. We could also stop the sequencing run when this desired read count has been achieved which further saves sequencing and data processing time.

4.1.3 Improvement in read quality

Average quality scores that were obtained during sequencing were 9.73 for samples run individually and 9.5 for samples run in a multiplex sequencing run. It is important to note that Nanopore sequencing has not been known for high read qualities (53–55). This, however, is only when considering single read quality scores. This method however has multiple single reads for the same areas; this has been shown to increase read quality as the combination of all the single reads increases the confidence in correctly basecalled bases. Nanopore sequencing is constantly trying to improve read quality and has been shown in some instances to be comparable to gold standard sequencing technologies (42,43). The use of the newly developed flow cell, the R10.4, can ensure higher accuracy as this flow cell is said to have read qualities of > Q30 (43,44). For data generated during the sequencing runs, this can ensure that we get up to 1000x coverage (ideal is 100x). We found that ± 120 000 reads are required for each individual to ensure adequate information for genotyping to ensure 100x coverage. ONT allows for real time basecalling and viewing of the sequencing run which allowed us to view if the desired regions were amplified and

being sequenced based on size selection. Gene targets, ranging from a few to hundreds, can also be sequenced in as little as 24 hours at half the cost (67,68).

4.1.4 Alternative approaches for genetic testing in comparison to AmpliSeq Nanopore sequencing

Colorimetric endpoint PCR, quantitative PCR (qPCR) (69) and cardiac autonomic neuropathy (CAN) screening (70) based points of care are alternative cheap and quick methods that can be used for genetic testing. These methods have been used extensively for T1D (71). These methods, however, are not suitable for polygenic diseases (72). AmpliSeq Nanopore sequencing allows us to have a method that has the benefit of the methods mentioned and the added benefit of screening for diseases that are polygenic in nature such as T1D. Because of its polygenic nature, it might be ideal that instead of focusing on whole genome sequencing the focus should be on highly multiplexed genotyping by sequencing methods. This ensures that we are covering a large number of variants whilst also being targeted.

4.1.4.1 Alternative methods to identify novel African variants which contribute to T1D susceptibility

As the South African black population has a highly polymorphic genome, whole exome sequencing (WES) which targets the protein coding region of genes (40) and whole genome sequencing, which target the entire genome of an organism (41) can be used as alternative methods to identify novel African variants. These variants can ultimately be used to develop a screening tool for T1D. These methods are superior for discovering novel variants especially in the South African population as unlike a targeted approach they sequence all regions of the genome/exome (42) and not only those previously associated with T1D. This is advantageous in this case since sequencing larger regions may help us discover novel variants that might have been missed using a targeted approach.

4.2 AmpliSeq Nanopore sequencing validated as a method for screening for T1D

We validated that this method can be used to viably sequence and screen for the various SNPs present in the target region. This was done by using Ion Torrent to sequence the same samples and taking sequencing data generated and comparing it to that generated by Ampliseq Nanopore sequencing.

One of the other ways validation could be carried out was comparing variant calling outputs from data from both sequencing platforms. RTG-tools were used to compare VCF files as they are highly accurate. We found that for data that was comparable the average percentage difference between data sequenced using Ion Torrent and Ampliseq Nanopore was 5.67 %, which is good since this means that > 93 % of the data obtained from sequencing using Ion Torrent was similar to that obtained using Nanopore sequencing, the discrepancy is as a result of mismatched bases between those generated using Ion Torrent and AmpliSeq Nanopore sequencing. As with most sequencing platforms errors occur during sequencing. Errors can occur during Nanopore sequencing due to poor sequence quality, high adapter concentrations or inadequately purified sequencing products (42). The adapters facilitate the aggregation of DNA to the membrane thus ensuring that the DNA can be captured into the nanopore (42), however high adapter concentrations can cause the adapters to saturate pores resulting in poor sequence qualities. In addition, the discrepancies may have arisen due to high GC and/or long polymer regions in the target regions which are known to cause errors when performing Nanopore sequencing (40-44). Ion Torrent sequencing was performed to validate the sequences obtained using Ampliseq Nanopore sequencing. The Ion Torrent sequencing generated high Qscores for all samples sequenced (Qscores >20) indicating a 99% accuracy in base calling. Thus it is likely that the discrepant base calling is due to the low Qscores (<10) obtained for samples sequenced using the Ampliseq Nanopore method which represent a 90 % accuracy in base calling. Ion torrent is the gold standard for NGS (40) and thus it can be assumed that these sequences are the correct result. However, Sanger sequencing can be done to confirm which platform generated the correct sequence.

4.3 Bioinformatics analysis

The analysis of whole HLA regions requires a lot of processing power thus requiring a large computer cluster (60). By sequencing selective variants for T1D using short reads, this has allowed for the reduction of the amount of sequencing required, which reduces the data generated during sequencing thus reducing the data needed for analysis and has simplified the bioinformatics analysis to a point that all that is needed is a laptop to carry out the bioinformatics analysis. Nanopore sequencing also allows for their data to be easily analysed with third party software as the user can specify ideal parameters for the analysis of their sequencing data. Third party software such as iGenomics (https://github.com/stuckinaboot/iGenomics) which is software designed to generated high confidence base calls with a low read depth covering a large number of variants can be investigated. This software can be used on an iPad or iPhone for analysis and can call risk genotypes thus allowing the user to generate GRS. This app requires no bioinformatics background thus ensuring that anyone can use it.

4.3.1 Variant calling quality scores

We had average Phred-scale quality scores of 73.89 for standards and 89.77 for participants. The ideal read quality for a Phred-scale quality score is ≥ 20 but the higher the quality score the better. By using the R10.4, instead of the R9.4.1 flow cell we can improve read quality thus increasing variant calling quality (73). Investigation of other variant callers more suitable for data generated during Nanopore sequencing such as CLAIR3 (https://github.com/HKU-BAL/Clair3) might also improve variant calling quality scores; just as Minimap2 was chosen for sequence alignment as it is designed to be better suited for data generated from Nanopore sequencing.

4.4 Genetic risk scores obtained from literature in comparison to this population

Odds ratios obtained from GWAS studies were used because they are statistically significant because of the study sizes as opposed to those that would be generated from this population currently, as the sample size is too small to generate a statistically significant beta coefficient thus would not generate statistically significant odds ratios. Which would give inaccurate GRS'.

The GRS calculator was not able to provide sufficient resolution between the non-diabetic control and T1D participants. This was expected as the population is too small to have a significant resolution. HLA subtypes form a very big proportion of the score to determine T1D in GRS calculators. In the GRS calculator, two SNPs (out of 30) account for 58 % of the GRS score, these are called tag SNPs. This is ideal in situations where the populations studied have a high occurrence of the target allele, this however is not true for this population. When considering the two SNPs that are used as tag SNPs in the GRS1 calculator: rs7454108 which is a SNP that has been found to have association with DR4 (34,65) and rs2187668 which is found on the *HLA-DQA1 gene (https://www.ncbi.nlm.nih.gov/snp/?term=rs2187668)* , in Caucasian populations their alternative alleles frequencies are 10.2 % and 12.0 % respectively while in African populations their frequencies are 3.8 % and 6.0 % respectively (https://www.ncbi.nlm.nih.gov/snp/). Populations statistics suggest that it is 5X less likely for both the SNPs to be heterozygous in the African population in comparison to the Caucasian population. The DR3/DR4-DQ8 HLA type is assigned to a patient when both the SNPs are heterozygous for the variant allele; this combination carries the highest GRS weighting. In this population only one participant was found to be heterozygous for the rs7454108. This is expected as the likelihood of finding these SNPs at high rates is unlikely. This poses the question of whether these GRS calculators that have a high rating for specific SNPS, especially those that are found significantly more in some populations than others are ideal to develop an unbiased bioinformatics pipeline to study and determine T1D susceptibility. Since the GRS that was used in this study has a more complex design because of the basis being with interactions for HLA tag-SNPS, this results in a higher error rate for identifying genotypes caused by technical factors. In order to observe more reliable scores for these calculators, the calculators themselves might have to be modified to account for differences in populations. This however, might alter the predictive ability of the GRS calculator. Perhaps genetic markers as opposed to tag SNPs should be investigated or developing this study's own GRS using a larger population size is ideal to generated better results.

4.4.1 Rationale for using 12 SNPs for GRS analysis

This was a pilot study to determine whether a European based GRS calculator would be able to successfully discriminate between participants with and without T1D in an African population. Data was only available for 12 of the 30 required SNPs and therefore for the remaining 18 SNPs we had to use the GRS from the literature. Thus it was not surprising that the GRS calculator was unable to correctly classify participants with T1D. To improve on results an African specific GRS would need to be developed using variants with a higher GRS in the African population.

4.5 PCR-RFLP of rs2040410 was discordant with the Ampliseq Nanopore sequencing results

The genotyping results for rs2040410 and rs7454108 from PCR-RFLP did not agree with the Ampliseq Nanopore sequencing results. Possible reasons for the discrepant findings between Ampliseq Nanopore sequencing and PCR-RFLP include incomplete digestion due to the restriction enzyme not working which may be a result of incorrect concentration of enzyme to PCR product, too much glycerol inhibiting the enzyme or incorrect variant calling for Nanopore Sequencing data which might be due to the sensitivity of the variant caller. Parameters, such as changing digestion enzymes, changing enzyme concentrations, changing digestion times, removing reagents that might cause inhibition during digestion, and using different variant callers, for the PCR-RFLP were changed to try to eliminate these discrepancies but the results remained the same. However, Sanger sequencing should have been done to confirm which method was correct.

4.5.1 Alleles that were found to confer T1D risk in this population

The A allele for rs2040410 was found to be more common in the T1D participants than the control participants (p=0.039), the TC genotype was found to be more common in the control participants (p= 0.0012) for the rs7454108, and the T allele was found to be more common in the T1D participants (p=0.000895) for the rs7454108. The above mentioned results were all statistically significant. The results obtained for the rs2040410 SNP correspond to those obtained from the Barker study (34) but those obtained for the rs7454108 do not. This might mean that

the T allele confers risk for T1D in this population but not particularly in the European population that the study was conducted on. A greater sample size is necessary to confirm these findings.

4.6 Study limitations

4.6.1 Primer coverage for the desired SNPs

The SNPs associated with T1D that were analysed in this study were obtained from the literature conducted on predominantly Caucasian populations from North America and Europe. Previous studies have shown that disease aetiology can be different and present differently in different populations dependant on race, gender, age, and environmental factors. Thus, this primer pool used in the current study might not have been appropriate for the South African black population. The investigated SNPs can also perform differently dependant on the relationship and association between other SNPs i.e., SNP A can perform differently dependant on whether it is present individually or concurrently with SNP B and this can either increase or decrease disease susceptibility. This is due to the linkage disequilibrium between genes implicated in T1D. This is also dependant on the GRS calculator that was used as each has its own parameters for conducting the calculations. Developing a calculator that is more suitable for this population could be designed and investigated to represent this population more accurately.

4.6.2 Small sample size

This investigation was for the validation of a novel screening method as well as a comparison for methods that are currently being used for T1D screening. Most of the SNPs that were investigated have a lower prevalence in the black population in comparison to the Caucasian population (https://www.ncbi.nlm.nih.gov/snp/); thus, we would need a greater sample size to identify these alleles in this population. Some of these SNPs had a greater weighting in the calculator comparatively which greatly affected the susceptibility result overall. This, however, presents an opportunity to identify novel SNPs as well as SNPs that weigh higher in this population. This will allow for better screening and identification of T1D in the black population overall.

4.6.2.1 Power of the study and sample size needed

The power of the study was calculated using a sample size calculator (https://clincalc.com/stats/samplesize.aspx) with SNPs that are used as tag SNPs in the GRS1 calculator: rs7454108 which is a SNP that has been found to have association with DR4 (34,65) and rs2187668 which is found on the *HLA-DQA1 gene* (*https://www.ncbi.nlm.nih.gov/snp/?term=rs2187668*) , and frequencies that the SNPs are found in the Caucasian populations and their alternative alleles frequencies (10.2 % and 12.0 %, respectively) and their frequencies in the African populations (3.8 % and 6.0 %, respectively). A dichotomous endpoint, one sample study with the alpha, beta and power values set at 0.05, 0.2 and 0.8, respectively. It was found that a sample size required would be 139 and 195 for each SNP. Thus a sample size of at least 200 participants would be required to get comparable, statistically significant results to compare all methods used for this study. This study had nowhere near the desired sample size thus results cannot be deemed conclusive as a larger sample size would be necessary to validate the results. Literature has shown that as few as 15 samples can be used for NGS method validation (75) and that the sample size was adequate to demonstrate precision, accuracy and sensitivity of the method. Thus, the sample size of 24 was considered sufficient to validate the new method.

4.6.3 Poor sample quality

The samples that were available for the study were of poor quality as they had been in storage for a very long time. This hindered proper amplification, sequencing and analysis which also affected data yield negatively. Thus for future studies better quality samples are pivotal for improved data yield and method validation.

4.6.4 Failure to generate sequencing data

Due to the poor quality of samples that were used for this research, some samples (n=4) did not pass the initial quality control (i.e. minimum DNA concentration of 15 Fmol) thus could not be sequenced. In addition, sufficient data (± 120 000 reads) was not obtained from six samples despite repeated sequencing attempts and therefore accurate base calling was not possible.

Samples that were sequenced in multiplex could not always be sequenced individually due to the concentration requirements for single vs multiplex sequencing runs. A concentration of 15-20 Fmol for each sample was required for an individual sequencing run and 20 Fmol total concentration was required for a multiplex run. Thus if 10 samples were run at a time, 2 Fmol of each sample was required for a multiplex run. Therefore, samples with a low DNA concentration could have had sufficient DNA for a multiplex sequencing run but not for an individual sequencing run. Multiple repeats were carried out for some samples but the results remained the same. Thus analyses were only carried out on samples that had sufficient sequencing data. Ideally the same samples would be used for all the analyses as this would be a better comparison for the different methodologies.

5 Conclusion

5.1 Summary of findings

A method was investigated to genotype participants for T1D associated SNPs with approximately 90% accuracy. The method combined two sequencing techniques i.e., Ion AmpliSeq and Nanopore sequencing to achieve this. A bioinformatics pipeline to analyse the generated data was also developed. The data can then be processed by in-house variant callers. The data was initially analysed on The Coriell Trio as this trio can be used as standards for validating unknown data. The required reads for sequencing data analysis are ± 120 000 reads which equates to approximately 17 hours of sequencing to obtained 100 x coverage. The average quality scores generated were 9.73 and 9.5 for sample sequences individually and samples sequenced in multiplex using barcodes (ideal 9), respectively. For this population, sequencing individually as opposed to in a multiplex sequencing run was ideal. The average percentage difference between Ion Torrent data and Nanopore sequencing data was 5.67 %. Average Phred-scale quality scores observed after variant calling were 73.89 and 89.77 for standards and participants (ideal ≥ 20). GRS calculators did not provide sufficient resolution between T1D participants and controls due to sample size. We identified incorrect allele calling that were generated using PCR-RFLP for these samples. The A allele for rs2040410 was found to be more common in the T1D participants than the control participants (p= 0.039), the TC genotype was found to be more common in the control participants (p= 0.0012) for the rs7454108, and the T allele was found to be more common in the T1D participants (p= 0.000895) for the rs7454108; these results were statistically significant.

Due to the low Qscores and discrepant results between Ampliseq Nanopore and PCR-RFLP we were unable to confirm the accuracy of this method. Currently this method cannot be used as a screening tool for individuals at risk of T1D development until further improvements in the methodology have been made. Once this data is available, a GRS calculator specific to black participants can be developed.

5.2 Implication of research

AmpliSeq Nanopore was designed to use minimal resources, time and space. The method also allows for third-party analysis to be carried out thus allowing for customisation by the user depending on their needs. This method was explored to improve upon the methods that are already out there for T1D screening. The aim is to ensure that an unbiased screening and analysis of patient data is carried out and even populations that have limited research data (such as the black population of South Africa) can have access to accurate T1D screening.

5.3 Future studies

A larger sample size would have to be investigated to further improve and validate these results and to have higher accuracy results. A wider ranged SNP primer pool can be investigated in future studies to involve SNPs that might have been overlooked due to the study population; this would then allow a more specific primer design for target region amplification. The bioinformatics pipeline is aimed to be automated so that all the user needs to do is enter sequencing data and the software will generate the probability of T1D susceptibility. A further simplified research methodology can also be investigated to further reduce time, resources, and cost. Use of the newly developed flow cells such as the R10.4 would allow for improved sequencing and variant calling accuracy. This research is very promising and will be very beneficial to the South African population.

5.3.1 Machine learning classifiers more suitable for T1D susceptibility identification

As part of a post-Doctorate research project machine learning (ML) to determine T1D susceptibility was investigated. This work is outside the scope of this research, but important information was discovered about the ML classifiers. The ML algorithm was first trained on the Teddy cohort as it is centred around T1D, and it is a large data set which is ideal for machine learning. Using a larger cohort like the TEDDY cohort for machine learning is ideal in situations that are not dependant on investigated parameters being reliant on races of study participants. While the Teddy cohort is a large data set it is made up of 73 % Caucasian participants and only 10 % African American participants. Classifiers use a probability of 0.5 as a cut-off to assign classes

in a binary output classifier, with a probability of >0.5 indication disease susceptibility. Classifiers, however, can be optimised to output probabilities instead of the binary output. This approach was employed in the study. We found that the probabilities were greatly affected based on the choice of input sample. When the 1D-CNN was trained using data from the parents, the probability for AmpliSeq samples was 0.11, 0.77 when using data from children and 0.37 when all samples were used for training. The classifiers were found to be more consistent even with missing data and other errors in variant calling. As a result, ML platforms have demonstrated to be an ideal choice for ONT sequencing data as ONT has been shown to have higher error rates as compared to other sequencing platforms. ML have also been found to be ideal for this study population in particular, as unlike GRS calculators, MLs are not heavily reliant of HLA haplotypes to determine susceptibility. A larger population is required however to further validate these results as well as to verify the cut-off values for T1D susceptibility. These ML have given an accuracy of 96.34 % with training, which considering the limited data set is not bad.

References

1. ADA. Diagnosis and classification of diabetes mellitus. American Diabetes Association. Diabetes Care. 2013.
2. Rose NR, Bona C. Defining criteria for autoimmune diseases (Witebsky's postulates revisited). Immunol Today. 1993;14(9):426–30.
3. Eisenbarth GS. Update: Update in type 1 diabetes. Journal of Clinical Endocrinology and Metabolism. 2007;92(7):2403–7.
4. Gianani R, Campbell-Thompson M, Sarkar SA, Wasserfall C, Pugliese A, Solis JM, et al. Dimorphic histopathology of long-standing childhood-onset diabetes. Diabetologia. 2010;53(4):690–8.
5. Classification and diagnosis of diabetes. Diabetes Care. 2017;
6. Gale EAM. Type 1 diabetes in the young: The harvest of sorrow goes on. Diabetologia. 2005;48(8):1435–8.
7. Leslie RD. Predicting adult-onset autoimmune diabetes clarity from complexity. Diabetes. 2010;59(2):330–1.
8. Diabetes DOF. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2012;35(SUPPL. 1).
9. Daniel Fulford SLJ. 基因的改变NIH Public Access. Bone [Internet]. 2008;23(1):1–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf
10. Steck AK, Johnson K, Barriga KJ, Miao D, Yu L, Hutton JC, et al. Age of islet autoantibody appearance and mean levels of insulin, but not GAD or IA-2 autoantibodies, predict age of diagnosis of type 1 diabetes: Diabetes autoimmunity study in the young. Diabetes Care. 2011;34(6):1397–9.
11. Notkins AL, Lernmark Å. Autoimmune type 1 diabetes : resolved and unresolved issues Find the latest version : Autoimmune type 1 diabetes : resolved and unresolved issues. J

Clin Invest [Internet]. 2001;108(9):1247–52. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC209446/pdf/JCI0114257.pdf

12. Fong P, Boss D, Yap T, Tutt A, Wu P, Mergui-Roelvink M. New England Journal CREST. Science (1979). 2010;609–19.
13. Hosokawa Y, Hanafusa T, Imagawa A. Pathogenesis of fulminant type 1 diabetes: Genes, viruses and the immune mechanism, and usefulness of patient-derived induced pluripotent stem cells for future research. J Diabetes Investig. 2019;10(5):1158–64.
14. Maahs DM, West NA, Lawrence JM, Mayer-Davis EJ. Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am [Internet]. 2010 Sep 1 [cited 2020 Feb 7];39(3):481–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20723815
15. Harjutsalo V, Sjöberg L, Tuomilehto J. Time trends in the incidence of type 1 diabetes in Finnish children: a cohort study. The Lancet. 2008;
16. Patterson CC, Dahlquist GG, Gyürüs E, Green A, Soltész G, Schober E, et al. Incidence trends for childhood type 1 diabetes in Europe during 1989-2003 and predicted new cases 2005-20: a multicentre prospective registration study. The Lancet. 2009;
17. Kalk WJ, Huddle KRL, Raal FJ. The age of onset and sex distribution of insulin-dependent diabetes mellitus in Africans in South Africa. Postgrad Med J. 1993;69(813):552–6.
18. Padoa CJ. The epidemiology and pathogenesis of type 1 diabetes mellitus in Africa. Vol. 16, Journal of Endocrinology, Metabolism and Diabetes of South Africa. South African Medical Association; 2011. p. 130–6.
19. Mobasseri M, Shirmohammadi M, Amiri T, Vahed N, Fard HH, Ghojazadeh M. Prevalence and incidence of type 1 diabetes in the world: A systematic review and meta-analysis. Vol. 10, Health Promotion Perspectives. Tabriz University of Medical Sciences; 2020. p. 98–115.
20. Vehik K, Dabelea D. The changing epidemiology of type 1 diabetes: Why is it going through the roof? Vol. 27, Diabetes/Metabolism Research and Reviews. 2011. p. 3–13.
21. Yoon JW, Jun HS. Autoimmune destruction of pancreatic β cells. Am J Ther. 2005;12(6):580–91.

22. Oram RA, Patel K, Hill A, Shields B, McDonald TJ, Jones A, et al. A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. Diabetes Care. 2016 Mar 1;39(3):337–44.
23. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451(7181):998–1003.
24. Buniello A, Macarthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D1005–12.
25. Janssens ACJW, Moonesinghe R, Yang Q, Steyerberg EW, Van Duijn CM, Khoury MJ. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genetics in Medicine. 2007 Aug;9(8):528–35.
26. Undlien DE, Lie BA, Thorsby E. HLA complex genes in type 1 diabetes and other autoimmune diseases. Which genes are involved? Trends in Genetics. 2001;17(2):93–100.
27. De RK, Tomar N. Preface. Immunoinformatics. Methods Mol Biol. 2014;1184:vii–vixi.
28. Nerup J, Platz P, Andersen OO, Christy M, Lyngs o e J, Poulsen JE, et al. HL-A ANTIGENS AND DIABETES MELLITUS. The Lancet. 1974;
29. Kumar N, Kanga U. Biomarkers of susceptibility to type I diabetes with special reference to the Indian population [Internet]. Article in The Indian Journal of Medical Research. 2007. Available from: https://www.researchgate.net/publication/6336320
30. del Guercio MF, Sidney J, Hermanson G, Perez C, Grey HM, Kubo RT, et al. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J Immunol [Internet]. 1995;154(2):685–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7529283
31. Lund O, Nielsen M, Kesmir C, Petersen AG, Lundegaard C, Worning P, et al. Definition of supertypes for HLA molecules using clustering of specificity matrices. Immunogenetics. 2004;55(12):797–810.
32. Doytchinova IA, Flower DR. In Silico Identification of Supertypes for Class II MHCs. The Journal of Immunology. 2005;174(11):7085–95.

33. Saha I, Mazzocco G, Plewczynski D. Consensus classification of human leukocyte antigen class II proteins. Immunogenetics. 2013;65(2):97–105.
34. Barker JM, Triolo TM, Aly TA, Baschal EE, Babu SR, Kretowski A, et al. Two single nucleotide polymorphisms identify the highest-risk diabetes HLA genotype; Potential for rapid screening. Diabetes. 2008;57(11):3152–5.
35. Bakker PIW De, Mcvean G, Sabeti PC, Miretti MM, Marchini J, Ke X, et al. Europe PMC Funders Group Europe PMC Funders Author Manuscripts A high resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2009;38(10):1166–72.
36. Borchers AT, Uibo R, Gershwin ME. The geoepidemiology of type 1 diabetes. Autoimmun Rev [Internet]. 2010;9(5):A355–65. Available from: http://dx.doi.org/10.1016/j.autrev.2009.12.003
37. Dyer PA, Tissue NR, Road H. Techniques used to define human MHC antigens : serology. 1991;29.
38. Bunce M, O'Neill CM, Barnardo MCNM, Krausa P, Browning MJ, Morris PJ, et al. Phototyping: comprehensive DNA typing for HLA-A, B, C, DRB1, DRB3, DRB4, DRB5 & DQB1 by PCR with 144 primer mixes utilizing sequence-specific primers (PCR-SSP). Tissue Antigens. 1995;46(5):355–67.
39. Raha O, Sarkar B, Vks Lakkakula B, Pasumarthy V, Godi S, Chowdhury S, et al. HLA class II SNP interactions and the association with type 1 diabetes mellitus in Bengali speaking patients of Eastern India [Internet]. 2013. Available from: http://www.jbiomedsci.com/content/20/1/12
40. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. 2008;5(12):1005–10.
41. Merriman B, Torrent I, Rothberg JM. Progress in Ion Torrent semiconductor chip based sequencing. Vol. 33, Electrophoresis. Wiley-VCH Verlag; 2012. p. 3397–417.

42. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Vol. 2022, BioMed Research International. Hindawi Limited; 2022.
43. ONT. Nanopore sequencing accuracy [Internet]. 2022 [cited 2023 Mar 20]. Available from: https://nanoporetech.com/accuracy
44. ONT. Nanopore devices cost, read quality and depth break down [Internet]. 2022 [cited 2023 Mar 20]. Available from: https://interpriseusa.com/wp-content/uploads/2021/05/MinION-Mk1C-brochure.pdf
45. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Vol. 34, Nature Biotechnology. Nature Publishing Group; 2016. p. 518–24.
46. Stockton JD, Nieto T, Wroe E, Poles A, Briggs D, Beggs AD, et al. Rapid , highly accurate and cost-effective open-source simultaneous complete HLA typing & phasing of Class I & II alleles using Nanopore sequencing . 1 = Institute of Cancer & Genomic Sciences , University of Birmingham 2 = NHS Blood & Transplant 3 = Quee.
47. Liu C, Yang X, Duffy BF, Hoisington-Lopez J, Crosby ML, Porche-Sorbet R, et al. High-resolution HLA typing by long reads from the R10.3 Oxford nanopore flow cells. Hum Immunol. 2021 Apr 1;82(4):288–95.
48. Liu C. A long road/read to rapid high-resolution HLA typing: The nanopore perspective. Vol. 82, Human Immunology. Elsevier Inc.; 2021. p. 488–95.
49. ONT. Sequencing machine comparison [Internet]. 2023 [cited 2023 Mar 20]. Available from: https://nanoporetech.com/products/minion
50. ONT. Nanopore Sequencing Mechanism. 2023 [cited 2023 Mar 20]; Available from: https://nanoporetech.com/applications/dna-nanopore-sequencing
51. ONT. Nanopore library preparation [Internet]. 2020 [cited 2023 Mar 9]. Available from: https://store.nanoporetech.com/ligation-sequencing-kit.html
52. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Vol. 39, Nature Biotechnology. Nature Research; 2021. p. 1348–65.
53. Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo JA, Manzano-Agugliaro F. DNA sequencing sensors: An overview. Vol. 17, Sensors (Switzerland). MDPI AG; 2017.

54. Meslier V, Quinquis B, Da Silva K, Plaza Oñate F, Pons N, Roume H, et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data. 2022 Dec 1;9(1).
55. Ranasinghe D, Jayadas TTP, Jayathilaka D, Jeewandara C, Dissanayake O, Guruge D, et al. Comparison of different sequencing techniques for identification of SARS-CoV-2 variants of concern with multiplex real-time PCR. PLoS One. 2022 Apr 1;17(4 April).
56. ThermoFischer Scientific. Ion AmpliSeq Custom Next-Generation Sequencing (NGS) DNA Panels [Internet]. [cited 2023 Jan 20]. Available from: Ion AmpliSeq Custom Next-Generation Sequencing (NGS) DNA Panels
57. Igo RP, Kinzy TG, Cooke Bailey JN. Genetic Risk Scores. Curr Protoc Hum Genet. 2019 Dec 1;104(1).
58. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: Markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet. 2009 Feb;5(2).
59. Sharp SA, Rich SS, Wood AR, Jones SE, Beaumont RN, Harrison JW, et al. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis.
60. Stockton JD, Nieto T, Wroe E, Poles A, Inston N, Briggs D, et al. Rapid, highly accurate and cost-effective open-source simultaneous complete HLA typing and phasing of class I and II alleles using nanopore sequencing. HLA. 2020 Aug 1;96(2):163–78.
61. Sharp et al.
62. Onengut et al.
63. Holmberg D, Ruikka K, Lindgren P, Eliasson M, Mayans S. Association of CD247 (CD3ζ) gene polymorphisms with T1D and AITD in the population of northern Sweden. BMC Med Genet. 2016 Oct 4;17(1).
64. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007 Jul;35(SUPPL.2).

65. De Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006 Oct;38(10):1166–72.
66. ONT. VolTRAX Nanopore Sequencing [Internet]. 2021 [cited 2023 Mar 23]. Available from: https://nanoporetech.com/products/voltrax
67. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases.
68. Gardner SN, Frey KG, Redden CL, Thissen JB, Allen JE, Allred AF, et al. Targeted amplification for enhanced detection of biothreat agents by next-generation sequencing. BMC Res Notes. 2015 Nov 16;8(1).
69. Oikarinen S, Krogvold L, Edwin B, Buanes T, Korsgren O, Laiho JE, et al. Characterisation of enterovirus RNA detected in the pancreas and other specimens of live patients with newly diagnosed type 1 diabetes in the DiViD study. Diabetologia. 2021 Nov 1;64(11):2491–501.
70. Pittasch D, Lobmann R, Behrens-Baumann W, Lehnert H. Pupil Signs of Sympathetic Autonomic Neuropathy in Patients With Type 1 Diabetes.
71. Diabetic retinopathy screening: a short guide Increase effectiveness, maximize benefits and minimize harm.
72. Pescarmona R, Belot A, Villard M, Besson L, Lopez J, Mosnier I, et al. Comparison of RT-qPCR and Nanostring in the measurement of blood interferon response for the diagnosis of type I interferonopathies. Cytokine. 2019 Jan 1;113:446–52.
73. ONT. Flow cells [Internet]. 2022 [cited 2023 Mar 23]. Available from: https://store.nanoporetech.com/flow-cells.html
74. Lawrence J. Jennings*, †. M.-R.-S. (2017). Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels:. *Molecular Diagnosis*.
75. Alisen Ayitewala*, I. S. (2021). Next generation sequencing based in-house HIV genotyping method: validation report. *AIDS Reseach and Therapy* .

Websites

https://www.ncbi.nlm.nih.gov/search/all/?term=rflp, n.d.

https://usegalaxy.org

https://www.snapgene.com

https://genome.ucsc.edu

https://www.ncbi.nlm.nih.gov

https://www.ncbi.nlm.nih.gov/genome/?term=human+genome

https://github.com/lh3/minimap2

https://www.htslib.org/download/

https://www.htslib.org/download/

https://research.google.com/colaboratory/faq.html

https://www.coriell.org/1/NIGMS/Collections/NIST-Reference-Materials

https://worldwide.promega.com/products/biochemicals-and-labware/nucleic-acids/human-genomic-dna/?catNum=G1521

https://www.socscistatistics.com/tests/chisquare2/default2.aspx

https://github.com/HKU-BAL/Clair3

https://github.com/stuckinaboot/iGenomics

https://www.ncbi.nlm.nih.gov/snp/?term=rs2187668

https://bio.tools/t?topicID=%22topic_3517%22

https://clincalc.com/stats/samplesize.aspx

Appendix

Appendix A

INFORMATION LEAFLET AND INFORMED CONSENT FORM FOR ADULTS FOR CLINICAL STUDY / INTERVENTION RESEARCH

1. STUDY TITLE

Pheno- and genotypic characterization of patients diagnosed with diabetes mellitus before the age of 40 in South Africa.

2. THE NATURE AND PURPOSE OF THIS STUDY

This is a research study. The aim of this study is to evaluate the role genes and antibodies play in causing various types of diabetes.

Diabetes affects different age groups:

Babies Kids Young Adults Older Adults

Some diabetics are very thin, and some are overweight. Why diabetes should affect these different types of people is not very clear.

Genes are present in every cell of your body, and they are responsible for the differences we see between people (like hair colour for example). Different genes may also explain why some people get diabetes and why some people don`t. Antibodies as molecules that your body makes to protect itself against strange proteins in your body (your body makes antibodies against germs for example). Sometimes the body makes antibodies against itself. It can make against antibodies against insulin for example and that may cause diabetes.

We want to study the genes and antibodies in people with diabetes who were diagnosed before 40 years of age as there are different types of diabetes possible in this age group.

3. EXPLANATION OF PROCEDURES TO BE FOLLOWED

This study involves answering some questions with regard to your diabetes weight, height, hip and waist measurements and blood. This will all be done on the same day as you are visiting the clinic. Three blood samples (5ml each) will be taken, and some blood will be analysed and some blood samples will be stored for testing at a later stage. Any tests done on these blood samples will only be related to diabetes.

4. RISK AND DISCOMFORT INVOLVED

The only risk and discomfort involved is the taking of blood from a vein and the measurement of your waist and hips.

5. POSSIBLE BENEFITS OF THIS STUDY

The findings of a study such as this will not benefit you personally but will lead to a better understanding of how diabetes and possibly its complications are caused and may lead to new treatments in future. For those who do not have diabetes we will be testing your glucose so that if you have high blood glucose (diabetes) when we test you, you will be referred for the right treatment.

6. I understand that if I do not want to partake in this study, I will still receive standard treatment for my illness.

7. I may at any time withdraw from this study.

8. INFORMATION

If I have any questions concerning this study, I should contact: Dr C Padoa, Tel: 011 489 8514.

9. CONFIDENTIALITY

All records obtained whilst in this study will be regarded as confidential. Results will be published or presented in such a fashion that patients remain unidentifiable.

CONSENT TO PARTICIPATE IN THIS STUDY

I have read or had read to me in a language that I understand the above information before signing this consent form. The content and meaning of this information have been explained to me. I have been given opportunity to ask questions and am satisfied that they have been answered satisfactorily. I understand that if I do not participate it will not alter my management in any way. I hereby volunteer to take part in this study.

I have received a signed copy of this informed consent agreement.

…… ……

Patient's name Patient's signature

Date……

…… ……

Investigator's name Investigator's signature

Date……

VERBAL PATIENT INFORMED CONSENT (applicable when patients cannot read or write)

I, the undersigned, Dr ……, have read and have explained fully to the patient, named …….. and/or his/her relative, the patient information leaflet, which has indicated the nature and purpose of the study in which I have asked the patient to participate. The explanation I have given has mentioned both the possible risks and benefits of the study and the alternative treatments available for his/her illness. The patient indicated that he/she understands that he/she will be free to withdraw from the study at any time for any reason and without jeopardising his/her further treatment.

I hereby certify that the patient has agreed to participate in this trial.

....

Patient's name Patient's signature

Date.....

....

Investigator's Name Investigator's Signature

Date.....

Appendix B

Questionnaire:

Demographics

Study number
Date of visit (dd/mm/yyyy):

5.

5.4 Biogram

Date of birth (dd/mm/yyyy):
Gender: M / F
Race:
Hospital number:
Clinic code:
Telephone number of self or best possible contact:

Family history of diabetes:

On Mothers side:
On Fathers side:
Of Siblings:

5.6 **5.**

5.7 Risk factors

Smoking: Y / N / Ex / U *Comments:* (ex > 1yr stopped)	**5.**
Snuff user: Y / N / Ex / U *Comments:* (ex > 1yr stopped)	**5.**

5.11

Random Capillary Glucose (from file):	**5.**
HbA1c within last 4 months:	**5.**
Urine dipstick (from file): *Glucose:* *Ketones:* *Protein:* *Blood:*	**5.**
Blood pressure (mmHg):	**5.**

Weight (kg):
Height (cm):
Waist circumference (cm):
Hip circumference (cm):
Acanthosis Nigricans: Y / N / U Where:

DM year of diagnosis:
Clinical judgement on type of DM: 1 / 2 T1 = age of diagnosis before 30 (and on insulin within 1 yr of dx) or on insulin within 1 yr of diagnosis regardless of age of diagnosis
Hypertension on treatment: Y / N / U Year of diagnosis:
HT year of diagnosis:

History:

1. Presentation at time of diagnosis DKA or severe hyperglycaemia requiring hospitalization Y / N / U Coincidental finding, e.g. at surgery or other illness Y / N / U Symptoms such as polyuria/polydipsia causing patient to visit a health care facility Y / N / U Weight loss at time of diagnosis Y / N / U
2. Insulin started at time of diagnosis? Y / N / U
3. If on insulin now, was it started within 1 year of diagnosis? Y / N / U
4.Complications Macro: MI Stroke Revascularization Amputation Micro: Lasered Nephropathy Neuropathy)

Medication (from list)

Insulin
Oral medication

Others

Other diseases

1.
2.
3.
4.
5.

Appendix C

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Dr C Padoa
School of Pathology
Department of Chemical Pathology
Medical School
University

E-mail: Carolyn.Padoa@wits.ac.za

CC: Supervisor: Not applicable <>
and <HREC-Medical ResearchOffice@wits.ac.za>

FROM: Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2020/02/27

REF: *R14/49*

PROTOCOL NO: **M200174** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *The phenotypic and genotypic characterization of patients diagnosed with diabetes mellitus in South Africa*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to the Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

Appendix D

Table A.1: Primer pairs for HLA gene region amplification including the amplicon sizes and the relative positions of the primers (60)

Forward primer	Sequence **(5' to 3')**	**Reverse primer**	**Sequence (5' to 3')**	**Size (bp)**
Name	**Sequence**	**Name**	**Sequence**	
HLA-A_F	ATCCTGGATACTCACGACGCGGAC	HLA-A_R	CATCAACCTCTCATGGCAAGAATTT	5398
HLA-B_F	AGGTGAATGGCTCTGAAAATTTGTCTC	HLA-B_R	AGAGTTTAATTGTAATGCTGTTTTGACACA	5296
HLA-C_F	CAGCACGAAGATCACTGGAA	HLA-C_R	TGAGGAAAAGGAGCAGAGGA	5906
DRB1_F1	CTGCTGCTCCTTGAGGCATCCACA	DRB1_R1	CTTCTGGCTGTTCCAGTACTCGGCAT	6150
				6146-
DRB1_F2	CTGCTACTCCTTGAGGCATCCACA	DRB1_R2_1	CTTCTGGCTGTTCCAGGACTCGGCGA	11478
		DRB1_R2_2	CTGCTGCTCCCTGAGGCATCCACA	
		DRB1_R2_3	CTTCTGGCTGTTCCAGTACTCAGCGT	6200
		DRB1_R2_4	CTTCTGGCTGTTCCAGTACTCCTCAT	7154
		DRB1_R2_5	CTTCTGGCTGTTCCAGTGCTCCGCAG	8326
		DRB1_R2_6	CTTCTGGCTGTTCCAGTACTCGGCGC	9428
				11041
DRB1_F3	GCACGTTTCTTGTGGCAGCTTAAGTT	DRB1_R3	ATGCACGGGAGGCCATACGGT	5010
DRB1_F4	GCACGTTTCTTGTGGCAGCTAAAGTT	DRB1_R4	ATGCACAGGAGGCCATAGGGT	5541
DRB1_F5	TTTCCTGTGGCAGCCTAAGAGG	DRB1_R5	ATGCATGGGAGGCAGGAAGCA	5769

DRB1_F6	CACAGCACGTTTCTTGGAGTACTC	DRB1_R6	CAGATGCATGGGAGGCAGGAAGCG	6218
DRB1_F7	AGCACGTTTCTTGGAGCAGGTTAAACA	DRB1_R7	CACAGCACGTTTCCTGTGGCAGGG	6218
DRB1_F8_1	CACAGCACGTTTCCTGTGGCAGGG			6245
DRB1_ F8_2	CACAGCACGTTTCTTGAAGCAGGA	DRB1_R8	TGGAATGTCTAAAGCAAGCTATTTAACATATGT	6218
DRB1_ F8_3	ACAGCACGTTTCTTGGAGGAGGT			6314
DQA1_F	GCCAGGGAGGGAAATCAACT	DQA1_R	ATCCAGTGGAGGACACAGCAC	6488
DQB1_F1	AAGAAACAAACTGCCCCTTACACC	DQB1_R1	TAGTATTGCCCCTAGTCACTGTCAAG	9093
DQB1_F2	AAGAAACAAACTGCCCCTTATACC	DQB1_R2_1	TAGTACTGCCCCTAGTCACTGCCAAG	9000
		DQB1_R2_2	TAGTACTGTCCCTAGTCACTGCCAAG	9100
DPA1_F	CTCTCTTGACCACGCTGGTACCTA	DPA1_R	TTGGCCTCTTGGCTATACCTCTTTT	6709
DPB1_ F1	CCTCCTGACCCTGATGACAGTCCT	DPB1_ R1	CCATCTGCCCCTCAAGCACCTCAA	8940
DPB1_F2	CTCAGTGCTCGCCCCTCCCTAGTGAT	DPB1_R2	CTCAGTGCTCGCCCCTCCCTAGTGAT	7272

Appendix E

Table A.2: PCR primers used for the amplification of HLA regions

Region	Primer	1x reaction volume (µl)
HLA-A	HLA-A_F	0.5
	HLA-A_R	1
HLA-B	HLA-B_F	1
	HLA-B_R	1
HLA-C	HLA-C_F	0.5
	HLA-C-R	1
DRB1	F1 Mix 1	1.5
	R1 Mix 1	3
DRB1	F1 Mix 2	3.6
	R1 Mix 2	3.4
DQA1	DQA1_F	1.5
	DQA1_R	1.5
DQB1	DQB1_ F1	1
	DQB1_ R1	1
	DQB1_ F2	2
	DQB1_R2_1	1
	DQB1_ R_2_2	1
DPA1	DPA1_F	1.5
	DPA1_R	1.5
DPB1	DPB1_F1	1.5
	DPB1_R1	1.5
	DPB1_F2	1.5
	DPB1_R2	1.5

*DRB1 forward primer mix 1 (F1 Mix 1) was made up of DRB1_F1, DRB1-F2, etc-F4, DRB1 forward primer mix 2 (F1 Mix 2) was made up of DRB1_F5-8, DRB1 Reverse primer mix 1 (R1 Mix 1) was made up of DRB1 R_1-4, DRB1 Reverse primer mix 2 (F1 Mix 2) was made up of DRB1_R5-8. Equal molars of primers were added into the mixes.

Appendix F

Table A.3: Primer sequences for rs2040410 and rs7454108

HLA polymorphism	Primer	Primer sequence 5'-3'
rs2040410 (G>A)	Forward primer	TGTGCTGAGAGTTCCAGCCT
	Reverse primer	CACAAGGACTCATGGCTTG
rs7454108 (T>C)	Forward primer	TCTCTGCTCTCACTGCACAC
	Reverse primer	ACCCCTAACACAAACCTAGAATTCT

Appendix G

Table A.4: The number of variants investigated in each method

Ion Torrent AmpliSeq	AmpliSeq Nanopore	PCR-RFLP
rs2040410	rs2040410	rs2040410
rs7454108	rs7454108	rs7454108
rs2476601	rs2476601	
rs2857595	rs2857595	
rs1980493	rs1980493	
rs9272346	rs9272346	
rs2647044	rs2647044	
rs12722495	rs12722495	
rs7100025	rs7100025	
rs689	rs689	
rs72743477	rs72743477	
rs12708716	rs12708716	

*SNPs found in all methods

*SNPs found in AmpliSeq Nanopore and Ion Torrent AmpliSeq

Appendix H

DRB1 Mix 1 Forward	Primer Number	volume for 1 reaction	ratio of primers	50 reactions using 4uM stock (Shiina)	250 reactions using 4uM stock (Shiina)	250 reactions using 10uM stock	
PE2-F1	9	0.5	1	25	125	50	
PE2-F2	11	0.5	1	25	125	50	
PE2-F3	13	0.5	1	25	125	50	
water						225	
total		1.5		75	375	375	**1,5ul per 20ul reaction**

DRB1 Mix1 Reverse	Primer Number	volume for 1 reaction	ratio of primers	50 reactions using 4uM stock (Shiina)	250 reactions using 4uM stock (Shiina)	250 reactions using 10uM stock	
PE2-R1	10	0.5	1	25	125	50	
PE2-R2	12	0.5	1	25	125	50	
PE2-R3	14	0.5	1	25	125	50	
PE2-R4	15	0.5	1	25	125	50	
PE2-R5	16	0.5	1	25	125	50	
PE2-R6	17	0.5	1	25	125	50	
water						450	
total		3		150	750	750	**3ul per 20ul reaction**

DRB1 mix 2 forward	Primer Number	volume for 1 reaction	ratio of primers	50 reactions using 4uM stock (Shiina)	250 reactions using 4uM stock (Shiina)	250 reactions using 10uM stock	
1.1-F	18	0.12	1	6	30	12	
1.2-F	20	0.12	1	6	30	12	
4-F	26	0.48	2	24	120	48	
2-F	22	0.24	4	12	60	24	
3568-F	24	0.24	4	12	60	24	
7F4	28	0.24	4	12	60	24	
10-F	31	0.24	4	12	60	24	
9-F	30	1.92	8	96	480	192	
water				0	0	540	
total				180	900	900	**3,6ul per 20ul reaction**

DRB mix 2 reverse	Primer Number	volume for 1 reaction	ratio of primers	50 reactions using 4uM stock (Shiina)	250 reactions using 4uM stock (Shiina)	250 reactions using 10uM stock	
4R	23	0.567	1	28.35	141.75	56.7	
12-R	19	0.142	2	7.1	35.5	14.2	
3568R	21	0.142	2	7.1	35.5	14.2	
7R-2	25	0.142	2	7.1	35.5	14.2	
10-R	29	0.142	2	7.1	35.5	14.2	
9-R	27	2.268	4	113.4	567	226.8	
water		3.403				510.45	
total				170.15	850.75	850.75	**3,4ul per 20ul reaction**

Appendix I

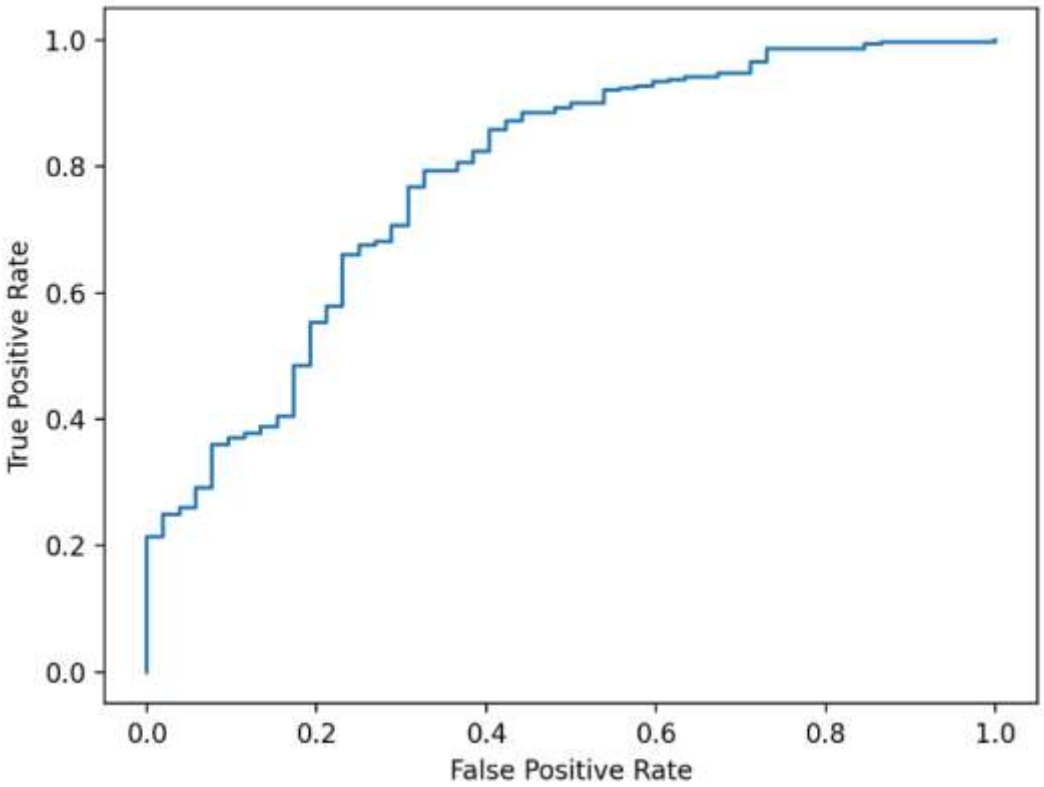


Figure 1: Area under the curve of 0.78 when training a 1D-CNN model on the TEDDY cohort children's genotypic data from 28 SNPs.

Final Master's dissertation NN Mathabela_highlighted(19 Sep-NN)-1.docx

ORIGINALITY REPORT

3%	3%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	hdl.handle.net Internet Source	1%
2	wiredspace.wits.ac.za Internet Source	1%
3	repository.up.ac.za Internet Source	1%
4	ElShorbagy, Khaled S.. "Metagenomic Analyses of Water Samples in Western New York Utilizing Oxford Nanopore Minion Sequencing Technology", State University of New York at Buffalo, 2021 Publication	1%