UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Using Oxford Nanopore technology to detect triplet expansion in Spinocerebellar Ataxia 7

DISSERTATION FOR A MASTERS DEGREE BY RESEARCH:

A Dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of

Master of Science in Medicine

by

Runesu Bakasa

Student Number: 2281063

Supervisor: Dr Stuart Ali

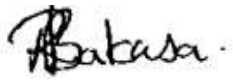
Co-supervisor: Prof Michèle Ramsay

Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Johannesburg, February 2022

Declaration

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



_____

(Signature of candidate)

On this the 15th day of February 2022 in Johannesburg

Acknowledgements

To my supervisors, Stuart and Michèle, thank you for your unwavering support and guidance. It has been a very obstacle-filled journey, but each hurdle has served as a character building opportunity. The chance to venture into the field of study has opened amazing opportunities, and I am fortunate to have been under the wing of the very best in the field. I look forward to more opportunities and growth, as we continue to work together.

To the SBIMB staff and students, thank you for holding my hand throughout the entire masters. David and his complete commitment to figuring things out with me, Jocelyn for being amazing at everything, Mahtaab for being an amazing friend and advisor. I felt very welcome in the walls of SBIMB and truly have become a better person and scientist.

To my family, I owe you all the thanks for your support through this entire Masters. My father, Runesu Bakasa Snr, being the reason I entered the science fraternity. Our conversations are always uplifting and you have helped me to remain strong. Your Winston Churchill quotes, and your ability to guide me through all situations, academic and social, have jointly made me a man of substance. My mother checking in daily, thank you for reminding me that no matter how old I get, I will forever be your "Chitopi" and will forever be protected by your prayers.

My siblings, for holding my hand through everything. You have been the driving force that kept me on track. Special shout-out to my brother Tseriwa, you have accommodated me in so many ways and are my first port of call when I feel the world crushing down over me. Your exemplary research achievements have shown how you are leading the way. My brother (in-law) Walter, you have given me extraordinary love and support. Stay blessed.

To my dearest Yevai, you are truly the best partner anyone could ever wish for. Your love and support has been a true reflection of what any person would pray for in a partner. At this stage, I feel your genetic crash course has given you enough fire power to present a journal club session of your own. I look forward to future achievements and greater heights that we can reach, on and off the pitch.

Finally, an acknowledgement to all the lives lost during this COVID-19 pandemic. Gone, but never forgotten. Rest in peace.

Financial Contributions which supported this work:

The National Research Foundation of South Africa, for the grant-holder linked bursary.

The Wits Health Consortium and the SBIMB for providing additional funding during my project extension to enable the completion of my degree.

Abstract

Spinocerebellar ataxia type 7 (SCA7), an autosomal dominant disease, is caused by the expansion of the CAG repeat sequence in the *ATXN7* gene. The current diagnostic test for SCA7 is by capillary electrophoresis which involves size characterization of the CAG repeat expansion. This provides a good estimate of the number of CAG repeats but gives no information on the sequence variation within the expansion. The aim of the project was to assess the effectiveness of using Oxford Nanopore long-range single molecule sequencing to characterize the length and sequence of the triplet repeat tract in the *ATXN7* gene. The first step was primer design and PCR optimization to develop a robust amplification protocol to ensure that African DNA samples would amplify, by minimising PCR failure due to variations in primer binding sites. The PCR optimization was designed for long stretches of G-C rich regions to accommodate the maximum size of a SCA7 triplet expansion. Ten control DNA samples and two patient samples were used in this project and following PCR, they were sequenced using the Oxford Nanopore MinION device. Data analysis protocols recommended by the manufacturer could not be used because the size of the reads was too short for the standard algorithm (<1kb). Alternative data analysis methods were used, which analysed the triplet number from the read size, but the analysis of the sequence remained sub-optimal. A cost analysis was attempted to compare the costs of conducting a size analysis by capillary analysis against using nanopore sequencing, but had many limitations. Future recommendations are to design primers that capture a longer fragment of the *ATXN7* gene to produce longer reads, as used in other successful, recently published analyses of nanopore data.

Nomenclature, List of Abbreviations and Symbols

BAM – Binary Alignment Map

BP – Base Pairs

(CAG)n – Number of consecutive CAG repeats in a given sequence.

CE – Capillary Electrophoresis

Chr – Chromosome

DMSO – Dimethyl Sulfoxide

DNA – deoxyribose nucleic acid

GRCh38 – Genome Reference Consortium Human Reference 38

LRS – Long Read Sequencing

NCBI – National Centre for Biotechnology Information

NGS – Next Generation Sequencing

NHLS – National Health Laboratory Services

ONT – Oxford Nanopore Technology

PCR – Polymerase Chain Reaction

Poly – Q – Polyglutamine

SAM – Sequence Alignment Map

SBIMB – Sydney Brenner Institute for molecular Bioscience

SCA/s – Spinocerebellar Ataxia/s

SMRT – Single-Molecule Real-Time

SNPs – Single Nucleotide Polymorphisms

SRS – Short Read Sequencing

TBE – Tris-Borate-EDTA

TE – Tris-EDTA

TRDs – Triplet Repeat Disorders

UCT – University of Cape Town

VCF – variant call file

ZMW - Zero-Mode Waveguide

Gene list

ATN1 – atrophin 1 gene / dentratorubral pallidoluysian ataxia 1 gene

ATXN7 –ataxin 7 gene / spinocerebellar ataxia 7 gene

DMPK – myotonic dystrophy protein kinase gene

FMR1 – fragile X mental retardation gene

HTT – huntingtin gene

JPH3 – junctophilin-3 gene

PP2R2B – protein phosphatase 2 regulatory subunit beta gene

Bioinformatics Toolsets:

ASIC - Application Specific Integrated Circuit

IGV – Integrative Genomics Viewer

Table of Contents

Table of Figures

Chapter 1

Table of tables

1. Chapter 1: Introduction

The evolution of molecular diagnostic approaches and methods have brought about improvements in accuracy, cost and convenience. The currently used molecular diagnostic assays have allowed for re-characterisation of conditions that had been previously diagnosed using biochemical methods (Zanetti *et al.*, 2019). Although molecular genetic diagnostic methods are designed to use the same basic principles when analysing DNA extracted from different populations, protocol adjustments may be required to develop protocols more suited to African populations. Africans have more genetic diversity in both coding and non-coding regions of the genome, so diagnostic methods developed for non-African populations are not always ideal for assessing samples from African individuals (Huang, Shu and Cai, 2015). For example, when using primers for PCR assays, it is important to check whether the primers overlap common polymorphisms in Africans and when doing mutation detection, it is important to first understand the mutation profile in African populations for the disease under consideration (Krause, Seymour and Ramsay, 2018). This project attempted to develop a molecular diagnostic protocol that uses Oxford Nanopore Technology (ONT) to characterise the SCA7 region in African individuals through long read sequencing.

Accurate molecular diagnosis is essential for clinicians to treat and manage their patients. There is a scarcity of readily availability data from African patients with monogenic traits which hampers the ability of clinicians when trying to make a diagnosis and requires further research (Pereira *et al.*, 2021).

The human genome is $3x10^9$ base pairs in length and includes many segments that contain repetitive DNA sequences located in pericentromeric, centromeric, or acrocentric regions (Jain *et al.*, 2018). A review on the biological implications of triplet repeat expansions showed that more than 40 neurological disorders are caused by triplet repeat expansions, prompting further understanding of the genetic structure of triplet repeat expansions (Adams and Wasserstein, 2021). Reviews of the nature of the various triplet expansion disorders have shown that when the threshold number of repeat sequences is exceeded within a gene, these variations cause monogenic disorders, which are recognized by their inheritance patterns in families (reviewed in Babar, 2017).

There are several types of triplet repeats found in the human genome, some of which cause monogenic disorders. A variety diagnostic methods are used to identify the different triplet repeat expansion disorders, and each comes with its own set of advantages and limitations. This project is looking at the triplet repeat expansion disorder spinocerebellar ataxia 7 (SCA7). By outlining the genotypic nature of the disorder, the location of the triplet expansion in the gene and the current diagnostic methods used to identify SCA7, I will explore the use of an alternative diagnostic method to counter some of the limitations of the currently used methods.

1.1. Triplet Expansions

Trinucleotide repeat sequences occur in coding or non-coding regions of particular genes. These repeats have no effect on the function of the gene as long as they do not exceed a particular threshold length (Almeida *et al.*, 2013). Expansion of the repeats above such a threshold is associated with a variety of neurological disorders. Table 1.1 shows the most prevalent triplet repeat disorders, the corresponding chromosomes where these expansion occur, locations, repeat types and the genes that the expansion sequences are found in. Expansion of GCG, CTG, CAG, GAA, and CGG repeats lead to a positive diagnosis for a range of diverse human monogenic diseases, depending on the gene the repeats are located in (Mitsuhashi *et al.*, 2018).

Triplet expansion disorders can be categorised based on clinical, pathological or triplet repeat sequence similarities. The prevalence of triplet expansion disorders also varies among populations and racial demographics, like spinocerebellar ataxia type 2 (SCA2) which has the highest prevalence among Indian families (Huang, Shu and Cai, 2015). This project will be focusing on spinocerebellar ataxia type 7 (SCA7).

Table 1.1 Selected common triplet expansion disorders

Disease Name	Location	Gene	Repeat Type	Prevalence	Unaffected	Reduced penetrance	Full penetrance
Friedrich Ataxia	Chr9	*FXN*	GAA	2 in 100 00 to 4 in 100 000	5-30 repeats	31-69 repeats	70-170 repeats
Fragile X Syndrome	ChrX	*FMR1*	CGG	1 in 4000 (males) 1 in 8000 (females)	6-40 repeats	61-200 repeats	<200-230 repeats
Huntington Disease	Chr4	*HTT*	CAG	2.7 in 100 000	4-35 repeats	36-39 repeats	40-121 repeats
Myotic Dystrophy	Chr19	*DMPK*	CTG	1 in 20 000	5-34 repeats	50-100 repeats	100- 1000 repeats

*(Chintalaphani et al., 2021)(Squitieri and Jankovic, 2012) (Adams and Wasserstein, 2021)(Young Gi Min, 2021) (Ramakrishnan et al., 2021. NCBI, accessed 3 January 2022)

Triplet repeat expansion disorders can also be categorized into non-polyglutamine disorders and polyglutamine (polyQ) disorders. The non-polyglutamine disorders are those that do not have CAG repeat sequences (Ramakrishnan et al., 2021. NCBI, accessed 3 January 2022). Common clinical and pathological attributes can be observed within disorders of a similar repeat sequence expansion. The CGG repeat expansion disorders, like Fragile X syndrome, involve methylation of the DNA sequence containing the expanded CGG repeats. This leads to the silencing of the *FMR1* gene (Sandberg and Schalling, 1997) (Adams and Wasserstein, 2021) .

As mentioned above, a commonly occurring group of triplet expansion disorders is the CAG repeat disorders. There are two types of CAG repeat expansion disorders. The less common CAG repeat expansions occur in untranslated regions of the associated gene, such as in spinocerebellar ataxia 12 (SCA12). The CAG repeat expansion associated with SCA12 is found 133 nucleotides upstream of the translation starting point of the *PPP2R2B* gene (Nalavade *et al.*, 2013). The more common CAG expansions have repeats occurring within the translated regions of target genes. These expansions are called polyglutamine (polyQ) disorders, such as Huntington disease, Huntington-like disease and spinocerebellar ataxia 7 (Nalavade *et al.*, 2013) (Chintalaphani *et al.*, 2021)

1.1.1. Polyglutamine (polyQ) disorders

Polyglutamine disorders are called polyQ disorders, because the CAG triplet codes for the amino acid glutamine, which is coded by the single letter Q (Sullivan *et al.*, 2019). PolyQ disorders are

characterised by a (CAG)n repeat expansion tract in a coding region and these disorders are characterized by proteins with enlarged sections containing stretches of polyQ, that potentially cause intracellular toxicity. CAG repeat expansions in coding regions, are associated with adult onset neurodegenerative diseases which are autosomal dominant, including the various spinocerebellar ataxias SCA1, SCA2, SCA3, SCA6 and SCA7, among other disorders (Martindale, 2017). PolyQ tracts are associated with the mis-folding of the three-dimensional structure of translated proteins. These misfolded proteins can form part of the translational machinery or be transcription factors, so their loss of function affects DNA transcription, RNA processing and cell survival pathways. These alterations cause activation of apoptosis (Nalavade *et al.*, 2013).

As mentioned above, one of the major groups of polyQ diseases are the spinocerebellar ataxias (SCAs) which were originally referred to as autosomal dominant cerebellar ataxias (ADCAs). SCAs are a heterogeneous group of neurodegenerative disorders. They are characterized by progressive degeneration of the brainstem, spinal cord and cerebellum. Each respective subtype is defined numerically in the name, by the order in which they were identified (Martindale, 2017). These SCAs demonstrate the phenomenon of anticipation, where the number of CAG repeats can increase in subsequent generations, due to the meiotic instability (Kraus-Perrotta and Lagalwar, 2016). The longer the repeat sequence tract, the earlier the onset of the disease and the more severe the symptoms (Lutz, 2007; Martindale, 2017). Anticipation is similar between sexes, whether inherited maternally or paternally (Gardiner *et al.*, 2019).

A study was conducted in 2011, where in a population of 50.5 million, the annual incidence of SCAs in South Africa was estimated to be 2.02/10 000 000 per year. Within the population of 50.5 million South Africans, 245 individuals presented with SCA symptoms and the individuals were from 165 families. Out of these 245 positive cases, the majority were confirmed to have SCA1 (48.8%) and SCA7 (26.6%). The study also showed that 50 of the 165 families that 245 positive cases came from were positive for SCA7. An ethnicity distribution analysis was conducted on the 50 families that were positive for SCA7 and 49 of these families were black South Africans (Smith *et al.*, 2012). This ethnicity distribution of SCA7 positive cases inspired the focus of this project to be solely on black South Africans when selecting the patient and control samples.

1.1.2. Spinocerebellar ataxia 7

Spinocerebellar ataxia type 7 (SCA7), an autosomal dominant disease, is caused by the expansion of the CAG repeat sequence in the *ATXN7* gene on chromosome 3. The *ATXN7* gene's cytogenetic location is 3p14.1 and the start coordinate on the GRCh38 reference sequence is position 63,898,361 (Mitsuhashi *et al.*, 2019). This project is focused on the number and sequence of CAG repeats in the *ATXN7* gene and the implications of expanded CAG repeats.

On the reference sequence (GRCh38) the *ATXN7* CAG repeat tract has 10 CAGs. This reference length is present in 70% of the human population in the *ATXN7* CAG repeat tract (Michalik, Martin and Broeckhoven, 2004). The full range of non-expanded alleles are recorded to be 4-35 CAGs (Mitsuhashi *et al.*, 2019), but this range covers the normal, mutable normal and reduced penetrance allele sizes. Affected individuals usually have more than 36 CAG repeats, although individuals with reduced penetrance of 34-35 CAG repeats may also present with symptoms (Salas-Vargas *et al.*, 2015). Table 1.2 below shows reference ranges for SCA7 and describes the implications of the number of CAG repeats in the *ATXN7* gene.

Table 1.2: CAG repeat reference ranges for SCA7 (Michalik, Martin and Broeckhoven, 2004) accessed from https://www.ncbi.nlm.nih.gov/books/ on 12 January 2022)

Gene	*ATXN7/SCA7*
Normal alleles	4-27 CAG repeats
Mutable normal alleles	28-33 CAG repeats
Reduced penetrance alleles	34-35 CAG repeats
Classic adult onset alleles	36-58 CAG repeats
Juvenile onset (below 10years of age) alleles	>70+ CAG repeats
Infantile onset alleles	130 - 200+ CAG repeats

SCA7 symptoms comprise of progressive cerebellar ataxia, and retinal and rod-cone dystrophy with progressive loss of vision that eventually results in complete blindness. Other symptoms include dysarthria and dysphagia. Early onset in childhood is associated with regression of motor milestones such as crawling, standing and walking (Smith *et al.*, 2012). SCA7 has very unstable CAG repeats and this results in a gradual increase in the triplet expansion through successive

generations. The increase in the number of repeats is characterized by an earlier onset of the disease and worsening of the symptoms, such that affected children may die before the affected parent or grandparent starts showing any symptoms (Smith, 2014) (Chintalaphani *et al.*, 2021).

1.2. Current Method used for SCA7 Diagnosis

At present, molecular diagnosis of SCA7 is performed at the National Health Laboratory Service (NHLS), Groote Schuur Hospital in Cape Town. Dr Esterhuizen (NHLS, Groote Schuur Hospital Cape Town) provided information on the diagnostic detection method, which is described in the following sections. The diagnosis process begins with DNA extraction from the patient blood sample. This project will not be discussing the DNA extraction steps, as it will be using pre-extracted DNA samples. The methodology of the current diagnosis of SCA7 involves the following steps:

1. PCR amplification of the SCA triplet repeat tract (singleplex and/or multiplex) with fluorescent probes.
2. Agarose gel electrophoresis to observe the PCR amplicons and identify triplet expansions.
3. Capillary Electrophoresis (CE), to accurately size the PCR fragments and determine the number of CAG repeats in each PCR fragment.
4. Confirmation test of any inconclusive results from the CE analysis.

1.2.1. PCR amplification of SCA Genes (Multiplex and Singleplex)

Since SCA symptoms have many similarities in how they present themselves, the physician often starts by requesting the patient's genetic history to determine the inheritance pattern. If the patient's family genetic history is known and has documented evidence of the SCA7 genetic mutation, then the physician will specifically request a singleplex polymerase chain reaction (PCR) for SCA7. If the patient presents with SCA symptoms, but their family's genetic mutation is unknown, then the physician will request a multiplex PCR reaction with loci for SCA1, 2, 3, 6 and 7. All the above mentioned disorders are characterised by CAG repeats (Dorschner, Barden and Stephens, 2002).

1.2.2. Capillary Electrophoresis (CE)

Amplicons from either the singleplex or multiplex PCR amplifications can be sized using CE. In this project, focus is put on the protocol of CE size analysis of the singleplex PCR amplicons of the SCA7 triplet repeat fragment. The CE protocol assessed in the Cape Town assay makes use of

the ABI 3100 Genetic Analyser. Although there are several other genetic analysers with fluorescence detectors available on the market, specific focus is placed on the ABI 3100 because it was used in the analysis of the patient samples that were analysed in my project, using an alternative method. The cost assessment, accuracy and time component assessed in this project makes use of the ABI 3100 Genetic Analyser method.

The sizing of PCR fragments using CE on the ABI 3100 Genetic Analyser involves very minimal sample preparation which takes less 30-45 mins. The fragment size analysis involves running GeneScan 500 Rox standards through a capillary. The prepared samples of DNA amplicons are also run through the same capillary and the sizes of the DNA fragments are determined by their migration through the capillary, relative to the migration of the Rox 500 standards. The analysis of the data produced is done using the GeneMapper software version 3.0 (Smith, Esterhuizen and Greenberg, 2013).

The interpretation of the results has three possible outcomes: a negative SCA7 result, a positive SCA7 result or an inconclusive result. Figure 1.2 below, shows a representation of an example of each of these results on an electropherogram. The top image (a) shows two alleles within the unaffected range of below 290bp. Since the amplicons are 262 + (CAG)n, and for both alleles the (CAG)n is less than 10 repeats, this represents a negative test result for SCA7. The second image (b) shows a peak of an allele within the unaffected range and expanded allele in the range of 350-370 bp. This is a positive result for SCA7 showing multiple peaks for the expanded allele as a result of mitotic instability of expanded alleles. The bottom image (c) shows one peak representing one allele within the unaffected range. This can be interpreted as homozygosity for an allele in the unaffected range, or a potential second outcome would be premised on the fact that there may be allelic dropout, where the expanded allele may have been too large to give an amplification product and therefore the outcome is inconclusive. Additional tests would need to be performed on the parents in order to confirm that they have the same sized alleles and homozygosity is a plausible outcome or alternative tests could be done to give conclusive results (Smith, Esterhuizen and Greenberg, 2013).

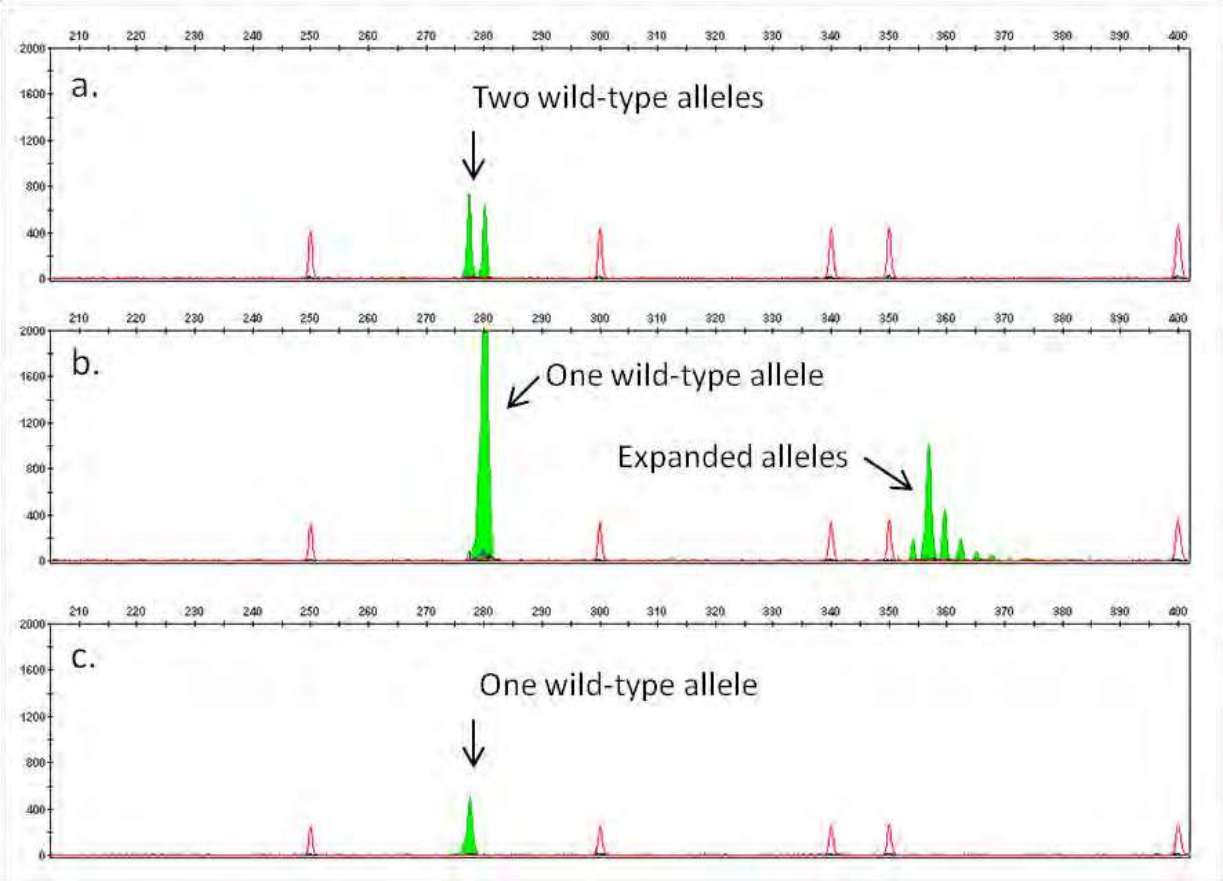


Figure 1.1: Example of electrogram results showing a negative result (a), a positive result (b) and an inconclusive result (c) (Smith, 2014). The x-axis represents the fragment size and the y-axis represents the height of the peaks (quantity of product).

1.2.3. Triplet Primed PCR

In the case of suspected homozygosity, alternative tests to give conclusive results are Southern blot analysis and triplet repeat primed PCR (TP PCR) (Smith, 2014). Southern blotting is labour and time intensive. The NHLS at Groote Schuur Hospital, UCT, does not have the facilities for Southern blotting as a confirmatory test for homozygosity, so TP-PCR is used as a confirmatory test for diagnoses that are deemed inconclusive due to suspected homozygosity (Smith, 2014).

The TP PCR test makes use of three primers. Primer 1 is the forward primer, and it has a fluorescent probe attached to it. Primer 2 is a reverse primer and is complimentary to the repeat sequence (CAG), so it binds onto any position on the repeat region. Primer 3 binds specifically onto the 5'-region or "tail" of Primer 2. It is important to ensure that Primer 3 does not bind anywhere else on the genome, so the design of Primer 3 should be very specific to the tail of Primer 2. In the PCR

reaction, Primer 2 is present at a 1:10 ratio to the forward and tail primers (Warner *et al.*, 1996). Since Primer 2 binds to the CAG repeat, it will have multiple binding sites within the triplet expansion region. This will result in amplicons of different sizes. (Smith, 2014).

The amplicons are analysed by CE and the results show a distinctive ladder-like electropherogram, and can be sized against the Rox internal standards. Unaffected range CAG repeat alleles will have the highest peaks. The presence of any expansion peaks will be indicated by the presence of larger peaks that diminish in height.

Figure 1.2 is a schematic representation that compares conventional PCR to TP-PCR.

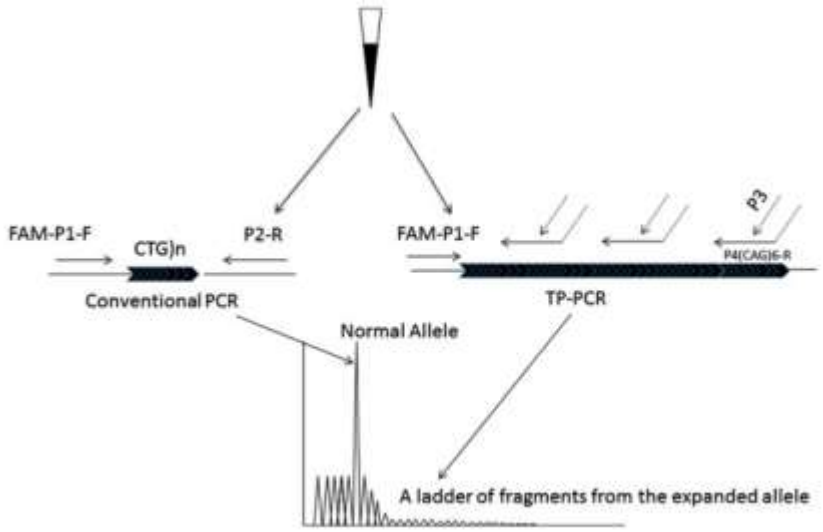


Figure 1.2. A schematic representation of the comparison between conventional PCR and TP-PCR and the CE eletrograms produced by each of the different PCR methods (Singh et al, 2014. Conventional PCR (left side of the schematic representation) shows the presence of two primers, a forward primer and a reverse primer. TP-PCR is shown on the right side of the schematic representation. The TP-PCR schematic shows the presence of 3 primers, a fluorescence labelled forward primer (FAM-P1-F), a reverse primer (P4-CAG-6-R) which binds onto the triplet repeat region and a tail primer (P3) that binds onto the tail of the reverse primer (P4-CAG-6-R). An electrogram comparison of the CE analysis of the conventional PCR and TP-PCR is also presented. The CE electrogram for TP-PCR shows multiple peaks of different sizes captured by the reverse primer and the tail primer as they amplify different portions of expanded triplet repeat region.

TP PCR protocols have also been developed to analyse the length of CAG repeats in other diseases such as Huntington Disease, Friedreich ataxia, SCA7 and SCA2 (Singh *et al,*. 2014). The effectiveness of this method was confirmed when samples previously tested using the Southern blot technique were also tested using TP PCR. This worked as proof of concept.

1.2.4. Limitations of Current Diagnostic Methods

The inconclusive aspect of the initial CE method was based on the possibility of the expansion being too large to be observed on the CE electropherogram, but with multiple fragment sizes, a

series of different sized fragments would be observed if indeed there is expansion. The accurate size of the CAG repeat alleles cannot be determined using TP PCR. The size range prediction is done by analysing the unknown samples in parallel with controls of a known genotype (Smith, 2014). The currently used analysis methods are dependent on the presence of standards and/or controls of a known genotype.

The current diagnostic test and the alternative confirmatory tests are effective, quick and easy to interpret, but the limitation is that they are only a determinant of size of the repeat region and not of the actual sequence of the triplet repeat tract. The analysis methods do not account for any variations within the triplet repeat sequences and are under the assumption that the triplet repeat region has no interruptions in the tandem repeat sequence. Different triplet repeat disorder studies have shown that interruptions within the tandem repeat sequence have various effects on the disease diagnosis, onset, prognosis, phenotype, progression and heritability (Musova *et al.*, 2009).

A variety of triplet expansion disorders with sequence interruptions present with different clinical implications. Fragile X Mental Retardation 1 *(FMR1*) is characterized by a CGG repeats, but when there are AGG interruptions, premature ovarian failure is prevented in carrier women. In the case of SCA10, although not a triplet repeat, the correlation between the number of ATTCT repeats and age of onset, does not apply in the presence of sequence variations within the repeat sequences (Musova *et al.*, 2009).

Narrowing down the focus to CAG repeat disorders, interruptions in the sequence of trinucleotide tracts have been shown to play a role in adding genetic stability to the triplet repeat regions in SCA1 and SCA2 (Gardiner *et al.*, 2019). These interruptions inhibit strand slippage, so the triplet repeat regions with sequence interruptions are less likely to expand upon transmission. Alleles without trinucleotide sequence interruptions are more predisposed to expansion and eventually to disease status (Musova *et al.*, 2009).

In SCA1, CAT interruptions in the CAG tandem repeat sequences serve as anchors that stabilize the repeat regions against slippage and expansion. Wild type alleles with repeats below the threshold have interrupted sequences of (CAG)n CATCAGCAT (CAG)n. Alleles with expansions above the threshold have pure CAG sequences in them . SCA2 has a known CAA interruption in the CAG repeat sequences, and yet most studies are still measuring the length of the repeat sequence. In SCA2, diseased alleles have pure CAG tracts while SCA2 alleles with in the normal

range have interrupted repeat configurations (Choudhry, 2001). This would suggest the importance of improving the current diagnostic methods of SCA7 by analysing the actual sequence of the triplet repeat sequences instead of simply measuring the size of the repeat sequence tract.

The possibility of interruptions within the triplet repeat sequence also has an effect on the current diagnostic methods used. In the case of suspected homozygosity, TP-PCR is used to confirm the inconclusive results. In the diagnosis of myotonic dystrophy, TP-PCR is used and interruptions in the CTG repeat sequence of the *DMPK* gene prevent the repeat primer from binding and this leads to false negative results (Musova *et al.*, 2009). This possibility reduces the faith in TP-PCR as a good alternative to inconclusive diagnoses of SCA7. It therefore becomes worthwhile to explore the possibility of using sequencing as the diagnostic method for SCA7.

1.3. Sequencing of triplet repeat expansions

DNA sequencing is a method for determining the exact order of nucleotides present in a section of DNA (Weinstock, 2013). Sequencing methods have evolved at an exponential rate since the start of the human genome project (Morey *et al.*, 2015). In this project, we shall be focusing on each generation of sequencing technology. Particular focus will be on how each of the sequencing methods fared in sequencing triplet repeat expansions. With each of the methods being analysed in this review, the suitability of each method for use in the diagnosis of SCA7 is assessed, and the cost and time implications that are attached to using each method to identify the triplet expansion in SCA7.

1.3.1. Classical (First Generation) Sequencing

The first sequencing methods that were developed are referred to as “first generation sequencing”, or “classical sequencing”, and include Maxam and Gilbert Sequencing, and the more popular and widely used, Sanger Sequencing (reviewed in Morey *et al.*, 2013). The Maxam and Gilbert sequencing method is complicated to undertake and makes use of toxic reagents. This made Sanger sequencing the more popular option. Sanger sequencing involves an enzymatic reaction which utilizes a DNA polymerase. Improvement of Sanger sequencing saw its development from the plus and minus Sanger sequencing, to a more efficient chain termination Sanger sequencing (Morey *et al.*, 2013).

In the selection of the best sequencing method for SCA7, several considerations were made on each sequencing method’s ability to be a sole diagnostic method of SCA7. Sanger sequencing is

good for sequencing short tandem repeats, with an accuracy of 99.99% for sequences of up to ~900bp. Long fragments cannot be analysed accurately as there is loss of resolution in the sequencing output due to the method's low power of discrimination (Tipu and Shabbir, 2015). The primers used in this project have an expected PCR amplicon size of 262 (CAG) n. Positive SCA7 alleles have 36-460 CAG repeats, so full penetrance alleles can be expected to be 1380 bases long. This is beyond the threshold size for accuracy using Sanger sequencing.

Sanger sequencing struggles with GC rich sequences (Tipu and Shabbir, 2015). The SCA7 triplet repeat tract being analysed in this project, is a CAG repeat expansion and constitutes mainly of GC bonds. Sanger sequencing has other limitations of low throughput and high costs and this makes this method a less preferred tool for SCA7 analysis. The limitations of the classical sequencing method paved way for the development of the next generation of sequencing approaches. This second generation sequencing is also known as short read sequencing (SRS) (Mardis, 2011).

1.3.2. Short Read (Second Generation) Sequencing

The concept of short read sequencing (SRS) is massively parallel reading of clonally amplified and spatially separated amplicons. The sequences produced by SRS are shorter (150 bp on average) (Paszkiewicz and Studholme, 2010) than those produced by classical (Sanger) sequencing methods, but there is a massive amount of these short sequences, which ensures that all sequences are represented many times (Morey *et al.*, 2013). The process of short read sequencing involves several sequential steps, which are the preparation of the template, sequencing and imaging process, and the data analysis pipeline (Grada and Weinbrecht, 2013).

Although there are different SRS methods, and each comes with specified data analysis processes, a general data analysis pipeline involves the pre-processing of the data to remove adapters and low-quality reads (Mardis, 2011). The data from high quality reads are mapped to a reference genome or one can do *de novo* alignment of the sequence reads (Grada and Weinbrecht, 2013). There is a wide variety of bioinformatics tools available for analysis of such data as well as identification of SNPs or indels.

In relation to sequencing triplet expansion regions, there are short falls that cannot be met by short read sequencing technology. SRS is dependent on an amplification step, which comes with the potential of PCR biases. Triplet expansion disorders can have long stretches of repeat sequences,

like Fragile X Syndrome (55-230 triplet repeats), which fall well above the read length limit of the Life Technologies SOLiD and the Illumina Hi-Seq 2000 technologies.

With repeat sequences, the reads potentially map onto several positions during alignment, so the algorithms will not accurately map them, or not map them at all, leading to read fall out (Etter *et al.*, 2011). SRS is difficult to implement as a diagnostic alternative for CAG repeat disorders because it cannot accurately sequence through GC rich regions of the DNA sequence is not optimally aligned due to a loss of sequence phase coherence (Loomis *et al.*, 2013).

1.3.3. Long Read (Third Generation) Sequencing

The next generation of sequencing that was developed is third-generation sequencing, also known as long read sequencing (LRS). This sequencing advance was spearheaded by developments from two groups, Pacific BioSciences and Oxford Nanopore Technologies (ONT). In relation to triplet expansion diagnoses, the long reads can cover the entire region of the triplet expansions and be used for both *de novo* genome assemblies and for mapping reads to a reference sequence. Figure 1.3 shows an illustration of how long reads can span entire regions which is convenient for sequencing and diagnosis of triplet expansion disorders like SCA7.

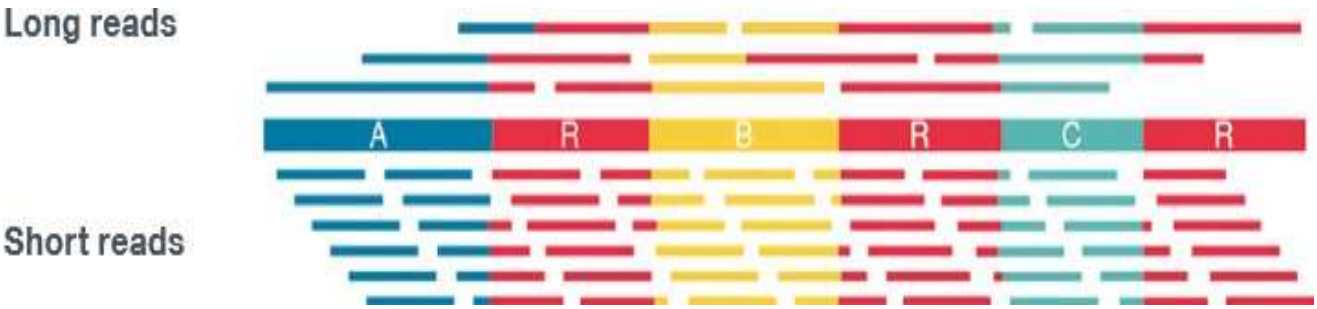


Figure 1.1: A representation of the alignment of long and short reads to a chromosomal region (each region represented by a letter). Genome assembly is simplified with long reads (top) and requires fewer sequences to cover the region, compared to short reads (bottom) (Ampe, 2019).

The first product released for LRS sequencing was the Pacific BioSciences (PacBio), which uses single-molecule real-time sequencing (SMRT) (Ampe, 2019). The SMRTbell is loaded onto a chip called the SMRTcell and diffuses to the bottom of the SMRTcell where there is the sequencing

unit called a zero-mode waveguide (ZMW). The ZMW has phospho-linked nucleotides labelled with four different fluorophores (Rhoads and Au, 2015). When the nucleotides are added to the ZMW chamber, DNA polymerase incorporates them and the phosphor chain is cleaved, releasing the attached fluorophore and emitting light (Rhoads and Au, 2015; Ampe, 2019). If the SMRTcell has many ZMWs, then many sequencing processes can be done in parallel. This platform requires minimal amounts of reagent and sample preparation and is very time-effective enabling results in minutes as opposed to days (Rhoads and Au, 2015).

In relation to sequencing of the triplet expansion disorder SCA7, LRS is ideal. For this project, however, PacBio was not appropriate because of the expensive set up costs of acquiring the PacBio Sequal II system, which costs 350 000 USD. This sequencing device is a cheaper alternative to the initial sequencing device, PacBio RS II, which costs 750 000 USD. An alternative method for LRS is ONT, and this is described in detail below, as it was my method of choice for this project.

1.4. Oxford Nanopore MinION

After analysis of the various sequencing methods, considerations of the advantages and limitations of each of the different methods were used in selection of the most ideal sequencing method to use for detecting the CAG repeat tract of the *ATXN7* gene. In this project, Oxford Nanopore MinION sequencing method is used to attempt to detect the size and sequence of the CAG triplet expansions in SCA7. Oxford Nanopore technologies has produced a range of sequencing devices, but the MinION was first to be released on the market, and as such, has been validated for different sequencing uses and has available data analysis pipelines on the GitHub repository.

Oxford Nanopore sequencing was used in the robust detection of tandem repeat expansions from long DNA reads. The analysis was done on artificial DNA plasmids of various copy numbers of the repeat unit. The analysis method was used to detect four different kinds of tandem repeats (CAG, CAA, GGGGCC, and CCTG) that are known to cause human diseases (Mitsuhashi *et al.*, 2019). This study proved the MinION sequencing protocol's ability to sequence tandem repeats in clonal DNA. Among the different tandem repeats analysed, the SCA7 triplet repeat sequence was included in the study. This served as a reference point for the choice of sequencing device and method used in this project.

Nanopore technology performs sequencing using a flow cell that consists of a sensor chip with an array of electrical channels. The sensor chip is covered with an electrically resistant membrane

with protein "nanopores" assembled above each channel. When an electrical potential is applied across the membrane, the sensor chip measures the electrical potential across each pore (Oxford Nanopore Technologies, 2017). Each DNA strand is ligated with an adaptor and a processive enzyme. The adaptor facilitates directing the DNA strand to the pore. As the DNA/enzyme complex approaches the nanopore, the enzyme unwinds the DNA, and a single strand of DNA is pulled through the nanopore. As the bases go through the pore, there are specific changes in electrical conductance across the pore (nanoporetech.com, 2017). The nanopore sequencing device can sequence nucleic acid bases in a molecule by measuring the interruption of current caused by each base passing through the pore on the membrane (Nanoporetech.com, 2017). Each interruption in the electric current is called an "event" and is recorded by a cloud based service called Metrichor which performs base calling using a Recurrent Neural Network (Cornelis *et al.*, 2017)

1.4.1. MinION Hardware

The MinION is a portable sequencing device measuring just 10 × 3 × 2 cm and weighing 90 g (Lu, Giordano and Ning, 2016). These dimensions allow it to be highly portable and convenient for both clinic, mobile, and point of care diagnostics. Apart from being the smallest sequencing device available, it is also relatively cheap to purchase ($1,000) and to use for clinical tests (Wilson, Eisenstein and Soh, 2019). Latest versions of the MinION have improvements on the protein pore and it uses a laboratory evolved E.coli CsgG mutant named R9.4 and sequencing speed is 450 bases/second (Jain *et al.*, 2018).

The MinION flow cell has 512 sequencing channels. This is ideal for sequencing many samples in parallel and simultaneously. Oxford nanopore technologies also produced an adaptor flow cell called the Flongle, which is used on the MinION. The Flongle has just 126 channels and is more cost-effective alternative, suitable for targeted regions, smaller genomes and diagnostics (Oxford Nanopore Technologies, 2017). The Flongle is also a reusable; washing the cell will allow the user to reuse it several times.

The only additional hardware the MinION requires is a computer with or without an internet source. The MinION can plug directly into a standard USB3 port on a computer. It has very minimal hardware requirements and simple configuration. The minimum requirements for the

compatible computer are Linux, OSX or Windows, with a solid-state drive, at least 8 Gigabytes of RAM and at least 128 Gigabytes of hard disk space (Lu, Giordano and Ning, 2016).

1.4.2. Library Preparation

DNA strands need to be converted to a library with suitable adaptors to enable sequencing and to guide the single DNA molecules to the pore. The library is prepared using fragmented native DNA or PCR products. DNA can be enzymatically fragmented using a transposase. This method is rapid (approximately 10 mins) and allows control of fragment length (Number *et al.*, 2019). For targeted sequence analysis or when sequencing limited amounts of starting DNA, there are several PCR-based library preparation protocols provided by Oxford Nanopore Technologies.

There are different methods for creating the DNA library and two common methods are 1D gDNA library preparation and 2D gDNA library preparation. The difference between the two library preparation methods is that a 1D library only contains the sense strand of the DNA, whereas the 2D library has both the sense and antisense strands that are joined together by a hairpin (Number *et al.*, 2019). In the earlier version of the MinION, 2D library preparation was most suitable for higher accuracy, but improvements that have been done to the protein pore of the sequencing (R9.4) have made 1D library preparation suitable for high accuracy (Jain *et al.*, 2018). The final step of the library preparation involves ligation of an "adapter" sequence to the ends of the DNA fragments and mixing with a processive enzyme that binds to the adapter.

1.4.3. Data Capture and Analysis

The MinION flow cell and the Flongle flow cell both have a sensor chip which is connected to an Application Specific Integrated Circuit (ASIC), where the signal processing and control happens. The ASIC measures the changes in electrical potential across the pores, and these signals are processed by a software package called MinKnow. This software also controls the MinION and can be used to change the parameters and workflows (Wei, Weiss and Williams, 2018).

The signals read by ASIC are recorded in the form of "squiggles". This is a presentation of the changing electrical signals as the DNA passes through the nanopore. MinKnow controls the flow cell, converting electrical output into a FAST5 file, which is a variation of HDF5 file format (Rang, Kloosterman and de Ridder, 2018). Various tools can be used to convert the FAST5 to FASTQ which is the human readable sequence that has been decoded from the FAST5 data – the most common tool being the Oxford Nanopore program called Guppy. FASTQ files can then be

processed by standard bioinformatics tools for interpretation of sequence data or can be uploaded through the internet for processing by the Oxford Nanopore Epi2Me service provided by Metrichor. Epi2Me will allow real-time sequence alignment/identification.

The Epi2Me service can be downloaded onto your computer and be used as a cloud-based data analysis platform, without having to download the different software required for analysis, instead just access them on this cloud-based platform. Epi2Me notebooks are available on the Epi2Me labs Index on the GitHub repository, where python codes for different pipelines are available. The notebook also has basic introductory tutorials, basic quality control pipelines, metagenomics pipeline, variant calling pipelines and many other analysis pathways specific to different users.

There are tools to improve base calling which work on both FAST5 and FASTQ. An example of such tools is Flip Flop base calling, which is used together with GUPPY for sequence extraction and base calling. Only pass reads (q score > 7) will be used for sequence alignment and genotyping analysis. The filtered FASTQ files obtained from NanoFilt will be mapped to the available reference sequences using minimap2 or NGMLR. Nanopolish is the error correction algorithm used to call variants from mapped BAM files to VCF file (Liau et al., 2019).

1.5. Rationale

The current diagnostic method for detecting the SCA7 triplet expansion length is PCR followed by capillary electrophoresis (CE) to separate fragments of different sizes that can be used to calculate the number of triplet repeats per allele. The CE method gives a good sizing estimate when only considering the number of repeats in the region, but it will not provide information on variations within the triplet repeat sequences themselves. Different triplet repeat disorder studies have shown that interruptions within the tandem repeat sequence have various effects on the diagnosis, onset, prognosis, phenotype, progression and heritability of the triplet expansion disorder (Musova *et al.*, 2009).

The current analysis method for SCA7, CE, has been able to give an adequate size estimate of the expanded alleles. This method, however has shown that the spectrum for size analysis has a limit and can give a false negative result if the expansion is too long. In this case, the result will show as apparent homozygosity. This brings about the introduction of a confirmatory test which is the TP-PCR. The TP-PCR comes accompanied by its own limitation, where the analysis would

produce a false negative result if there are variations in the triplet repeat sequence and the primers do not align to the target sequence as expected.

The search for an ideal diagnostic method that can characterise the size and sequence of the triplet repeat tract in the *ATXN7* gene, led to the conclusion that sequencing would be the most ideal method. After analysis of the different sequencing technologies that have been used to detect triplet sequences, the preferred sequencing approach with the most accuracy and high throughput is the LRS approach. Sanger and SRS methods cannot be used as the only diagnostic method as the sequencing accuracy is hampered by the difficulty of sequencing the regions of GC rich repeat sequences, challenges in aligning of short read sequences, potential PCR stuttering and somatic instability of the triplet repeat units. These shortcoming can be addressed by long read single molecule sequencing.

The sequencing devices available on the market for LRS are the PacBio and ONT. The LRS method selected for this project was ONT. The latest PacBio sequencing machines on the market is known as the Sequal II 5000 and the Sequal II 6000. Considering the costs of such sequencing machines and the laboratory space required to house the machine, considerations for the more appropriate LRS method to use in this project led to the selection of the Oxford Nanopore MinION.

The Oxford Nanopore MinION is a long-read sequencer, which costs $1000 (USD). A report by Jain *et al.* in 2018 showed the assembly of a human genome with Nanopore sequencing data. The data was collected from sequencing using MinION R9.4 1D chemistry and accuracy was above 99.8%. More recent studies have made use of the Oxford Nanopore MinION to sequence artificially cloned DNA with various sizes of the CAG triplet repeat (Mitsuhashi *et al.*, 2019). The protocols are results from these recent studies, and suggest that the MinION is ideal for the analysis performed in this project. The cost implications and accuracy justifies proposing the Oxford Nanopore MinION as a sequencing instrument to characterise the size of the triplet repeat region in the diagnosis of SCA7.

1.6. Justification

This project aims to explore whether the Oxford Nanopore MinION can accurately determine the size and DNA sequence of triplet expansions in triplet expansion disorders. The choice of disorder was based on discussions with Prof Krause, taking into consideration other projects in the Division of Human Genetics and the availability of samples. SCA7 was chosen, not based on prevalence,

but based on sample availability and the fact that SCA7 tends to be more common among black patients who present with suggestive symptoms to the clinicians at the Division of Human Genetics. The project was conducted in the midst of the COVID-19 pandemic and sample availability was more limited than originally anticipated.

This project aimed to assess whether the MinION is faster than the currently used technology, which also involves a PCR step. The project will assess the Oxford MinION's ability to accurately size and sequence the triplet repeat sequence tract in the SCA7 gene and offer a diagnostic alternative to the currently used method, CE. The sequencing approach further allows for more information about each triplet expansion because it would provide data on the exact sequence of the triplet repeat tract. The outcome of the project could provide information into understanding the effects of sequence variation, if any, on the heritability, prognosis and other clinical implications of SCA7.

1.7. Aim

To assess the effectiveness of using Oxford Nanopore long read single-molecule sequencing technology to characterize the triplet repeat length and DNA sequence in the *ATXN7* gene in a group of controls and SCA7 affected individuals with known triplet expansions.

1.8. Objectives

1. To establish methods for Oxford Nanopore Sequencing to assess the number and sequence of *ATXN7* trinucleotides per allele in unaffected individuals
2. To assess whether the Oxford Nanopore MinION platform can detect and characterize the triplet expansions in SCA7 affected individuals
3. To assess the cost/benefit implications of using the Oxford Nanopore MinION for the detection and characterization of the triplet expansion in SCA7

2. Methods

2.1. Study participants

The study was conducted using 10 control samples and 2 patient samples. The 10 control samples were anonymized African individuals with no reported history of SCA, provided by the SBIMB Biobank, with informed consent for broad use. The 2 patient samples were provided by the NHLS Human Genetics laboratory in Johannesburg. The study was given ethics clearance by the Human Research Ethics Committee (Medical) with the study number M190938, before commencing any research.

2.1.1 DNA samples (controls and patients)

Four commercial reference samples of human genomic DNA were used as laboratory controls for optimization. The laboratory controls included a Promega control DNA, and 3 samples from a family that were obtained from the Coriell biobank, referred to as the Coriell trio. The sequences of the laboratory reference samples are provided as open-source data on several online genomic resources. The Coriell samples are referred to as Sample 1, 2 and 8 as detailed below.

1- NA12891 – Sample 1 (Father)

2- NA12892 – Sample 2 (Mother)

8-NA12878 – Sample 8 (Daughter)

P- Promega Control DNA

The details of the 2 patient samples T1 and T2 from individuals with triplet expansions in the SCA7 alleles are shown in Table 2.1, and includes the number of CAG repeats as estimated using capillary electrophoresis. Unfortunately, no other information on the patients was provided, such as age at onset of symptoms, sex, or disease progression.

Table 2.1: Details provided from the diagnostic laboratory on the two SCA7 positive patient samples

Sample Name	Given ID	CAG Repeats
MISC873	T1	7/58
SCA464	T2	17/37

2.2. Design of SCA7 PCR Primers

Initially, primer sets were obtained from previous studies on the SCA7 triplet expansion analysis (Dorschner, Barden and Stephens, 2002). Evidence suggested that these primer sets were not optimal, because occasionally PCR allelic dropouts were observed (Fiona Baine-Savanhu, personal communication). This could indicate heterogeneity in the primer binding sites of the primers used.

It is understood that primer design is often based on the human reference genome versions HG19 or HG38 which was from a group of unknown individuals of predominantly Caucasian ancestry. Although there was no information about the ancestral background of individuals whose samples showed PCR allelic dropouts, it was reasoned given the genomic diversity of individuals of black African ancestry, the PCR dropouts could be associated with black South African patients.

To test this hypothesis, a bioinformatic analysis of the primer binding regions in sequence data from individuals of black African ancestry was conducted using the VCF files of the samples' genome sequence data and analysing variances in the targeted region using VCFtools. These sequence data were obtained from the 1000 Genomes Project.

Sequence variations were seen in the primer binding sites in the black African genome data sets. These variations would likely result in failure to amplify the SCA7 triplet expansion as the PCR primers would not bind to that region. This finding could explain why PCR dropout had been seen in the diagnostic laboratory, and meant that it was necessary to design a new set of primers for this project.

The general criteria for primer design is that they should be 20-24 nucleotides in length, have a melting temperature of between 55-65ºC and a GC content of 45-55%. Other features are the absence of primer dimers, no secondary priming sites and absence of hairpins (>3bp). Using these criteria, optimal primers were designed as follows. The Ensembl genome browser was used to obtain the FASTA sequence of the *ATXN7* gene. The sequence of the target region (Exon 3, Codon 30) where the CAG triplet repeats begin was identified, and the search for suitable primer binding sites was constrained to 50 bases before and 50 bases after the position of the triplet repeats. The decision was made to select primers that captured a region close to the targeted repeat sequence even though this meant that the target sequence length would be small. This was done to avoid confusion with other CAG tandem repeats that are found in non-coding regions of the gene within the target chromosome. Forward and reverse primers were generated using the software Primer3, which would amplify a 280-nucleotide segment of *ATXN7* reference sequence that includes the repeat sequence region.

The sequence of the forward and reverse primers generated, is shown below:

ATXN7 (CAG)n Forward: 5' GAGCGGAAAGAATGTCGGAG 3'

ATXN7 (CAG)n Reverse: 5' ATCACTTCAGGACTGGGCAG 3'

Fragment size: 280 bases.

The expected amplicon size is 250 bases plus three times the number of CAG repeats, i.e., 250 + (CAG)n. A BLAST analysis using the new primer sets was done using the GRCh38 reference sequence and gave an expected fragment size of 280 bases as expected because the reference sequence has 10 CAGs in the triplet repeat tract and so the amplicon size would be 250 + (CAG)10.

The annealing temperature of the new primers was calculated to be 58ºC using the NEB Tm calculator (https://tmcalculator.neb.com/).

While polymorphisms in the primer binding sites were observed in sequence data using the Ensembl genome browser, none of these polymorphisms was recorded to be found in African individuals from the 1000 Genomes dataset. This would indicate that the primer set should not fail

with samples from individuals of black African ancestry, something that was particularly important given the ethnicity of the control and patient samples and small sample size.

The newly designed primers were ordered from Inqaba Biotech, where they were manufactured and cartridge purified. The importance of the purification step was to remove all truncated fragments that could negatively impact PCR amplification.

2.2.1. SCA7 PCR design and optimization

Designed SCA7 primers were received in a lyophilized state. The lyophilized primers were diluted to a concentration of 100µM in nuclease free water. A series of trials to optimize of PCR conditions was done using the Coriell trio genomic DNA as controls. The Coriell trio consist of DNA extracted from a family with a father, mother and daughter given the sample names 1, 2 and 8, respectively.

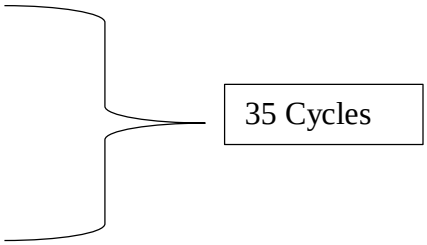
The first trial involved a temperature gradient which was within the range of annealing temperatures recommended by the NEB Tm calculator. This trial included 4% DMSO as suggested by (Smith *et al.*, 2012) using a PCR protocol was provided by Fiona Baine-Savanhu (Personal communication) with one exception – a different DNA polymerase and buffer was used. In the Baine-Savanhu protocol, the Taq polymerase used was the Promega GoTaq Long DNA polymerase. However, because the targeted region (*ATXN7*) is GC rich, the TaKaRa PrimeSTAR GXL DNA polymerase was used in my project, because it provides more robust amplification of GC rich regions, and has a proof reading ability which Taq does not (Jia *et al.*, 2014). The TaKaRa PrimeSTAR GXL DNA polymerase comes with additional reagents of a 5X PrimeSTAR GXL Buffer (Mg^{2+} plus) and PrimeSTAR GXL dNTPs. These additional reagents come at the manufacturer-designed concentrations

Table 2.2a.: *ATXN7* PRC protocol master mix, first optimization trial with a temperature gradient, adapted from *(Smith* et al., *2012)*

Reagent	1X	6X
DMSO 50%	$2\mu L$	12µL
PrimeSTAR GXL dNTPs	$2\mu L$	$12\mu L$
5X PrimeSTAR GXL Buffer (Mg^{2+} plus)	$5\mu L$	$30\mu L$
PrimeSTAR GXL DNA Polymerase	$0.5\mu L$	$3\mu L$
ATXN7 Primer 10µM (Forward)	$1\mu L$	$6\mu L$
ATXN7 Primer 10µM (Reverse)	$1\mu L$	$6\mu L$
dH_2O	$12.5\mu L$	$75\mu L$
Template/ dH_2O	1µL	

Table 2.2b.: *ATXN7* PCR profile thermocycler settings for temperatures and durations of each step for the first optimization trial with a temperature gradient, adapted from *(Smith* et al., *2012)*

Step	Temperature	Duration
Activation	95℃	3mins
Denaturing	94℃	45secs
Annealing	55-57-59℃	45secs
Amplification	72℃	30secs
Extension	72ºC	10mins
Holding	4ºC	∞



The samples were loaded at the ratio of 5µL PCR product: 5µL tracking dye. The expected band is 250+(CAG)n. Number of CAG repeats ranges from 4-34 so expected fragment size is from 274 to 364bp

The second trial introduced alterations to the master mix, in accordance to a SCA7 analysis protocol from (Cagnoli *et al.*, 2018), where the master mix had the addition of DMSO and betaine, and the polymerase used was the Go-Taq long DNA polymerase.

Table 2.3a.: *ATXN7* PCR protocol master mix, for the second trial, adapted from *(Cagnoli* et al.*, 2018)*

Reagent	1X	6X
DMSO 50%	$2\mu L$	12µL
Betain 5M	$5\mu L$	$12\mu L$
Go Taq Long DNA Polymerase (25U)	$12.5\mu L$	$3\mu L$
ATXN7 Primer 10µM (Forward)	$1.25\mu L$	$6\mu L$
ATXN7 Primer 10µM (Reverse)	$1.25\mu L$	$6\mu L$
dH_2O	$2\mu L$	$75\mu L$
Template/ dH_2O	1µL	

Table 2.3b.: *ATXN7* PCR profile thermocycler settings for temperatures and durations of each step for the second trial, adapted from *(Cagnoli* et al.*, 2018)*

Step	Temperature	Duration
Activation	95℃	3mins
Denaturing	94℃	45secs
Annealing	57℃	45secs
Amplification	72℃	30secs
Extension	72ºC	10mins
Holding	4ºC	∞

35 Cycles

The final trial was an adaptation of the protocol by (Cagnoli *et al.*, 2018) with the alteration of the Taq polymerase used. In this trial, the TaKaRa master mix was used, for increased sensitivity to

GC rich regions. There was no change in the PCR profile used, but the PCR master mix volumes were scaled up to 50µL in order to produce more DNA for the sequencing step that followed.

Table 2.4.: *ATXN7* PCR Protocol master mix, third and final trial, adapted from *(Cagnoli* et al., *2018),* **with the alteration of the TaKaRa Taq polymerase.**

Reagent	X1/µL	X6/µL
X5 Buffer	5	30
DNTPs	2	12
Forward primer 10µM	1	6
Reverse primer 10µM	1	6
Taq polymerase	0.5	3
DMSO 50%	2	12
Betaine 5M	5	30
dH2O	7.5	45
Total	24	144

Amplicons sizes from all the trials were analysed using agarose gel electrophoresis on a 2% agarose gel, using 1X TriTrack loading dye. The amplicon sizes were measured using the molecular weight marker, Low range GeneRuler Ladder.

2.2.2. Purification of PCR products

Amplified laboratory controls were purified using the ProNex Size Selective Purification Kit (Promega) which is a magnetic bead-based system. The Promega website (www.Promega.com) provides a manual that stipulates the bead to sample ratio to use when conducting size selective purification using the ProNex Size selective purification Kit. Amplicons were purified using ProNex beads from Promega at a ratio of 1:1.5 (v/v) sample to beads, as stated by the manufacturer. This ratio of beads to sample was chosen because the expected amplicon size is between 250-300 bases.

A 2% agarose gel was prepared to assess the purified amplicons. This was to ensure that DNA was not lost during the purification and to assess the concentration of amplicons after purification. As a concentration comparison, the purified DNA was run on the same 2% agarose gel as unpurified

amplicons. This provided a side-by-side comparison of the DNA concentration difference between the purified and unpurified amplicons. DNA concentrations in each of the purified amplicons was measured using the Qubit4 1x dsDNA HS assay kit, according to manufacturer instructions.

2.3. Sanger sequencing approach

Sanger sequencing was used to confirm that the amplified fragments from the controls contained the expected SCA7 sequence. The sequencing was performed by Inqaba Biotech. The Sanger sequencing results from Inqaba Biotech were files of sequence data which were viewed as chromatograms. Forward and reverse sequence results for each sample were presented in separate files and viewed using the SnapGene software. The Sanger sequence chromatograms were mapped to the GRCh38 reference sequence in the SnapGene software.

2.4. 2.4 Sample preparation for PCR

2.4.1. Preparation of control samples

Initial concentrations of each of the control samples were provided in the sample information. Each sample was diluted to 50ng/µL by adding TE Buffer (calculations shown in Appendix 1)

The above-mentioned dilution protocol was used to produce control DNA stock samples of approximately 50ng/µL each. The final concentrations for each of the diluted samples were assessed using a Qubit4 fluorometer from Thermo Fisher. The integrity of the control sample DNA was investigated by running 1µL DNA plus 4µL tracking dye for each of the diluted control samples on a 0.8% agarose gel. Expected band for high molecular weight genomic DNA is around 5Kb. High integrity DNA is expected to only have one thick band rather than a smear, which would be indicative of DNA degradation. A 1Kb Plus ladder was run on the same gel to assess fragment size.

2.4.2. Preparation of patient samples

The DNA concentration in the patient samples was assessed using a Qubit4 fluorometer, where 1µL of each sample was analysed. Because the initial Qubit readings were out of range, indicating high DNA concentrations, a 50x dilution was done for each of the DNA samples (1µL stock DNA

and 49µL TE buffer). An assessment of DNA concentration was done using a Qubit4 fluorometer, where 1µL of the 50x dilution was measured.

1µL of 50x dilution of patient sample T1 = 23.6ng/µL

23.6 x 50 = 1 180 ng/µL = Stock T1 DNA

1µL of 50x dilution of patient sample T2 = 9.20ng/µL

9.20 x 50 = 460 ng/µL = Stock T2 DNA

The patient DNA samples T1 and T2 were diluted to 50 ng/µL by adding TE Buffer, using the calculations in Appendix 1.

The volumes required to make a stock at 50ng/µl were calculated, and final concentrations confirmed using a Qubit4 fluorometre. A consistent 50ng/µL concentration of DNA template was prepared to serve as the template DNA for the PCR master mix, to maintain uniformity with the concentrations of the control sample templates that were used in this project.

Qubit quantification of T1 working Stock = **50ng/µL**

Qubit quantification of T2 working Stock = **45.8ng/µL**

The integrity of the test DNA samples was observed by running 1µL sample with 4µL of tracking dye, on a 0.8% agarose gel along with a 1µL 1Kb plus ladder with 4µL tracking dye.

2.4.3. Patient and Control sample PCR

The optimized *ATXN7* PCR protocol, stated above, was used to amplify the SCA7 region of the 10 control samples and 2 test samples. The PCR products were run on a 2% agarose gel at the ratio of 5µL PCR product: 5µL tracking dye. The DNA ladder used was 5µL of GeneRuler low range ladder. After identifying the bands of the expected size (approximately 300bp) in the control samples, as well as the expected test sample amplicon sizes, the test and control PCR product DNA was quantified using Qubit analysis. The DNA quantification was done before amplicon purification, so as to determine percentage yield when comparing it to the amount of DNA after

purification. The test and control samples were purified using ProNex beads, as described earlier, and the purified amplicon concentrations were assessed using a Qubit4 fluorometer.

Purified patient and control DNA samples were run on a 2% agarose E-gel from Invitrogen. This is a highly sensitive gel which produces sharp bands. For a visible band on the E-gel, a minimum of 10ng of DNA is required. Given the expected size of fragments, the NEB calculator was used to calculate the fmol needed for each sample (54fmol). The purified control and test samples were each diluted to 54fmol in a final volume of 20µL including loading buffer, and loaded onto the E-gel. The gels were run in 2 batches, with the first batch containing a DNA ladder, 5 control samples (1A, 1B, 1C, 1D and 1E), and the 2 test samples. The second batch contained a DNA ladder, 5 control samples (1F, 1G, 1H, 2A and 2B), and the 2 test samples.

The remaining volumes of test and control sample amplicons were stored at 4°C for Nanopore library preparation.

2.5. Nanopore sequencing

2.5.1. Nanopore Library Preparation

Library preparation was done according to manufacturer instructions using the provided protocol from Oxford Nanopore Technologies. The protocol used was the "Amplicons by Ligation" (SQK-LSK109). After ensuring that all the materials, consumables and equipment were available, as stated on the "Before start checklist", the outlined steps in the protocol were followed according to manufacturer's instructions:

Step 1- DNA Repair and end preparation

Step 2 - Bead Clean Up

Step 3 - Adaptor Ligation

2.5.2. Oxford Nanopore sequencing using the Flongle flow cell

The Oxford Nanopore sequencing device is connected to a controller called the MinIT, which is essentially a miniature Linux computer that runs the MinKNOW software from which the settings are controlled, and the sequencing initiated and monitored. Sequencing data can be observed in real time. All the steps were followed according to manufacturer's instructions.

2.5.3. Oxford Nanopore Sequence Data Collection

The MinION sequencing device is controlled by the MinKNOW interface which is a collection of software that allows for real-time base calling, conversion of fast5 reads into fastq reads, as well as real filtering. The software versions used in this project were:

MinKNOW 21.02.2
MinKNOW Core 4.2.4
Bream 6.1.10
Guppy 4.3.4

After each sequencing run was completed, data files were transferred from the MinIT to a laptop computer using the Bitvise SSH Client 8.46 software. These data files included the following:

fast5_fail: This file contains the fast5 reads with a q-score less than 7
fast5_pass: This file contains the fast5 reads with a q-score greater than 7
fastq_fail: This file contains the fastq reads with a q score less than 7
fastq_pass: This file contains fastq reads with a q-score greater than 7
Duty time: This is a text file that shows the activity of each channel throughout the sequencing run. The text file is the form of an excel spreadsheet and shows the active pores at each time in the sequencing run and the number of reads each pore produced
Final summary: This is a text file that contains all the information about the files generated during the run, including the date and time of the sequencing run. This information indicates the sequencing device's identity, the flow cell identity and contents of all the read files (pass and fail reads in both fast5 and fastq)
Report: This is a report in PDF format, and provides detailed information about pore activity during the sequencing run. This includes the number of pores that were active over the sequencing run and when the pores were active. The report gives the total number of reads produced, the mean length of reads, the temperature and voltage changes throughout the sequencing run and the sequencing yields. The reports also contains histograms that allow the analyst to observe mean lengths of reads produced and the reads that were successfully base called.

Sequencing summary: This is a text file that gives the detailed activity of each channel. The details include the number of reads each pore produced, the amount of time the sequencing process took to produce a read, whether the reads were pass or fail, the temperature of each pore during the sequencing and the identity of each pore. The mean q scores of the reads produced by each of the pores is also given in this summary

2.5.4. Analysis of sequencing quality

With the purchase of an Oxford Nanopore sequencer there is the option to become a member of the Nanopore Community. This resource provides a platform to communicate with other Nanopore users and share experiences and ideas. Most importantly, this gives access to useful analysis tools that can be used for basic quality control as well as other types of analyses. These tools are provided on a Nanopore cloud infrastructure called EPI2ME, and is accessed using the EPI2ME desktop agent that is installed on a local computer. For the basic quality control analyses in this project, EPI2ME was used to process fastq-pass reads, align them to the reference sequence GRCh38, determine the total number of reads, the number of reads that map to the reference sequence GRCh38, which chromosomes these reads map to, the average q scores of the reads and the average lengths of the reads.

During Nanopore sequencing, a data folder is created with fastq_pass reads that are stored in individual files, each containing 1000 reads. These files were joined to make a single file, joint.fastq, which contained all the fastq_pass reads for each of the respective samples sequenced in this project. This process was done using the MobaXterm terminal, navigating to the fastq_pass folder and using the command:

cat *fastq_pass > joint.fastq

Before QC analysis, the EPI2ME desktop agent requires the following steps be completed:

1. Upload of the reads being analysed
2. A declaration that if data is of human origin, that the user has relevant permission
3. Setting parameters of minimum and maximum length of the reads being analysed
4. Selecting the alignment reference GRCh38

5. Accepting and running the alignment.

Figure 2.1 shows the interface of the EPI2ME desktop agent. In this example we see the representation of the data. The data contains the number of reads analysed, the number of reads successfully aligned, the percentage accuracy and the yield. The analysis also gives the average size of reads, average q-score of the reads and the specific chromosomes the reads align to.

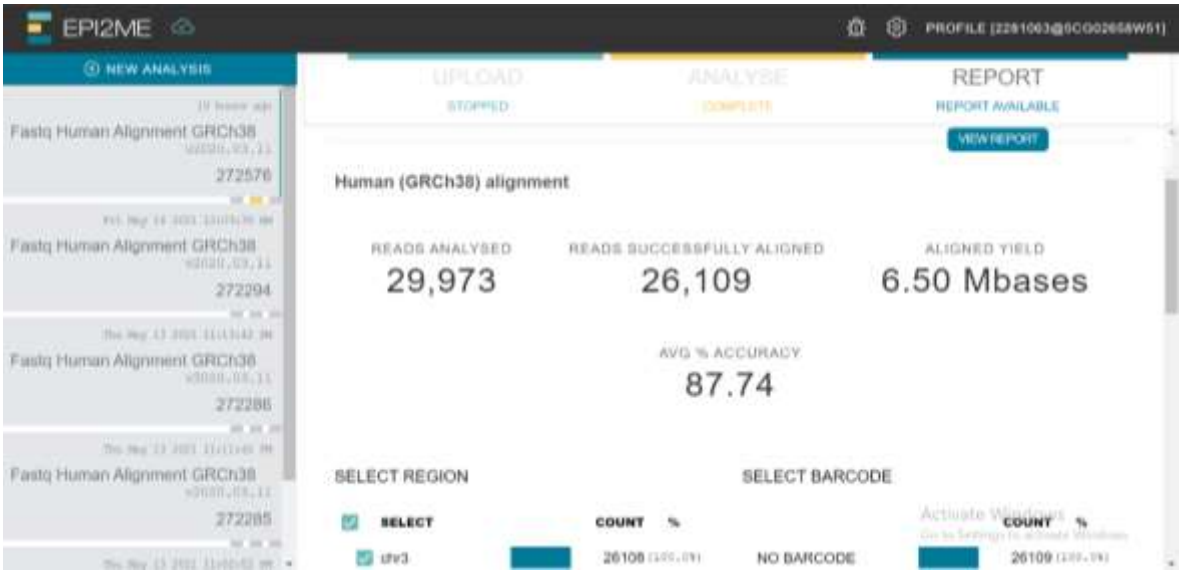


Figure 2.1: Visualization of the EPI2ME desktop agent giving basic QC alignment of the Control 1C to reference GRCh38

Oxford Nanopore provide an analysis platform called the EPI2ME Labs notebook and server. This is a Docker-run container that allows the user to make use of different predesigned sequence analysis protocols. The interface to the EPI2ME Labs notebook is a Jupyter terminal that uses the Python scripting language. This terminal can be used to run and modify Nanopore analyses, and create their own pipelines by creating a separate Notebooks on EPI2ME Labs and installing downloading different analysis software onto the server using the !pip command.

Due to the MinION's high depth of sequencing, all the samples had a very large number of reads at a minimum q-score 7. As part of the basic QC, the reads were filtered using the nanofilt software. Nanofilt was downloaded onto the EPI2ME Labs server, and used to filter the reads in the joint.fastq files. Data was filtered to remove reads that were smaller that 250bp, larger than 500bp

and with a q score less than 12. The size filter removes any truncated reads, or reads that resulted from artefacts of library preparation (such as ligation of PCR of PCR products during adapter ligation). The q-score filter was used to remove reads of low quality. The command used was:

!cat joint.fastq | NanoFilt -q 12 -l 250 --maxlength 350 > jointf.fastq

The redirection ensured the creation of a filtered read file named jointf.fastq.

There was a significant reduction in the number of passed reads after this filtration, and these filtered reads were once again aligned onto the GRCh38 reference to see the improvement in aligned read percentage and the alignment onto chromosome 3 (Chr 3) where the *ATXN7* gene is located.

The filtered reads were then processed on the EPI2ME cloud to align against GRCh38 as described earlier. Figure 2.2 shows the EPI2ME basic QC alignment to GRCh38 of the same sample aligned in Figure 2.1. This illustrates the smaller number of reads produced, the higher average q-score, the higher accuracy of alignment and all sequences aligning to Chr 3.

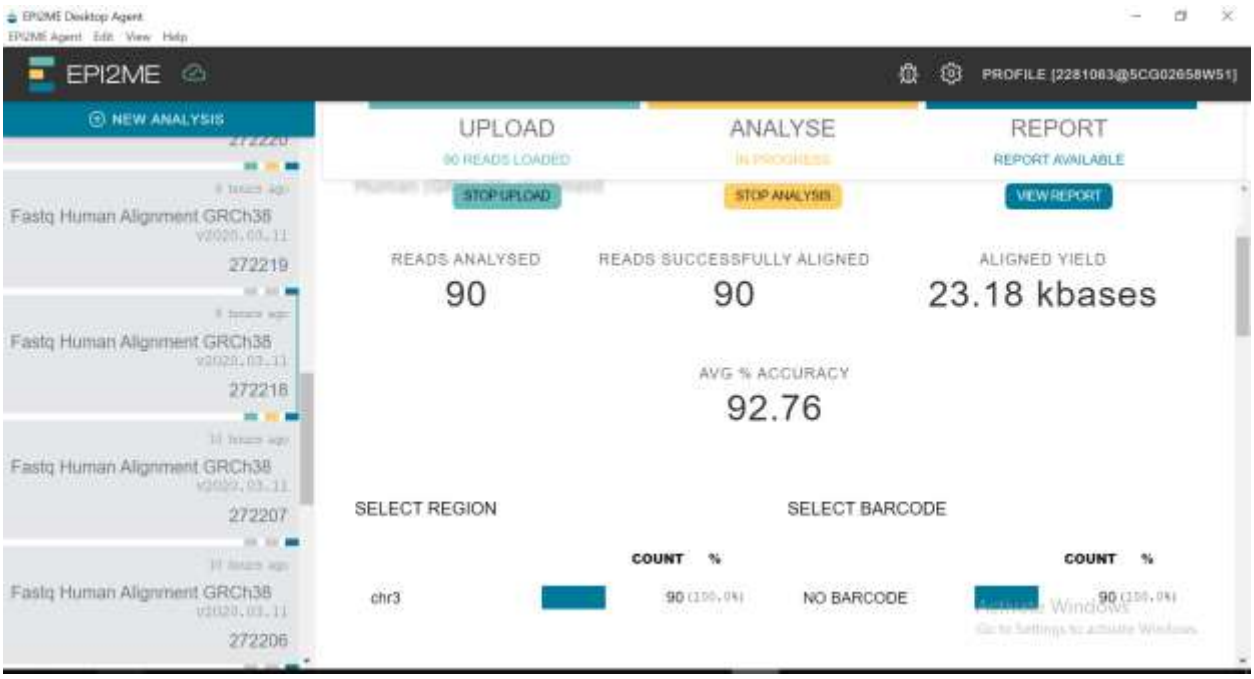


Figure 2.2: EPI2ME cloud QC alignment to GRCh38 for Control 1C, showing the lower number of reads, the increased accuracy and the higher quality of reads.

As shown in figure 2.2, all filtered reads could be aligned to Chr3. Although a large number of reads were lost after filtering, the accuracy (a measure of average sequencing quality) increased from 87.74 to 92.76% accuracy. These findings support the use of a high minimum q-score filter, such as minimum q-score of 12.

The next QC step was to check whether the reads aligned to the SCA7 CAG triplet repeat region in the *ATXN7* gene in Chr 3. The GRCh38 reference FASTA file was downloaded from the NCBI genome browser and named **sca7fullref.fasta**. Minimap2 was downloaded and installed onto the EPI2ME Labs server and used to analyse Nanopore reads. Minimap2 was used to convert the filtered sample data files (jointf.fastq) into SAM files using the Python command - **!minimap2 -ax map-ont /EPI2MElabs/sca7fullref.fasta jointf.fastq > jointf.sam.**

The SAMtools software was downloaded and installed onto the EPI2ME Labs server and jointf.sam files were zipped into BAM files and the BAM files were sorted and indexed. This was done using the commands below:

!samtools view -bS jointf.sam > jointf.bam

!samtools sort jointf.bam > sortedjointf.bam

!samtools index sortedjointf.bam

The IGV software was used to visualize the indexed BAM files mapped to the downloaded GRCh38 reference sequence. The IGV software allows a user to align sorted and indexed BAM files to a reference sequence of the user's choice. In this analysis, the GRCh38 reference sequence was downloaded from the Ensembl genome browser and uploaded onto the IGV software to serve as the reference sequence. The sorted and indexed bam files produced from the filtered reads in the above steps, were aligned to the reference sequence to visualise alignment and presence of indels. The reference sequence GRCh38 has 10 CAG repeats in the SCA7 triplet repeat region, from position 53128 to position 53158. All the sequences mapped perfectly to the SCA7 triplet

repeat region and showed indels at position 53128 or position 53158, which are the start and the end of the triplet repeat sequence in the SCA7 region.

2.6. Nanopore data analysis to measure the size of the CAG triplet repeat tract

EPI2ME Labs contains an index of notebooks that can be used for a variety of analysis methods. The notebooks each contain different software and use the Jupyter command line for analysis. The appropriate notebook for identifying the size and identity of the insertions and/or deletions observed in the basic QC step, is the variant calling workflow. The software tools recommended by Nanopore for EPI2ME analyses are flye, canu and medaka. However, these tools are specifically built to analyse long read data and not shorter reads, like the ones that were generated in this project. Attempts to analyse SCA7 data were with the message “Implausibly small genome size”.

Given that the variant calling software that is recommended by the manufacturer (Oxford Nanopore Technologies), was incompatible with the size of reads produced through the sequencing processes done in this study, an alternative workflow was required. The new approach examines size of the region that contains the triplet repeats.

Figure 2.3 below shows the designed workflow for quantification of the size of the triplet repeat tract in the test and control samples. The first step was to filter the reads using nanofilt. A q-score of 10 was used to remove low quality reads, and size specifications were included in the nanofilt command to exclude short- and long-read artefacts, as described earlier.

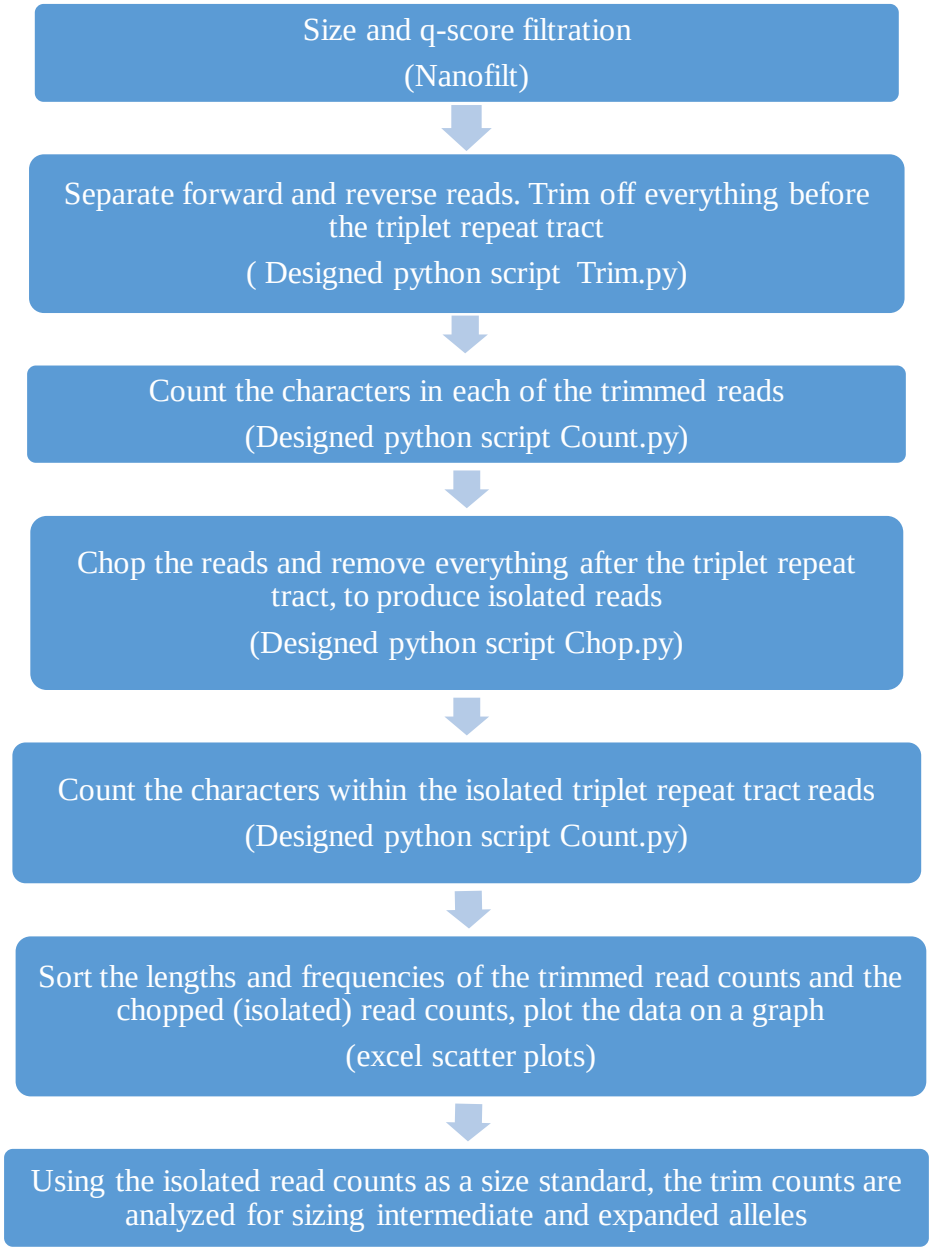


Figure 2.3: Data analysis workflow to determine the fragment size of the triplet repeat tract in the sequence data.

The command used to filter the reads was:

!cat joint.fastq | NanoFilt -q 10 -l 250 --maxlength 350 > jointf.fastq for control samples

!cat joint.fastq | NanoFilt -q 10 -l 250 --maxlength 800 > jointf.fastq for patient samples

The output was the filtered read files, which were named jointf.fastq for each of the control and patient samples.

The trimming process was done by using a python script that simultaneously separates forward and reverse reads, and trims everything before the occurrence of the triplet repeats. The forward reads begin with the repeat sequence CAGCAGCAG and the reverse reads end with the repeat sequence CTGCTGCTG. The python script designed for the separation and trimming process was named Trim.py (Appendix 2). A different command was run to obtain the reverse and the forward sequences.

Python3 Trim.py CAGCAGCAG CTGCTGCTG input.fastq Trimmed_Forward.fastq Trimmed_forward.fasta
Python3 Trim.py CTGCTGCTG CAGCAGCAG input.fastq Trimmed_Reverse.fastq Trimmed_Reverse.fasta

The input files for these commands are the filtered reads which were named jointf.fastq. The output files are the trimmed reverse and forward read files separated. These files are produced in fastq and fasta format.

A python script was designed to count the number of characters (i.e., nucleotides) in the trimmed reads, and calculate the frequency of read lengths. The command used is:

```
Python3 count.py Trimmed_Reverse.fasta > Tim_counts.txt
```

The read count txt file was sorted into numerical order using the Sort command:

```
Sort -n -k1 Trim_ counts.txt > sorted_Trim_counts.txt
```

The sorted Trim_count text files were opened with MS Excel, and a scatter plot was created for each of the control samples to visualise the distribution of the read frequency. The positions of peaks in these plots is related to the size of the triplet repeat tract in the sample, but do not represent the exact size of the repeat because they also contain flanking sequences. Unlike capillary electrophoresis-based methods where an internal standard is added during analysis to determine length of triplet expansion, here there is no internal size standard. Instead, an alternative approach was developed where the trimmed reads were taken through a "chopping" process to remove the sequences upstream and downstream of the triplet repeat.

The repeat sequence tract in the wild type alleles is conserved. The assumption was made that the sequences have no variations within the triplet repeat tract and all have the same 3 bases following the end of the repeat sequence tract. This allowed for the design of the chopping python script (Appendix 3). The trimmed reads were chopped at the end of the triplet repeat tracts and this produced reads that were made only from the triplet repeat tract. The command used to chop the trimmed sequences was:

Python3 Chop.py CTGCTGCCG CAGCAGCAG Trimmed_reads.fastq chop_reads.fastq chop_reads.fasta

This command truncates sequences that occur after the 3 bases that flank the end of the triplet repeat tract. The output files contain the triplet repeat tract and the 3 bases that flank it. These reads are termed "isolated reads". A python script was designed to count the number of characters in these isolated reads, and hence determine the number of triplet repeats each read had (Appendix 4). The command used is:

Python3 count.py chop_reads.fasta > counts.txt

The read count txt file was sorted into numerical order using the Sort command:

Sort –n –k1 counts.txt > sorted_counts.txt

The sorted count text files were opened with MS Excel and a scatter plot was produced for each of the control samples to visualise the distribution of the read frequency.

While the chopping process does not allow for the identification of intermediate of expanded alleles, it does accurately count the triplet repeats in the conserved region. To calculate the length of any intermediate or expanded alleles, the known length of the conserved allele provides a sizing standard.

The graphical representation of the trimmed_counts showed the peaks of the conserved allele reads, as well as the intermediate or extended allele reads. Since the chopped read counts provide the exact number of repeats in the first peak, the number of bases between the first peak and the peak that follows, is the difference in the number of triplet repeats between the two allele reads.

This workflow is effective for both control and patient samples as it requires no external size standards and uses the sample's own conserved region as a size standard to measure any intermediate or expanded allele read lengths.

2.7. Assessing the costs associated with the two assays

2.7.1. Analysis of time required for CE relative to Oxford Nanopore Sequencing

The amount of time it takes to complete each step of the respective analysis methods is important to document. This assessment helps to determine the costing of each method, as it stipulates the amount of time a device would be in use, as well as the amount of time a scientist would take to complete an assessment in order to determine their salary expense for each assessment.

CE steps and duration of each step

From personal communication with Fiona Baine-Savanhu, from the NHLS, information was provided on the list of steps taken in CE, and the time taken for each.

The steps listed were as follows:

- PCR, with fluorescent primers (master mix preparation and thermocycler run-time)
- PCR product purification

- Sample preparation for CE and CE run-time
- Data collection and analysis to determine a result

The time required for report writing was not included as it is relative to individual laboratory protocols and will be similar irrespective of which analysis method is used.

Nanopore sequencing *steps and duration of each step*

The duration of each of the steps done in this project was recorded throughout the project and average times determined.

The steps required for Nanopore sequencing are as follows:

- PCR (master mix preparation and thermocycler run-time)
- PCR product purification
- Nanopore sequencing library preparation
- MinION sequencing run-time
- Data collection and analysis to determine result

2.7.2. Cost assessment of CE relative to the Oxford Nanopore approach

To assess the costs of the currently used method (PCR followed by capillary electrophoresis to determine the fragment size and number of triplets) and Oxford Nanopore Sequencing is complex and needs to consider several different costs related to equipment, consumables, time and overheads.

Samples are usually tested in batches and therefore the costing was done by calculating the cost of running 10 samples using each of the respective methods. The cost comparison will take into consideration the equipment used and the reagents required for the procedure in question. The set-up costs for the laboratory infrastructure required for these tests will be considered separately, as items of equipment, such as pipettes, vortex and centrifuge equipment, would already be present in a molecular diagnostic laboratory and would therefore not need to be purchased specifically for these assays. In addition, the costs for extracting the DNA will be the same between the two methods and will therefore not be included in the assessment.

The elements that would be factored into costing the assessment and comparison are divided into three sections

1. Laboratory consumables
    a. PCR costs:

Although the PCR costs are similar between the two methods, it is an essential first step and there are some notable differences. The reagents, tubes, tips and gloves will be factored in, but not the costs related to the PCR machine.

    b. Downstream analysis costs:

Current method: costs for capillary electrophoresis and reagents were obtained from Thermo Fisher Scientific.

Nanopore method: Oxford Nanopore sequencing was costed using the Oxford Nanopore MinION device, the Flongle flow cell and the associated reagents.

2. The labour costs:

The labour costs included the salary of the person doing the work and the duration of time spent on different activities. For example: laboratory work is performed by a Medical Scientist and the calculations included the salary of the scientist (calculated on an hourly rate). The same was done for a senior scientist who would review the results and the report to the referring doctor or genetic counsellor.

3. Overheads:

After calculating the combined costs for laboratory consumables and labour, the overheads were calculated as follows: Wastage (10% of total cost of the test) + Laboratory fees (based on NHLS rates as 20% of the cumulative total) + VAT (15% of the cumulative total).

3. Results

3.1. Primer Design

Primers designed using the software Primer3 for the amplification of the triplet repeat tract in the *ATXN7* gene were as follows:

ATXN7 (CAG)n Forward: 5' GAGCGGAAAGAATGTCGGAG 3'
ATXN7 (CAG)n Reverse: 5' ATCACTTCAGGACTGGGCAG 3'

The expected fragment size is 250 + (CAG)n. Tools that perform *in silico* PCR runs, such as BLAST and the UCSC genome browser, use the GRCh38 reference sequence, which has 10 CAG repeats. The Primer3 tool also uses the GRCh38 reference sequence for primer generation, hence the expected fragment size is 250 + (CAG)10 = 280 bp

The primer design matched the criteria for good primers, as they comply with reviewed recommendations of PCR primer design (Dieffenbach, Lowe and Dveksler, 1993). The forward primer is 20nt in length with a melting temperature of 58.7^{o}C, a GC percentage of 55% and no hairpins structures are predicted. The reverse primer is 20nt in length with a melting temperature of 59.4^{o}C, a GC percentage of 55% and no hairpins.

```
PRIMER PICKING RESULTS FOR

Template masking not selected
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any_th  3'_th  hairpin  seq
LEFT PRIMER          3   20   58.72   55.00     0.00   0.00     0.00 GAGCGGAAAGAATGTCGGAG
RIGHT PRIMER       282   20   59.38   55.00     0.00   0.00     0.00 ATCACTTCAGGACTGGGCAG
SEQUENCE SIZE: 335
INCLUDED REGION SIZE: 335

PRODUCT SIZE: 280, PAIR ANY_TH COMPL: 0.00, PAIR 3'_TH COMPL: 0.00

    1 AGGAGCGGAAAGAATGTCGGAGCGGGCCGCGGATGACGTCAGGGGGGAGCCGCGCCGCGC
        >>>>>>>>>>>>>>>>>>>

   61 GGCGGCGGCGGCGGGCGGAGCAGCGGCCGCGGCCGCCCGGCAGCAGCAGCAGCAGCAGCA


  121 GCAGCAGCAGCCGCCGCCTCCGCAGCCCCAGCGGCAGCAGCACCCGCCACCGCCGCCACG


  181 GCGCACACGGCCGGAGGACGGCGGGCCCGGCGCCGCCTCCACCTCGGCCGCCGCAATGGC


  241 GACGGTCGGGGAGCGCAGGCCTCTGCCCAGTCCTGAAGTGATGCTGGGACAGTCGTGGAA
                           <<<<<<<<<<<<<<<<<<<

  301 TCTGTGGGTTGAGGCTTCCAAACTTCCTGGGAAGG


KEYS (in order of precedence):
>>>>>> left primer
<<<<<< right primer
```

Figure 3.1: An extract of the primer design outcome in Primer3 (https://primer3plus.com/). The CAG repeat sequence embedded in the sequence, in rows marked 61 and 121.

A BLAST analysis of the primers using NCBI genome browser (https://blast.ncbi.nlm.nih.gov/Blast.cgi) showed specificity to *ATXN7*. A virtual PCR using the UCSC *in silico* PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr) against reference genome GRCh38 resulted in specific amplification of *ATXN7*, with no alternative PCR products predicted. DNA sequences of African individuals that are in the 1000 Genomes project database were analysed by downloading the VCF files for the genome data, and using the VCFtools software to assess the presence of SNPs within the coordinates of the primer binding sites. This was to ensure that there would be no SNPs in the primer binding regions, which could cause PCR failure. No sequence variations were observed in the primer binding sites of the sequence data from the 1000 Genomes datasets.

3.2. PCR Optimization

A detailed process of PCR optimization was then carried out to produce a robust efficient PCR protocol for production of pure and highly concentrated amplicons. The laboratory control DNA template was used for the PCR optimization and involved 3 trials. The first trial made use of the published SCA7 protocol from (Smith et al., 2012), with a temperature gradient to obtain the optimum temperature.

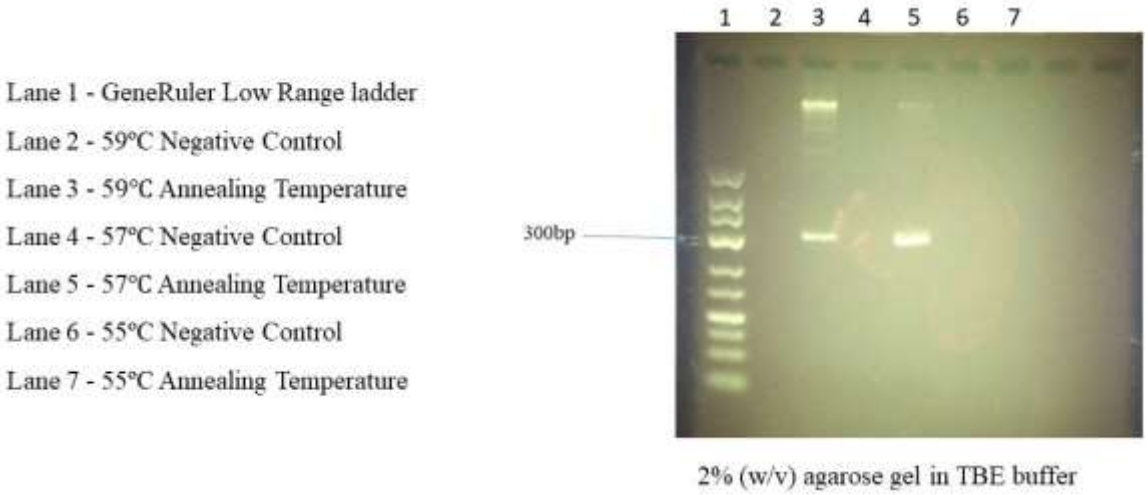


Figure 3.2: Agarose gel image of the *ATXN7* PCR products of the first optimization trial, adapted from *(Smith et al., 2012)*. A fragment of ~300bp was present in lanes 3 and 5, at annealing temperatures of 59 and 57°C, indicating that the correct fragment was amplified. Additional high molecular weight fragments were observed.

Lane 3 showed the expected band with an annealing temperature of 59°C, but also showed non-specific binding that generated a high molecular weight PCR fragments. Lane 5 showed that with a 2°C drop in annealing temperature, there is an increased yield of the expected PCR product, and very feint higher molecular weight bands.

The presence of nonspecific PCR products indicates that the PCR protocol was not optimal. The second trial included the same thermocycler settings with a temperature gradient, with an addition of betaine (1M final concentration), to improve annealing. The addition of betaine was inspired by the PCR protocol from a study of diagnostic method for spinocerebellar ataxias (Cagnoli *et al.*, 2018). The paper by Cagnoli *et al.* used a combination of DMSO and betaine, with the Go-Taq Long polymerase enzyme. My next optimization trial comprised of similar reagents to the Cagnoli *et al*. paper,

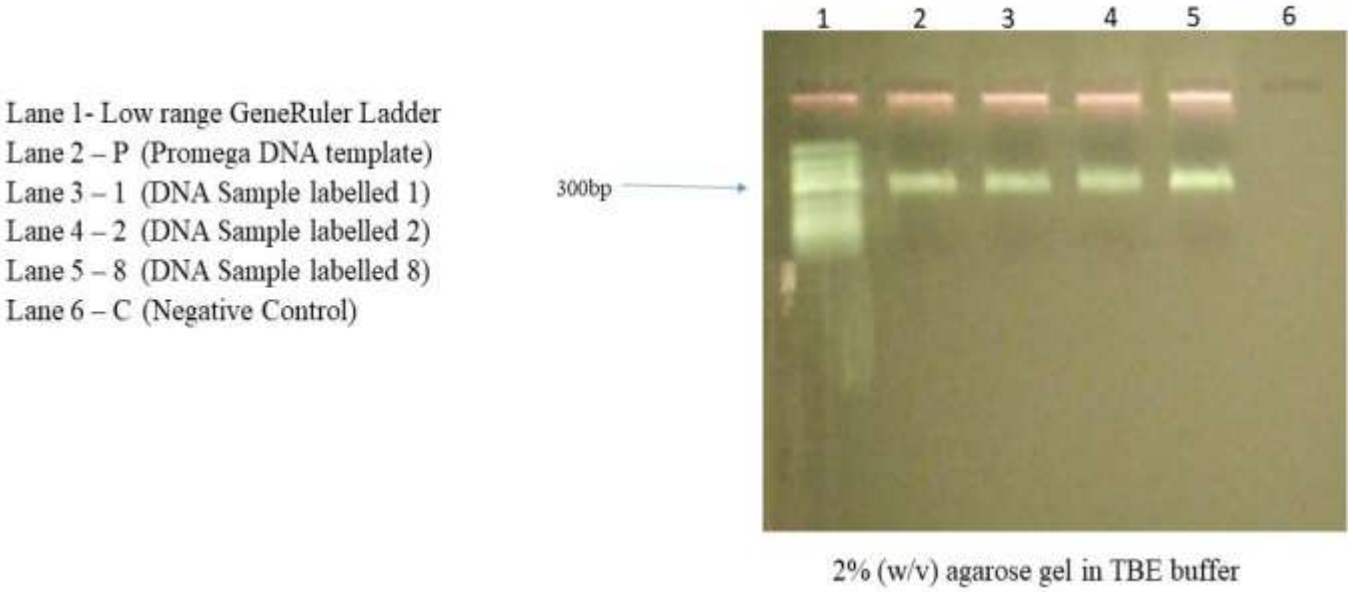


Figure 3.3: Agarose gel image of the *ATXN7* PCR products of the second optimization trial, at an annealing temperature of 57ºC, adapted from*(Cagnoli* et al., *2018).* **Fragments of ~300bp were present in lanes 2, 3, 4 and 5 and compared to the previous trial, no additional high molecular weight fragments were observed.**

The samples were loaded at the ratio of 5µL PCR product: 5µL tracking dye. A single clean band was observed in the lane containing the positive control and in lanes containing the control samples 1, 2 and 8. No band was observed in the lane containing the negative control. It is important to note that the image is not very sharp. This is because the original imaging device for the agarose gel was not available so a low-resolution cell phone camera was used instead.

The third and final trial was done using the optimum annealing temperature observed in the first and second trials. The addition of betaine (1M) and DMSO (4%) proved to be part of optimal conditions in the second trial, and the TaKaRa Taq polymerase was used because of its ability to amplify long GC rich regions with minimal errors.

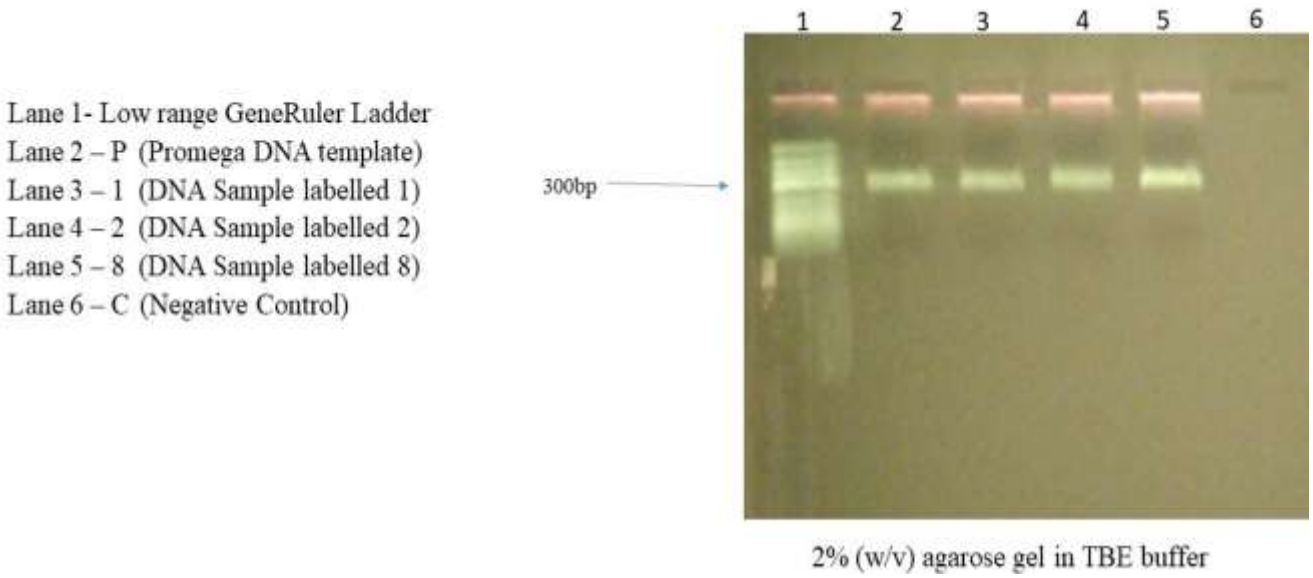


Figure 3.4.: Agarose gel image of the *ATXN7* PCR products of the third and final optimization trial, at an annealing temperature of 57°C, adapted from *(Cagnoli* et al., *2018),* **with the alteration of using the TaKaRa Taq polymerase. Fragments of ~300bp were present in lanes 2,3,4 and 5.**

The samples were loaded at the ratio of 5µL PCR product: 5µL tracking dye. Clear bright bands observed at the expected size for each of the PCR products from the third and final *ATXN7* PCR protocol. The gel image showed that there was no non-specific binding and only one band of the expected size was observed. The negative control shows no amplification, showing that there was no contamination in the PCR preparation.

3.3. Purification of PCR products

The purified *ATXN7* PCR amplicons were assessed on a 2% (w/v) agarose gel. This was to ensure that DNA was not lost during the bead purification. Unpurified *ATXN7* PCR amplicons were assessed on the same gel to provide a visual side-by-side comparison of DNA concentrations between the purified and unpurified amplicons. Figure 3.5 shows the agarose gel image of the above mentioned electrophoresis run outcome.

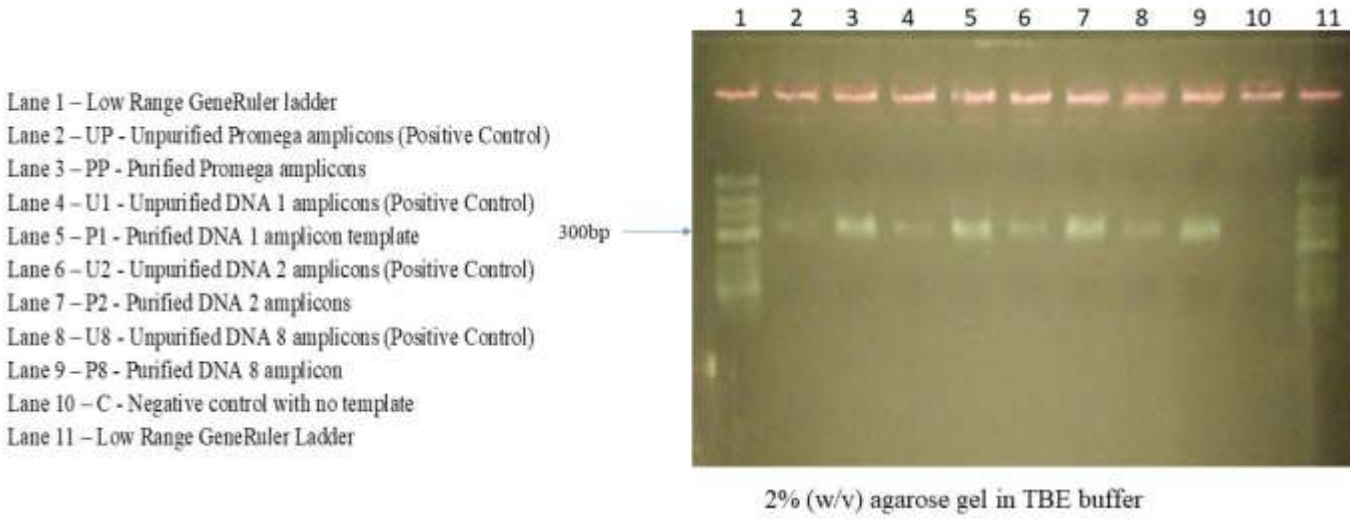


Figure 3.5: Agarose gel image of the *ATXN7* PCR products, showing the purified and unpurified PCR products.

Bands are visible in all the lanes, signifying a successful purification of amplicons. The purified amplicons for each of the samples has a brighter band that the unpurified samples, showing the significant increase of DNA concentration after amplicon purification.

Each purification gave 20µL of clean amplicon. Purified amplicon concentrations measured using the Qubit4 1x dsDNA HS assay kit and the results were as follows:

P - 6.24ng/µL x20 = 124.8ng of DNA

1 – 5.04ng/µL x20 = 100.8ng of DNA

2 – 4.76ng/µL x20 = 95.2ng of DNA

8 – 6.27ng/µL x20 = 125.4ng of DNA

3.4. Sanger sequencing

The purified PCR products from the control sample 8, were sent to Inqaba Biotech for Sanger sequencing using the designed *ATXN7* primer sets. Sanger sequencing was done to confirm that the primers had indeed amplified the SCA7 triplet repeat region, to characterize the length of the triplet expansions, and examine first-hand the reported challenges of using this approach to sequence triplet expansion tracts.

The initial attempts of Sanger sequencing done by Inqaba, were unsuccessful. Given the difficulty faced in order to successfully optimise the *ATXN7* PCR, it was suggested that Inqaba attempt to

repeat the Sanger sequencing, using a combination of betaine at DMSO. The repeated Sanger sequencing attempt was successful and sequencing was done for the control sample amplicons submitted to Inqaba Biotech.

The sequencing reports for the laboratory controls, were received from Inqaba Biotech, for both the sequencing with DMSO and the sequencing with a combination of betaine and DMSO. Sanger sequencing was done in both the forward and reverse direction. The sequencing reports were chromatograms that were analysed using the SnapGene application. Sanger sequences produced by Inqaba were mapped onto the GRCh38 reference sequence and the chromatograms were viewed in a multiple sequence alignment, in order to simultaneously observe the forward and reverse sequences.

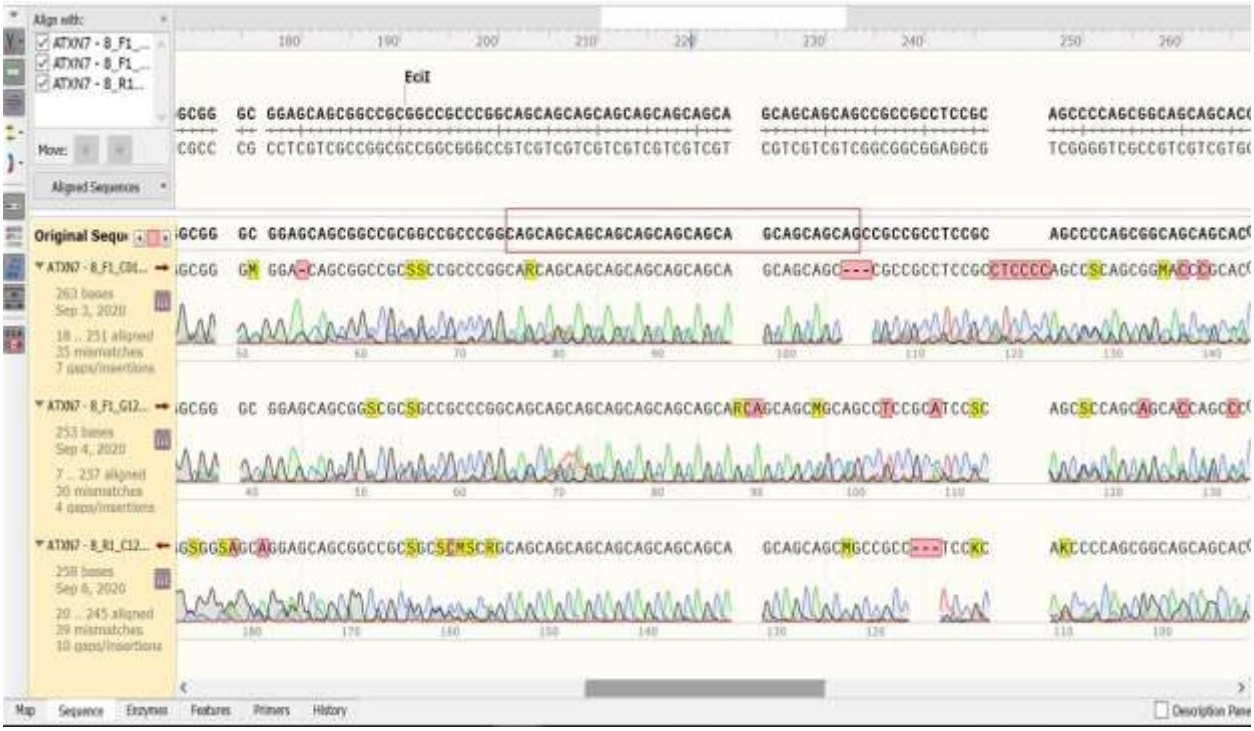


Figure 3.6: A SnapGene image of the visualization of Sanger sequence chromatograms of control sample 8 *ATXN7* PCR amplicons, mapped to the GRCh38 reference sequence

1. The upper sequence alignment *ATXN7*-8_F1_C01 is the forward Sanger sequence, using betaine. The sequence shows 9 conclusive repeats, 3 gaps, and 1 ambiguous base.

2. The middle alignment ATXN-8_F1_G12 is a forward Sanger sequence, using DMSO. The sequence shows 9 conclusive CAG repeats with 2 ambiguous bases

3. The third alignment ATXN-8_R1_C12 is the reverse Sanger sequence, using DMSO. The sequence captured almost all the CAG repeats accurately, with one ambiguous base (CMG) in place of the last CAG position on the reference sequence. This appears to be the closest to the reference

Sanger sequencing confirmed that the amplified region is indeed the SCA7 region of the *ATXN7* gene. Sequencing of this region was most effective using the reverse primer and the addition of DMSO. This was done to analyse the effectiveness of Sanger sequencing as a sole diagnostic tool for triplet repeat sequences. The forward sequences had gaps and ambiguities and could not accurately identify the number CAG repeats. The reverse sequencing gave the best result, although it too had sequence ambiguity.

3.5. Preparation of SCA7 negative control samples

The 10 SCA7 negative control genomic DNA samples provided by the SBIMB Biobank were each diluted to concentrations of approximately 50ng/µL as indicated in the Table below.

Table 3.1: Control sample stock DNA dilution and quantification

Sample Name	Sample ID	Concentration ng/µL	DNA Volume (µL)	TE Buffer Volume (µL)	Final Concentration ng/µL
XETOW	1A	138.155	10	17.63	51.4
XCHOF	1B	110.949	10	12.19	45.0
XAFOY	1C	139.452	10	17.85	43.9
XCCOA	1D	125.902	10	15.18	50.2
XINOA	1E	137.301	10	17.46	47.8
XIJOV	1F	115.765	10	13.15	47.6
XIPOB	1G	110.668	10	12.13	55.4
WZEOT	1H	199.383	10	29.88	49.0
XAROK	2A	149.158	10	19.83	51.2
XKXOM	2B	105.586	10	11.12	50.0

To observe DNA integrity 1µL DNA (50ng/µL) in 4µL fluorescent dye of each diluted DNA sample was run on a 0.8% (w/v) agarose gel in TBE. Although the resolution of the 1kb plus ladder was poor, it is still visible that the DNA samples each have a high molecular weight.

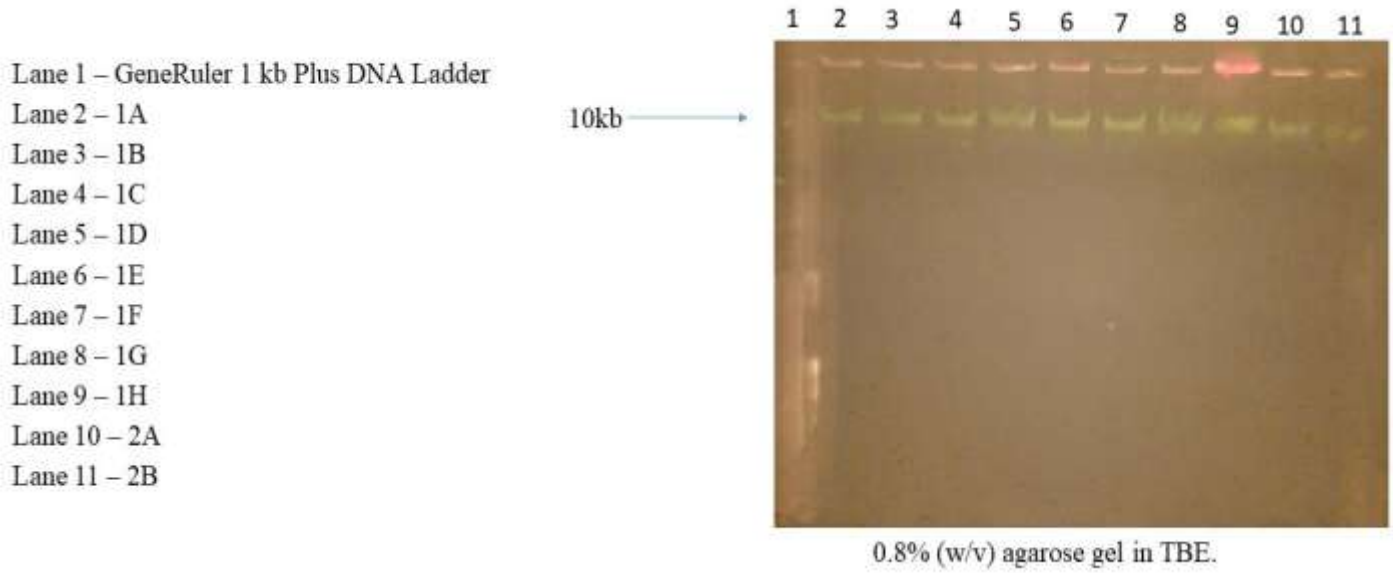


Figure 3.7: Agarose gel image of control sample DNA template stock solutions to confirm the integrity of the DNA in the control samples and ensure it has a high molecular weight, is intact and is not degraded.

3.6. Preparation of SCA7 positive samples

Unlike the SCA7 negative samples, the two positive patient samples, T1 and T2, were provided without any information about DNA concentrations. Quantification of DNA in 50-fold dilutions of each sample was performed using a Qubit4 analysis.

Patient sample T1 = **1 180 ng/µL Stock T1 DNA**

Patient sample T2 = **460.0 ng/µL = Stock T2 DNA**

Each of the patient samples T1 and T2 were diluted in calculated TE buffer volumes to make up 50ng/µL concentration for each of the patient samples, as presented in Table 3.2 below.

Table 3.2: Summary of details of SCA7 positive DNA samples

Sample Name	Sample ID	Concentration ng/µL	DNA Volume (µL)	TE Buffer Volume (µL)	Final Concentration ng/µL
SCA7	T1	460	5.43	44.57	50.0
SCA464- Male	T2	1180	2.12	47.88	45.8

To observe DNA integrity 1µL DNA (50ng/µL) in 4µL fluorescent dye of each diluted DNA sample was run on a 0.8% (w/v) agarose gel in TBE. The resolution of the 1kb Plus ladder is poor, but the largest visible band on the ladder is 10kb. It is visible that the sample T1 and T2 have high molecular weight DNA.

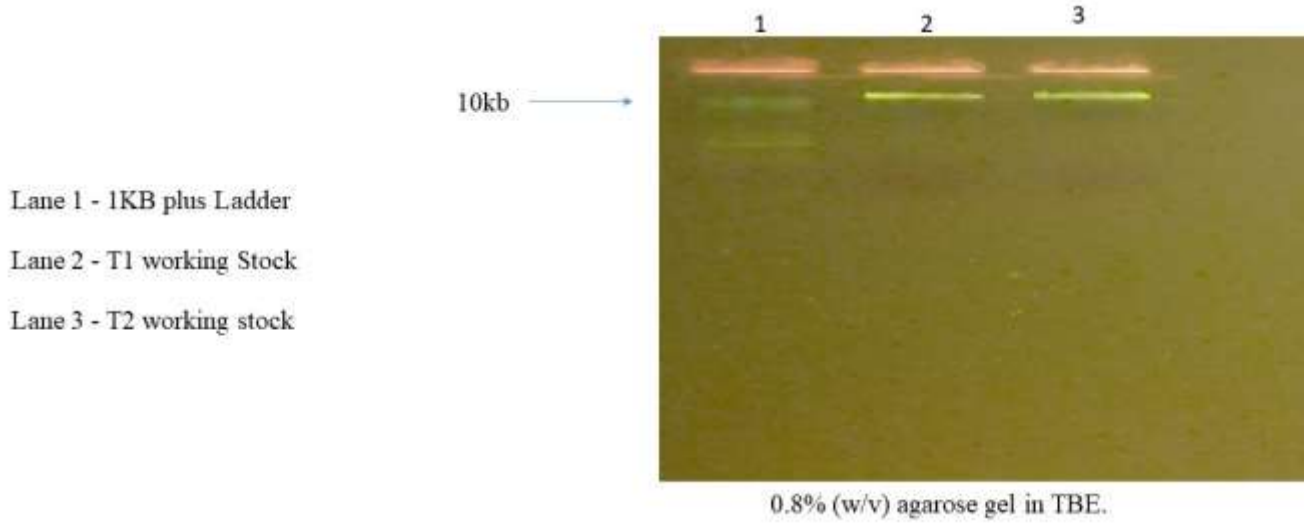


Figure 3.8: Agarose gel image of patient samples' DNA template stock solutions (T1 and T2), to confirm the integrity of the DNA, ensure it has a high molecular weight, is intact and is not degraded.

Both patient samples (T1 and T2) showed single bands in lane 2 and 3, this showed that the genomic DNA was intact and not degraded.

3.7. Patient and Control Sample PCR

The SCA7 negative control samples were amplified by PCR using the optimized *ATXN7* primer set as described earlier. This was performed in two batches.

The first batch of PCR runs was done on control samples labelled 1A, 1B, 1C, 1D and 1E. An aliquot of 5µL of each PCR product with 5µL tracking dye were run on a 2% (w/v) agarose gel in TBE.

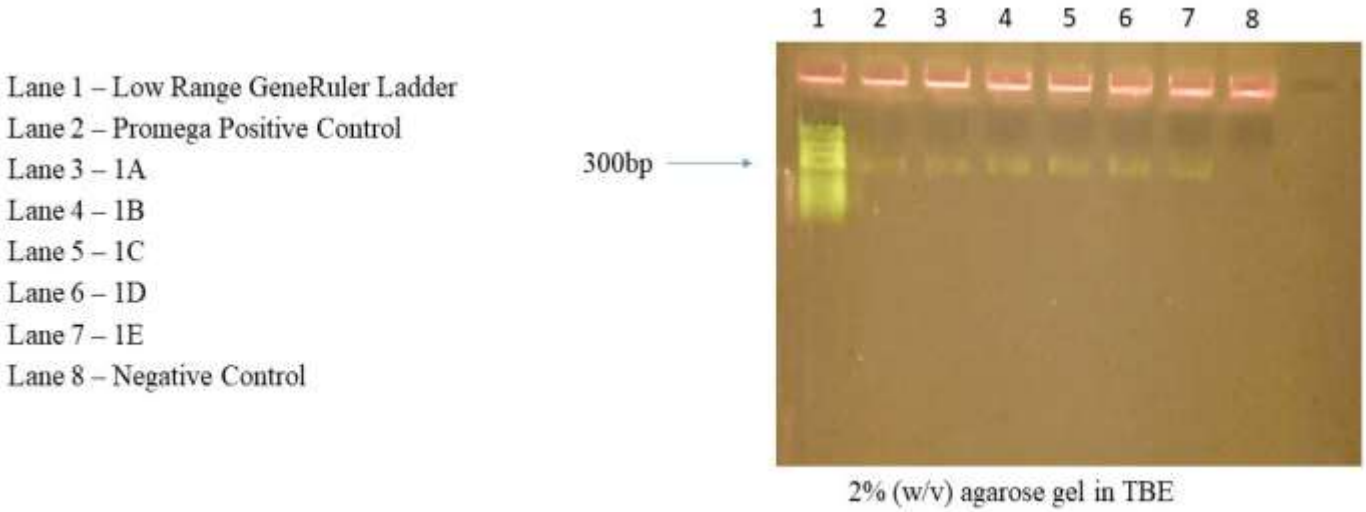


Figure 3.9: Agarose gel image of the *ATXN7* PCR products for the first batch of control samples. A fragment of ~300bp was present in 2, 3,4,5,6 and 7, as expected in a successful PCR amplification.

Bands were observed in the positive control and patient samples which was the expected size of approximately 300bp. Qubit quantification was done on the amplicons before cleaning. The Amplicons were cleaned using the ProNex protocol described earlier. Qubit4 fluorometer analysis of the purified DNA to determine concentration and yield.

Table 3.3: Quantification of DNA amplicons of the first batch of control samples, before and after amplicon purification

	Before purification				After purification			
Sample	**ng/µL**	**Total ng**	**Total fmol**		**ng/µL**	**Total ng**	**Total fmol**	**% Yield**
1A	3.54	70.9	360		2.64	50.16	250	69.4
1B	3.98	79.6	400		2.22	42.18	210	52.5
1C	3.80	76.0	380		1.76	33.44	170	44.7
1D	2.96	59.2	300		2.02	38.38	190	63.3
1E	3.50	70.0	350		1.94	36.86	190	54.3

The percentage yields were above 50% on all, except for sample 1C which had a 44.7% yield. This however is a concentration of 170fmol, which is above the minimum concentration required for Nanopore sequencing (range 100-200fmol). Each sample has 18µL of DNA after purification.

The second batch of the PCR runs was done for control samples labelled 1F, 1G, 1H, 2A and 2B and reaction products analysed by agarose gel electrophoresis in an identical way to the first batch.

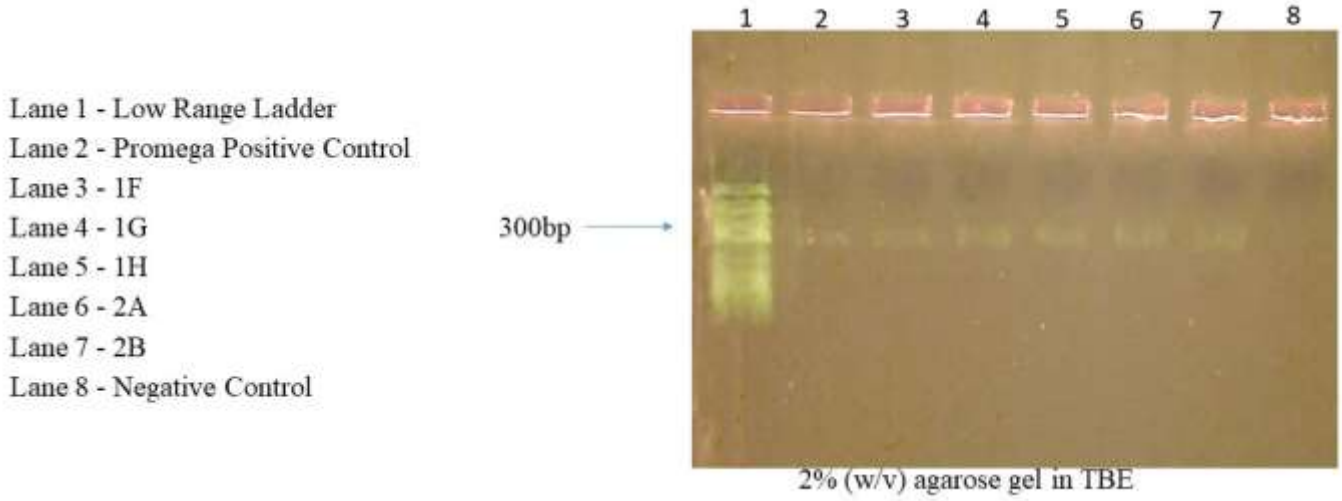


Figure 3.10: Agarose gel image of the *ATXN7* PCR products for the second batch of control samples. A fragment of ~300bp was present in 2, 3,4,5,6 and 7, as expected in a successful PCR amplification.

The percentage yields of DNA were lower in the second batch, and the amounts barely enough for Nanopore library preparation and sequencing.

Table 3.4: Quantification of DNA amplicons of the second batch of control samples, before and after amplicon purification

	Before purification				After purification			
Sample	**ng/µL**	**Total ng**	**Total fmol**		**ng/µL**	**Total ng**	**Total fmol**	**% Yield**
1F	4.68	93.6	470		1.22	24.4	120	25.5
1G	4.54	90.8	460		1.20	24.0	120	26.1
1H	5.04	100.8	510		1.27	25.4	130	25.4
2A	5.04	100.8	510		2.02	40.4	200	39.2
2B	4.82	96.4	490		1.84	36.8	190	38.8

The PCR amplification for the second batch was repeated to produce a larger amount of DNA. The repeat PCR was done using the optimized *ATXN7* PCR protocol, but volumes of the master mix were doubled to 50µL. The PCR products were run on a 2% (w/v) agarose gel in TBE buffer at the ratio of 5µL PCR product: 5µL tracking dye.

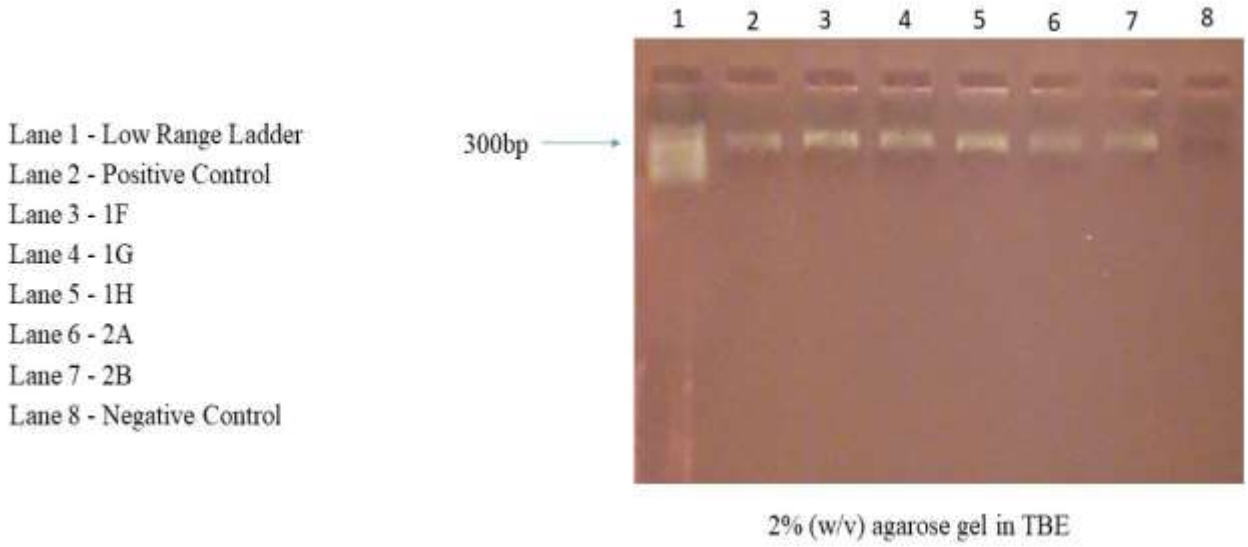


Figure 3.11: Agarose gel image of the *ATXN7* PCR products for the second batch of control samples. A fragment of ~300bp was present in 2, 3,4,5,6 and 7, as expected in a successful PCR amplification.

Bands were observed in the positive control and the batch of control samples 1F, 1G, 1H, 2A and 2B, which were the expected size. The Amplicons were purified using ProNex beads. Qubit analyses was used to determine yield and concentration.

Table 3.5: Quantification of purified *ATXN7* PCR amplicons for the repeated second batch of controls

Sample ID	Concentration (ng/µL)	Volume	Total ng	Concentration (fmol)
1F	2.88	39	112.32	605.8
1G	3.16	39	123.24	664.7
1H	3.10	39	120.9	652.0
2A	2.70	39	105.3	567.9
2B	3.10	39	120.9	652.0

The amounts of DNA recovered were much higher at a larger PCR volume, and adequate for several Nanopore sequencing experiments.

The patient samples T1 and T2 were amplified using the optimized *ATXN7* protocol, in a volume of 25µL. A PCR master mix was prepared for 5 PCR samples. The PCR products were run on a 2% (w/v) agarose gel in TBE buffer, at the ratio of 5µL PCR product: 5µL tracking dye.

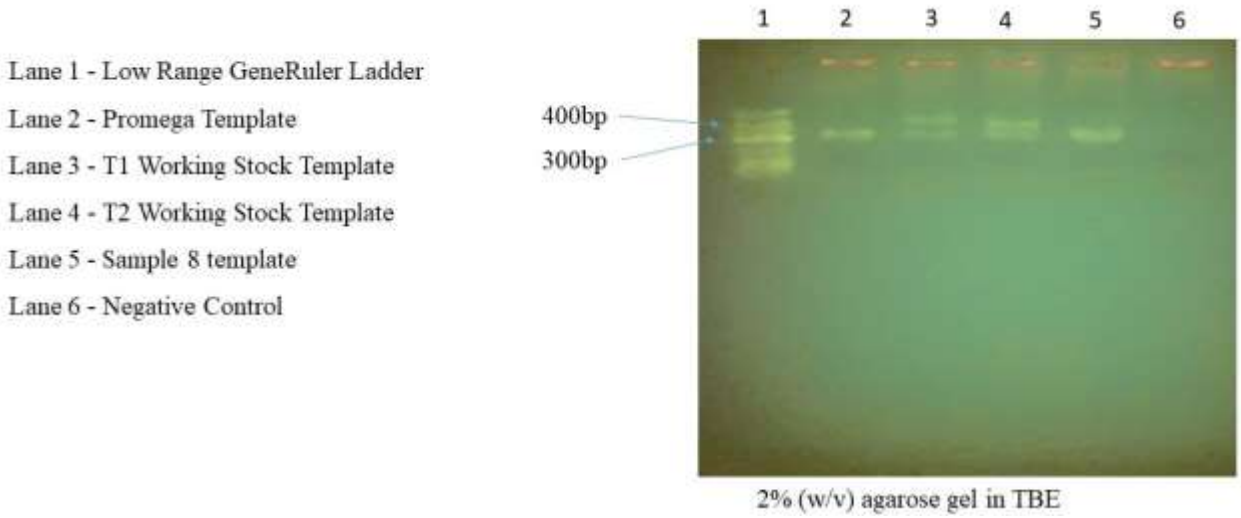


Figure 3.12: Agarose gel image of the *ATXN7* PCR products of the patient samples and positive controls. Single bands of ~300bp were observed in lanes 2 and 5 which contained the positive control sample amplicons. Two bands were observed in lanes 3 and 4, which possessed the patiend sample amplicons with one normal allele (~300bp) and one expanded allele

Lane 1 contains the DNA Ladder which gave the expected fragment size demarcations. Lane 2 contains the Promega positive control, with a single band at the expected size of approximately 300bp. Lane 3 contains patient sample T1 which has two bands below and above 300 bp. expected sizes of 271/424bp. The sizes of the 2 bands observed in lane 3 comply with the known sizes of sample T1. Lane 4 contains patient sample T2 which has expected sizes of 307/361bp. The sizes of the 2 bands observed in lane 4 comply with the known sizes of sample T2. Lane 5 contains the control sample 8 from the Coriell trio. The band observed in lane 5 complies with the expected band size of approximately 300bp.

The patient samples T1 and T2 amplicons were cleaned using the ProNex bead cleaning protocol. Qubit was done on the cleaned DNA to analyse the produced DNA concentration for sequencing. The DNA concentrations obtained after amplicon purification were:

T1 - 109fmol

T2 - 302fmol

After purifying all the 10 control samples and the 2 patient samples, they were all run on Invitrogen E- gel EX 2% agarose gels. These are highly sensitive gels with high resolution. A minimum of 10ng of DNA is required for running samples on the E-gel EX. The control and patient sample

DNA amplicons were each aliquoted into 10ng (54fmol) tubes and made up to 20µL with loading buffer.

The sample preparation and running on the E-gel x was done in 2 batches. The first batch contained the purified amplicons of the 5 control samples 1A, 1B, 1C, 1D, 1E, and the 2 patient samples T1 and T2. The required 54fmol of each of these amplicons was calculated and loaded onto the E-gel EX agarose gel. The gel was run for 20 mins.

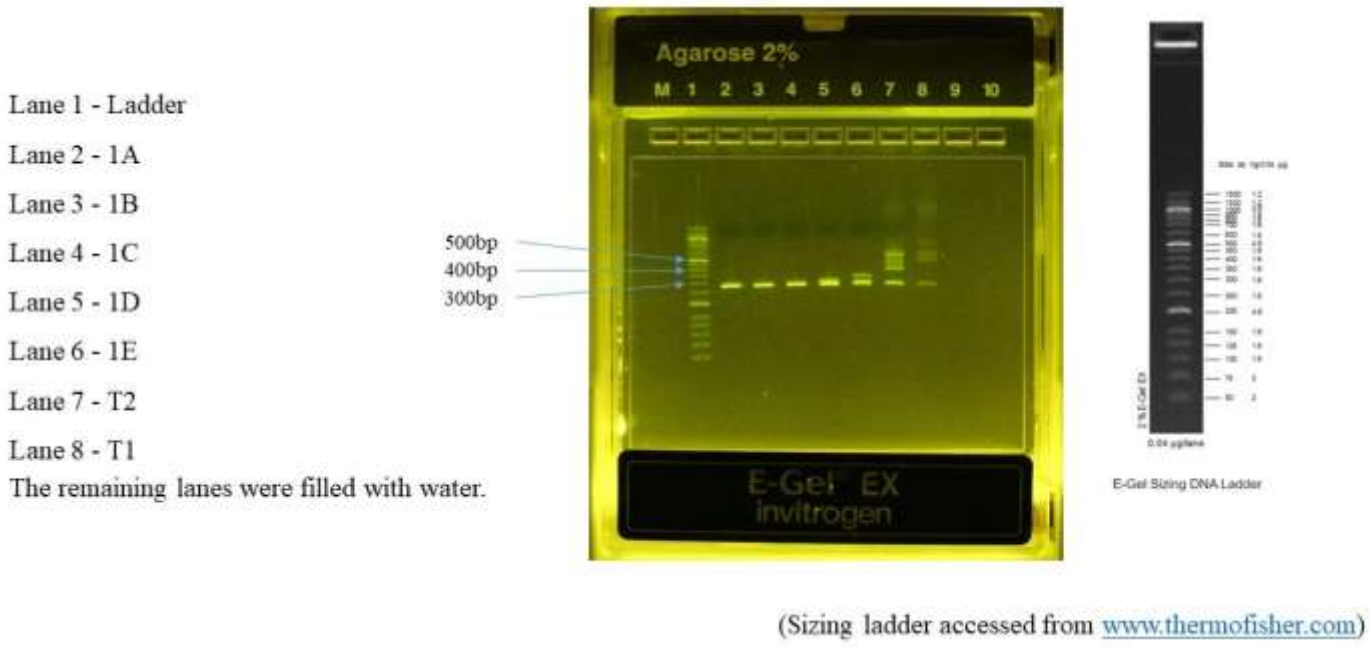


Figure 3.13: Agarose gel image of the *ATXN7* PCR products of the first batch of control samples and the patient samples. An image of the sizing ladder placed next to the gel, for scale

Lane 1 contains the E-Gel ladder which demarcates the fragment sizes and migration down the gel. Lane 2 contains control sample 1A, which shows a singular sharp band of the expected size which shows no triplet expansion. Lane 3 contains control sample 1B, which shows a singular sharp band of the expected size which shows no triplet expansion. Lane 4 contains control sample 1C, which shows a singular sharp band of the expected size which shows no triplet expansion. Lane 5 contains control sample 1D, which shows 2 bands, which suggest heterozygosity, but both bands within the size range which shows no triplet expansion. Lane 6 contains control sample 1E, which shows 2 bands, which suggest heterozygosity, but both bands within the size range which shows no triplet expansion. Lane 7 contains patient sample T2 which shows several bands with the smallest band in the lane was matches the 307bp prediction and a band matching the 361bp

prediction. There are an additional 2 bands observed, which are larger than the predicted size. These are not PCR artefacts, but were explained as mitotic instability, where at least one other larger band is observed.

Lane 8 contains patient sample T1 which shows several bands. The smallest band in the lane matches the 271bp expected band and there is another band matching the expected 424bp. Additional bands are also observed in the lane and are explained as a result mitotic instability.

The second batch to be run on a 2% Agarose E-Gel EX contained the purified amplicons of the 5 control samples 1F, 1G, 1H, 2A, 2B and the 2 patient samples T1 and T2. The required 54fmol of each of these amplicons was calculated and loaded onto the E-gel EX agarose gel. The gel was run for 20 minutes.

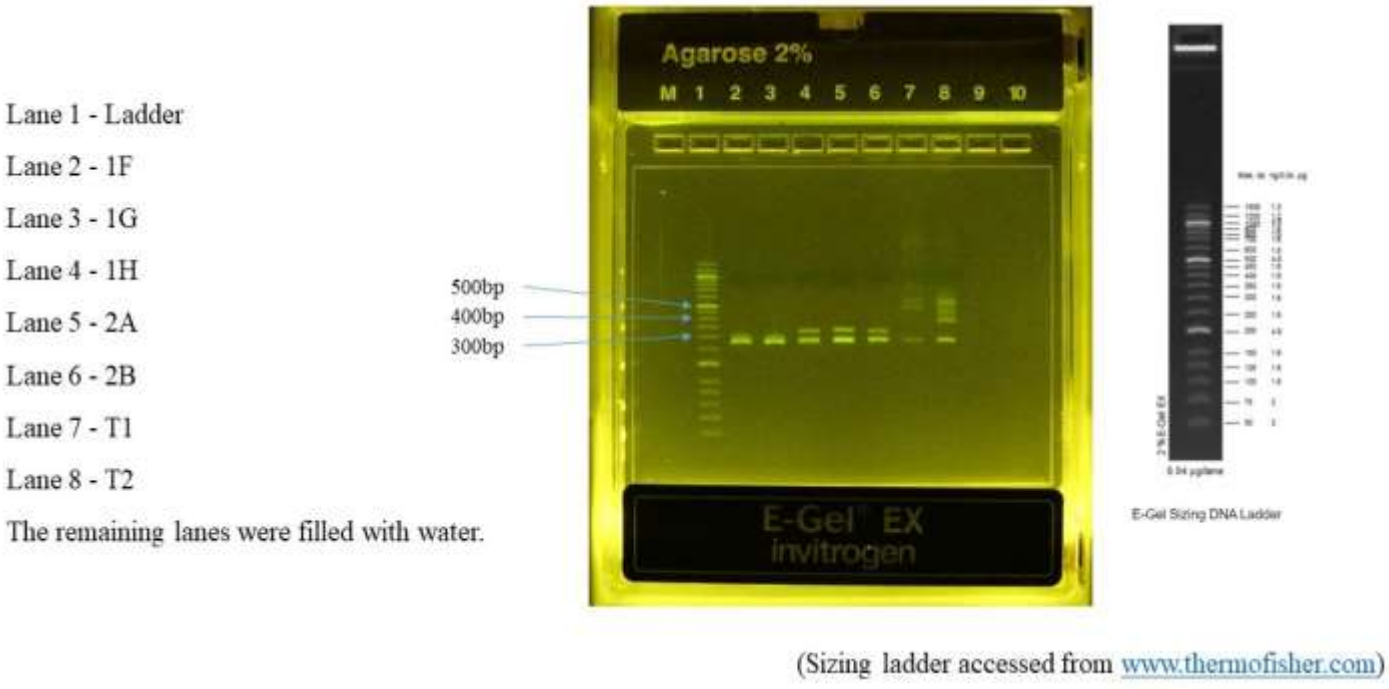


Figure 3.14: Agarose gel image of the *ATXN7* PCR products of the second batch of control samples and the patient samples. An image of the sizing ladder placed next to the gel, for scale

Lane 1 contains the E-Gel ladder which demarcates the fragment sizes and migration down the gel. Lane 2 contains control sample 1F and shows 2 bands that suggest heterozygosity, but both showing no evidence of triplet expansion. Lane 3 contains control sample 1G and shows 2 bands

that suggest heterozygosity, but both showing no evidence of triplet expansion. Lane 4 contains control sample 1H and shows heterozygosity. There are 2 bands with one band at the expected 280bp and the other band showing an intermediate allele of approximately 330bp, which is still below the threshold number of triplet repeats. The 2 heterozygous bands, with an intermediate allele, is also observed in lanes 5 and 6, which contain control sample 2A and 2B respectively. Lanes 7 and 8 contain the patient sample T1and T2 respectively and are showing the same sizes of bands as seen when they were run in the first batch, along with the first batch of controls. This shows repeatability and robustness of the findings.

3.8. Oxford Nanopore sequence data collection

Sequencing data files for the 10 control samples and the 2 patient samples were collected and the reports in the data files showed the number of passed reads (fastq_pass) with a minimum q-score of 7, and the run time for each sequence.

Table 3.6: Flow cell identities, number of fastq_pass reads and runtimes for the control and patient samples

Sample ID	Flow Cell ID	Run Time	Number of fastq_pass reads
1A (control)	AES204	17hrs 8mins	90 758
1B (control)	AEX897	1hr 50mins	56 675
1C (control)	AFL682	16hrs 35mins	31 258
1D (control)	AFF848	4hrs 1min	99 925
1E (control)	AFK822	3hrs 40mins	36 358
1F (control)	AFL626	2hrs 47mins	52 852
1G (control)	AFG111	1hr 27mins	101 076
1H (control)	AFK805	3hrs 5mins	91 761
2A (control)	AES427	1hr 48mins	107 488
2B (control)	AFG127	2hrs 51mins	171 954
T1 (patient)	AEY223	2hrs 16mins	114 760
T2 (patient)	AFI328	4hrs 30mins	27 689

The number of reads produced in each of the sequencing runs were not dependent on the runtime. The other variable was the number of active pores, but that too fluctuated as the sequencing run proceeded where the report in each sequence data file would show variations in number of active pores throughout the sequencing process. The functional assumption would be that the number of reads produced is a combined effect of the active pores and the quality of the library preparation done on each sample.

3.9. Basic Quality Control

Table 3.7 The EPI2ME desktop Agent basic quality control reports for the fastq_passed reads

Sample ID	Number of Reads	Successfully mapped reads	Reads mapped to Chr3	Average q-score	Average length	Percentage accuracy (%)
1A	90 758	85 686	85 669	9.97	332bp	91.75
1B	56 675	54 304	54 304	9.88	328bp	91.16
1C	31 258	26 611	26 561	8.72	349bp	87.70
1D	99 925	93 042	93 009	9.15	347bp	90.80
1E	36 358	33 056	33 043	8.98	337bp	89.33
1F	52 852	47 775	47 756	9.73	344bp	90.75
1G	101 076	97 039	97 003	9.76	322bp	90.13
1H	91 761	82 705	82 698	9.41	328bp	90.13
2A	107 488	99 761	99 726	9.60	323bp	90.72
2B	171 954	163 969	163 927	9.48	331bp	89.35
T1	114 760	102 086	98 683	9.82	406bp	90.73
T2	27 689	20 813	20 339	9.05	384bp	88.69

The average q scores are high, with the highest scores observed in samples that had a higher sequencing depth. The higher Q score average samples, also had the higher percentage accuracies. This would support the notion that sequencing accuracy increases as the sequencing runtime increases.

To obtain the highest quality of reads, a filtration run was done on each of the reads to trim them by size and by q-score. The samples were filters by size, from a minimum of 250bd to a maximum of 350bp. This was done to all the control sample, but the patient samples were filtered to a

minimum of 250bp and a maximum of 500bp. The reads of all the control and patient samples were filtered to a minimum of q-score 12. These parameters were set on the nanofilt tool, using the EPI2ME labs server command line.

Table 3.8 The EPI2ME desktop Agent basic quality control reports for filtered fastq passed reads

Filtered Sample ID	Reads Analysed	Reads Aligned	Aligned to Chr3	Average Size	Average q-score	Percentage Alignment to Chr3
1A	6768	6764	6764	297	12.69	100%
1B	3171	3171	3171	293	12.53	100%
1C	90	90	90	301	12.59	100%
1D	1270	1266	1264	301	12.59	99.8%
1E	418	418	418	294	12.65	100%
1F	3191	3186	3186	308	12.65	100%
1G	3636	3630	3628	299	12.57	99.9%
1H	1791	1784	1784	298	12.51	100%
2A	3628	3628	3627	302	12.58	99.9%
2B	3665	3661	3659	309	12.57	99.9%
T1	6901	6820	6720	332	12.70	97.4%
T2	167	152	143	339	12.48	94.1%

The change in minimum q-score filter, from 7 to 12, greatly decreased the number of passed reads for each sample. The filtered reads mostly aligned 100% to Chr 3. As this is the location of the

SCA7 repeat region, a further quality control step would be to ensure that the reads are aligning specifically to the SCA7 region.

The filtered reads for each of the samples were sorted into SAM files, then the SAM files were compressed into BAM files and the BAM files were indexed. The indexed Bam files for each of the samples aligned to the GRCh38 reference sequence, using the Integrative Genomics Viewer (IGV) tool. Filtered reads from all the samples all aligned to the SCA7 repeat sequence region. The CAG repeats in the SCA& region of the reference sequence GRCh38, are located between the coordinates 53128 and 53 158. This is a stretch of 10 CAG repeats. Since the number of CAG repeats are different in each sample, the alignment in IGV shows the presence of indels at position 53128 or position 53158. The indels show an insertion where a sample has a more triplet repeats that the reference, and a deletion where the sample has less CAG repeats than the reference. Figure 3.15 shows a representation of the IGV interface and the presence of a consistent tower of indels in the filtered reads. This both confirms the alignment of the reads to the SCA7 region, and the presence of indels in the CAG repeat region of the filtered reads.

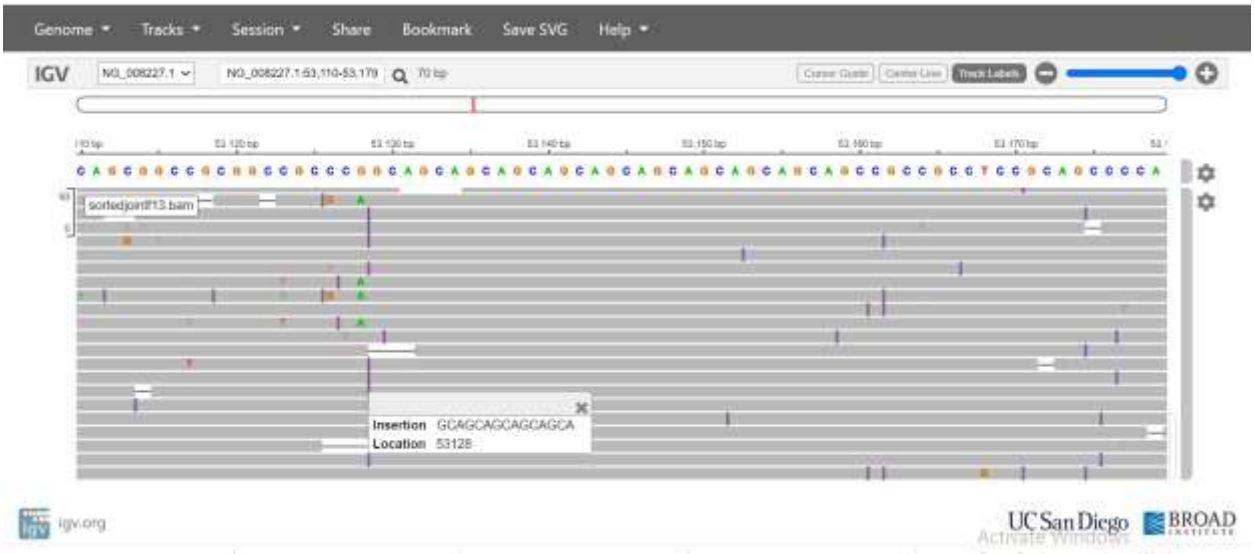


Figure 3.15: Visualization of sorted Control 1A reads mapped against the GRCh38 reference sequence in the IGV tool, showing an insertion location and identity.

3.10. Data analysis

For each of the control samples, the read length distribution was represented on a scatterplot. For each sample, there are 2 plots. The plot with isolated reads represents the length of reads that have a conserved sequence following the triplet repeat tract. Next to the isolated read plot is a trimmed read plot, which shows the length of reads after truncating the sequences that come before the triplet repeat tract.

Control – 1A and 1B

Isolated read counts and trimmed read counts for samples 1A and 1B were plotted on scatter plots and presented in figure 3.16a.

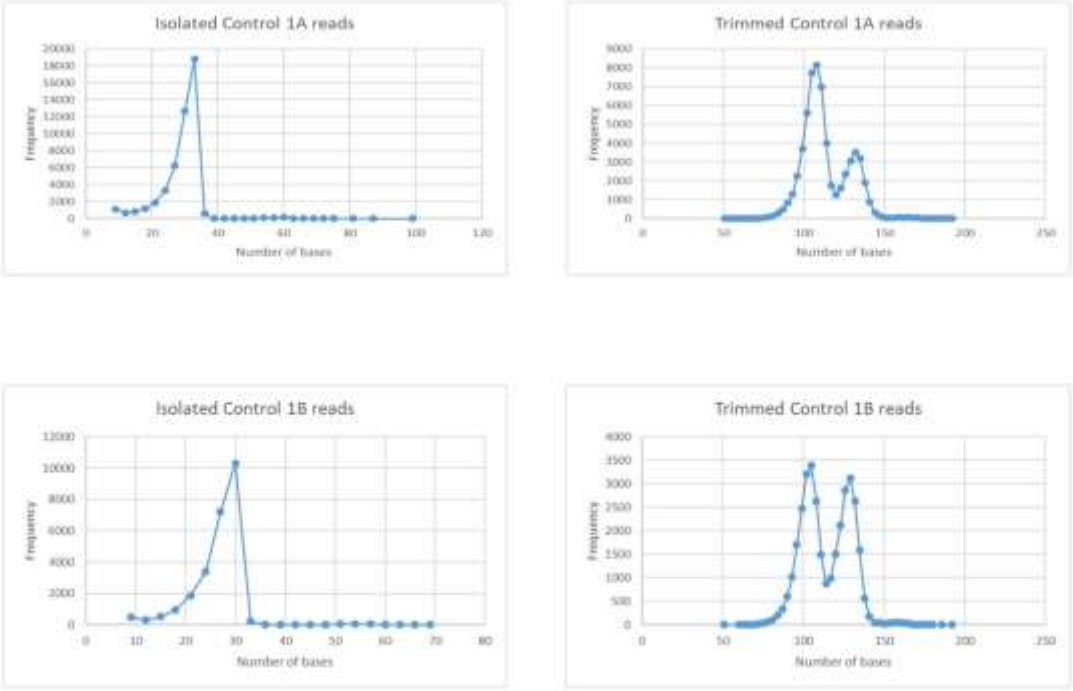


Figure 3.16a: Scatter plot of isolated read counts and trimmed read counts for sample 1A and 1B, showing the fragment sizes and frequency of the sizes within the reads' data files.

For control 1A, the isolated read peak represents the highest frequency of reads at 33 bases, including the CCG flank. The number of bases in the conserved triplet repeat tract is therefore 30bases = 10 CAG repeats. This gives a sizing standard that states the number of repeats in the

first peak of the trimmed reads. The first peak in the control 1A trimmed read scatter plot is 108 bases, and the second peak is 132 bases. A difference of 24 bases is 8 CAGs. Allele sizes for control 1A are therefore heterozygous - 10/18. The same concept was used to calculate the lengths of all the control samples and the patient samples.

Control - 1C and 1D

Isolated read counts and trimmed read counts for samples 1C and 1D, were plotted on scatter plots and presented in figure 3.16b.

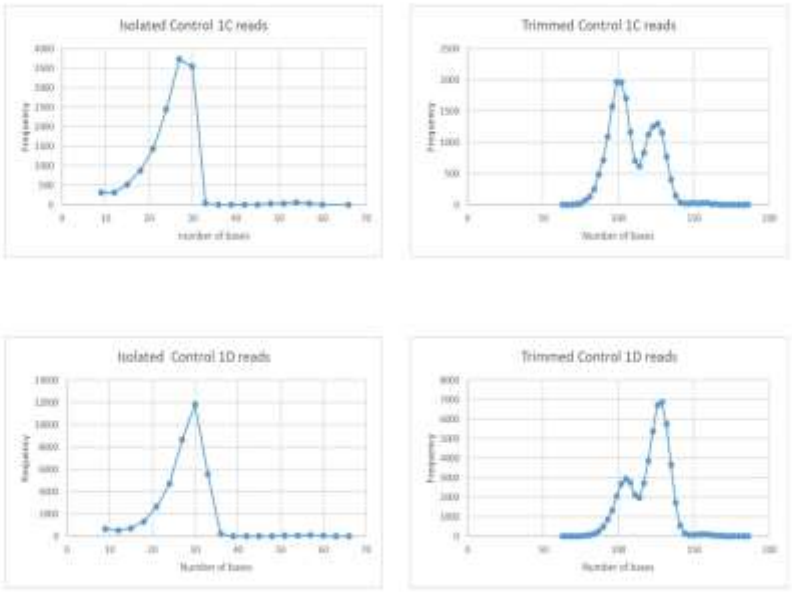


Figure 3.16b: Scatter plot of isolated read counts and trimmed read counts for sample 1C and 1D, showing the fragment sizes and frequency of the sizes within the reads' data files.

Control - 1E and 1F

Isolated read counts and trimmed read counts for samples 1E and 1F were plotted on scatter plots and presented in figure 3.16c.

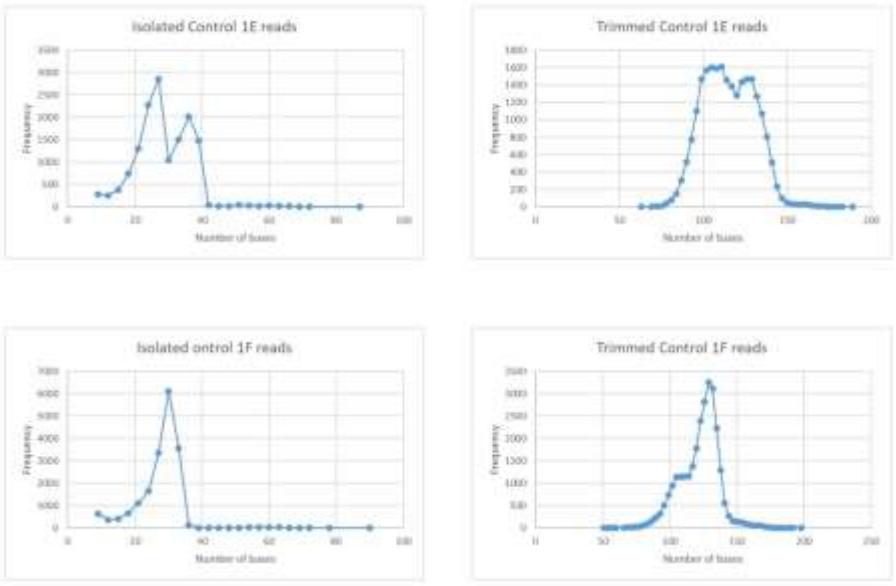


Figure 3.16c: Scatter plot of isolated read counts and trimmed read counts for sample 1E and 1F, showing the fragment sizes and frequency of the sizes within the reads' data files.

Control – 1G and 1H

Isolated read counts and trimmed read counts for samples 1G and 1H were plotted on scatter plots and presented in figure 3.16d.

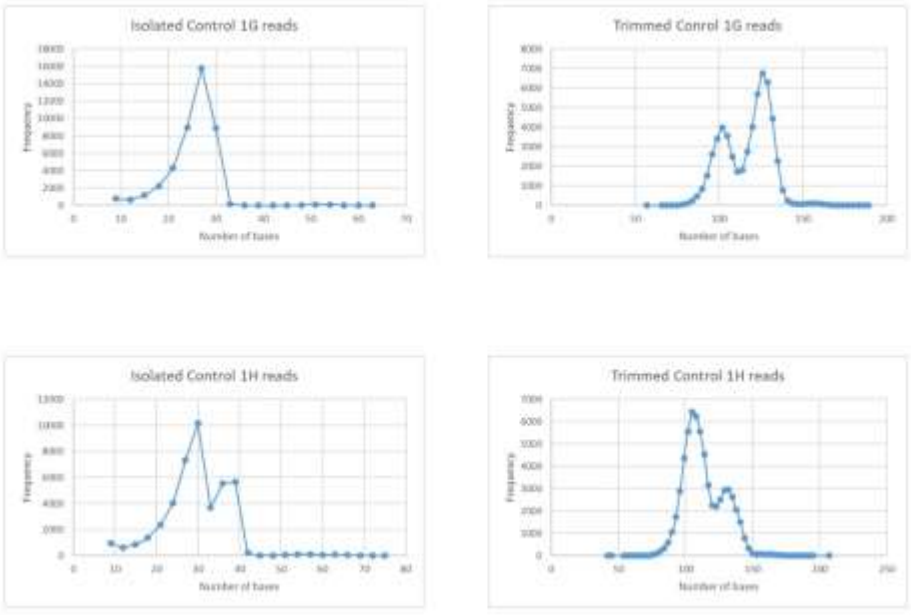


Figure 3.16d: Scatter plot of isolated read counts and trimmed read counts for sample 1G and 1H, showing the fragment sizes and frequency of the sizes within the reads' data files.

Control - 2A and 2B

Isolated read counts and trimmed read counts for control 2A and 2B were plotted on scatter plots and presented in figure 3.16e.

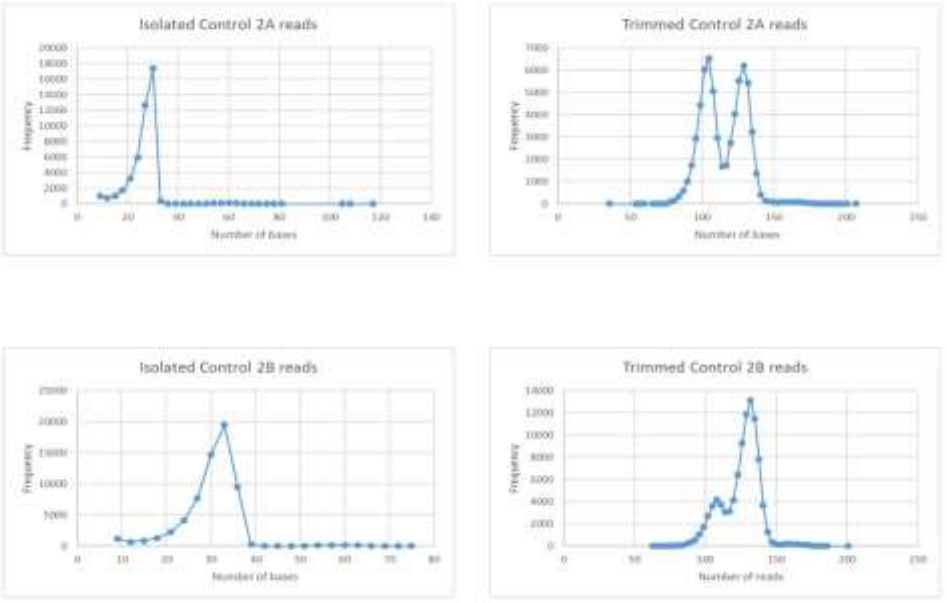


Figure 3.16e: Scatter plot of isolated read counts and trimmed read counts for sample 2A and 2B, showing the fragment sizes and frequency of the sizes within the reads' data files.

Observed allele sizes for the control samples were tabulated and presented in Table 3.9. The calculation method stated above showed heterogeneity in all the control samples used in this study.

Table 3.9. The number of CAG repeats in triplet repeat tracts of the two alleles in each of the control and patient samples.

Sample ID	Number of CAGs
1A	10/18
1B	9/17
1C	8/16
1D	9/17
1E	8/16
1F	9/17
1G	8/16
1H	9/18
2A	9/17
2B	10/18

Patient Samples T1 and T2

Isolated read counts and trimmed read counts for Patient samples T1 and T2 were plotted on scatter plots and presented in figure 3.16f.

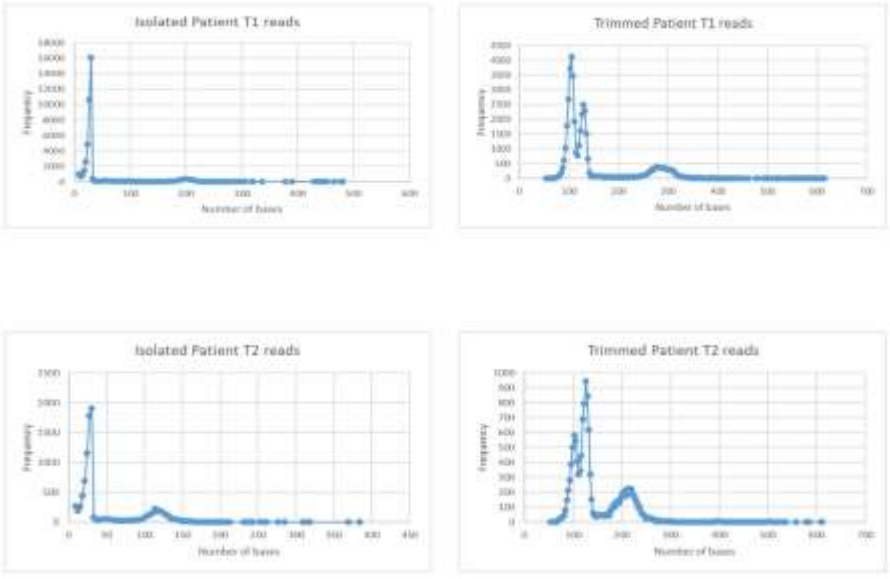


Figure 3.16f: Scatter plot of isolated read counts and trimmed read counts for sample T1 and T2, showing the fragment sizes and frequency of the sizes within the reads' data files.

For patient sample T1, the isolated read plot has a single peak representing the highest frequency of reads at 30 bases, including the CCG flank. The number of bases in the conserved triplet repeat tract is therefore: 27 bases = 9 CAG repeats. This gives a sizing standard that states the number of repeats in the first peak of the trimmed read counts. The plot for the patient sample T1's trimmed read counts presented 3 peaks.

The first peak is 105 bases long = 27 bases in the triplet repeat tract = **9CAG repeats**

The second peak is 129 bases = 24 bases more than the first peak = 8CAGs more = **17CAG repeats**

The third peak is 279 bases = 58 bases more than the first peak = 58CAGs more = **67 CAG repeats**

Taking the highest unexpanded allele read peak and the longest allele read peak. Allele sizes are therefore heterozygous – 9/67.

For patient sample T2, the isolated reads plot showed 1 peak, representing the highest frequency of reads at 30 bases, including the CCG flank. The number of bases in the conserved triplet repeat tract is therefore: 27 bases = 9 CAG repeats. This gives a sizing standard for the number of repeats in the first peak of the trimmed read counts. The plot for the patient sample T2's trimmed read counts presented 3 peaks.

The first peak is 102 bases = 27 bases in the triplet repeat tract = **9CAG repeats**

The second and highest peak is 126 bases = 24 bases more than the first peak = 8CAGs more = **17CAG repeats**

The third peak is 213 bases = 111 bases more than the first peak = 37 CAGs more = **46 CAG repeats**

Taking the highest unexpanded allele read peak and the longest allele read peak. Allele sizes are therefore heterozygous – 17/46.

The calculated allele sizes for the patient samples T1 and T2 were concluded as:

T1 – 9/67

T2 – 17/46

3.11. Duration assessment of currently used methods and the Oxford Nanopore approach

In this assessment, we analysed the duration it takes to complete each step in the different analysis methods being compared in this project. The currently used method, CE and the Nanopore sequencing method, which was undertaken in this project. The importance of this duration assessment is to provide context for the costing comparison of the 2 methods, as it facilitates the calculation of the salary of the scientist who performs the analysis. The time taken in each step also provides the time taken while each machine is in use. This is a vital part of method costing. The duration for each step of the 2 methods, is stated for an analysis of 1 sample and for analysis of 10 samples. The cost assessments were split into hands-on duration and machine-use duration. CE is performed on 10 samples at a time, while Nanopore sequencing in this project was done on 1 sample at a time. The Nanopore duration analysis for both hands-on and machine use was tabulated for the duration of analysing 1 sample and analysing 10 samples. Some steps have the same duration for 1 and 10 samples, as several samples are run at once.

Table 3.11. CE steps and hands-on duration, as well as the machine-use duration of each step

CE step	Hands-on Duration	Machine-use Duration
PCR - master mix preparation and thermocycler run time	30 mins	2hrs 30mins
Gel electrophoresis	10 mins	20 mins
PCR product purification	40mins	40 mins
CE sample preparation and run time	15 mins	30 mins
Data collection and analysis	30 mins	30 mins
TOTAL	2hours 5mins	4hours 30mins

Table 3.12. Nanopore sequencing steps and durations (hands-on and machine-use) of each step for analysis of 1 sample and of 10 samples

Nanopore sequencing step	Hands-on Duration		Machine-use Duration	
	1 Sample	**10 Samples**	**1 Sample**	**10 samples**
PCR - master mix preparation and thermocycler run time	30 mins	30 mins	2hrs 30mins	2hrs 30mins
Gel Electrophoresis	10 mins	10 mins	20 mins	20 mins
PCR product purification	40mins	40mins	40mins	40mins
Nanopore library preparation	2 hours	2 hours	2 hours	2 hours
Nanopore sequencing run time	10 mins	1 hr 40mins	2hrs 30mins	25 hours
Data collection and analysis	45 mins	7hrs 30mins	45 mins	7hrs 30mins
TOTAL	4hrs 15mins	12hrs 30mins	8hrs 45mins	38 hours

3.12. Cost Comparison

A cost assessment was done for the currently used CE method and the Nanopore sequencing method, which was attempted in this project. The reagents used in each step of the respective analysis methods were quantified and costed, relative to analysis of 10 samples. The machinery used in each analysis was costed by taking the retail price and dividing it by 5years, which is the stipulated machine life span. Dividing the machine cost price by the number of hours in 5 years gives a cost of using the machine per hour. The duration assessment of each analysis method provided context on how long each of the machines was used for the assessment of 10 samples. The duration of use of each device was used to determine the machinery cost of each assessment method.

The costs of some of the reagents and machinery was priced in US Dollars. The exchange rate constantly fluctuates, an average over the past year was extracted from the online currency exchange platform called Exchange rates.org (https://www.exchangerates.org.uk/USD-ZAR-exchange-rate-history.html) and it states that the average exchange rate over the past year is 1

USD = 14.50 Rand. This is the constant exchange rate that was used in this costing analysis for the Nanopore sequencing device and the reagents required to undertake the Nanopore sequencing analysis.

Table 3.13. A representation of the base costs in Rand (ZAR) for SCA7 analysis using CE in comparison to using Nanopore sequencing

Oxford Nanopore					**Capillary Electrophoresis (CE)**				
PCR					**PCR**				
ITEM	**Unit**		**Per 10 Samples**		**ITEM**	**Unit**		**Per 10 Samples**	
	Qty	**Cost**	**Qty**	**Cost**		**Qty**	**Cost**	**Qty**	**Cost**
Primers sets	300µL	523.57	20µL	34.90	Fluorescent Primer sets	300µL	682.92	20µL	45.53
TaKaRa Prime Star GXL Polymerase*	250µL	4411.40	5µL	88.22	TaKaRa Prime Star GLX Polymerase	250µL	4411.40	5µL	88.22
Betaine 5M	1500µL	354.82	50µL	11.83	Betaine 5M	1500 µL	354.82	50µL	37.50
DMSO	5000µL	541.91	10µL	1.08	DMSO	5000 µL	541.91	10µL	1.08
Nuclease Free water	5000µL	330	75µL	4.95	Nuclease Free water	5000 µL	330	75µL	4.95
Loading dye	1000µL	416	10µL	4.16	Loading dye	1000 µL	416	10µL	4.16
DNA Ladder	500µL	1685	5µL	16.85	DNA Ladder	500µL	1685	5µL	16.85
Agarose Tablets (0.5g each)	50 Tablets	2336	1 pellet	46.72	Agarose Pellets	50 Pellets	2336	1 Pellet	46.72
PCR Cleaning	1	28.00	10	280.00	PCR Cleaning	1	28.00	10	280.00
Total				**488.16**					**498.79**
Nanopore Sequencing					**Capillary Electrophoresis Sizing**				

ITEM	Unit		Run time cost (10 samples)		ITEM	Unit		Run time cost (10 samples)	
	Qty	Cost	Qty	Cost		Qty	Cost	Qty	Cost
Oxford Nanopore MinION	1 (5 years - 0.33/hr)	14500	25hrs	8.27	3500 Genetic Analyzer for Fragment Analysis	1 (5 years - 2.05/ hr)	89784.77	1hr	2.05
Library Prep Kit	6	8985	10	14 975	DS-30 Matrix Standard Kit (8 runs)	8 runs	3 125.00	1 run	390.63
NEB Companion Library Module	24 reactions	16474	10 reactions	6 864.16	GeneScan 500 ROX Dye Size Standard (800 sample runs)	800 runs	9 190.00	1 run	11.50
SPRI Beads	5000µL	1435.50	1000 µL	287.10	Hi-Di formamide	5ml	370.00	2.5µL	0.19
Flongle Flow Cell	1	1350	10	13500					
Control DNA expansion Kit (6rxns)	6	900	10	1500					
TOTAL			**37 134.53**		**TOTAL**				**404.37**

* This includes all the PCR reagents (buffer and dNTPs)

Labour

Tests are performed by a Scientist:

Salary of Scientist/month R25 000.00

Salary of Scientist/Hour R156.25

Capillary Electrophoresis

Hands-on duration (10 samples): 2hours 5 mins

Salary cost per 10 tests: R312.52

Nanopore Sequencing

Hands-on duration (10 samples): 12hours 30 mins

Salary cost per 10 tests: R1953.13

All patients done are reviewed by a Pathologist:

Salary of Pathologist /month R120 000.00

Salary of Pathologist/ Hour R750.00

Salary cost/ test (15mins) R187.50

Same amount for both analysis methods.

Overheads and Total

Capillary Electrophoresis:

Analysis base costs = PCR costs + CE sizing costs + Scientist salary + Senior Scientist salary

= 498.79 + 404.37 + 312.52 + 187.50

= 1 403.18

+ Wastage (15%) = 1 613.63

+ Laboratory fees (20%) = 1 936.36

+ Vat (15%) = 2 226.81

Total R 2 226.81 (10 samples)

Nanopore Sequencing:

Analysis base costs = PCR costs+ Nanopore sequencing + Scientist salary + Senior Scientist salary

= 488.16 + 37 134.53 + 1953.13 + 187.50

= 39 763.32

+Wastage (10%) = 43 739.65

+Laboratory fees (20%) = 52 487.58

+Vat (15%) = 60 360.72

Total R 60 360.72 (10 samples)

4. Discussion

A prevalence study was done in 2011 for all the SCAs in South Africa and the study was conducted on individuals presenting with SCA symptoms with individuals from 165 families. From the 165 families, 50 families presented with SCA7 and out of those 50 families, 49 were black South Africans (Smith *et al.*, 2012). This study showed the higher occurrence of SCA7 in black Africans in South Africa and presented the need for further study and development of assays that are more appropriate to African populations. Africans have more diverse genetic variations in both coding and non-coding regions of the genome, where the use of diagnostic methods developed for non-African populations could result in African individuals being mistakenly considered to have disease risk variants (Huang, Shu and Cai, 2015). Some variants that are considered to be disease risk variants in European populations, are actually common non-pathogenic variants in African populations (Mulder, 2017).

The currently used method of SCA7 diagnosis is CE, which involves a PCR amplification with fluorescent primers, followed by a size analysis using a genetic analyzer. Given the high occurrence of DNA sequence variations that are specific to African populations, the published primers were assessed through a series of bioinformatics analysis tools and variations were observed in the primer binding sites. *In silico* analysis of the published primers, using bioinformatics tools, showed they were unable to capture some of the DNA sequences in the 1000 Genomes database of African genetic data. In this project, a set of primers was designed, ensuring there are no variations in the primer binding sites and that the *ATXN7* gene could be amplified in all African DNA samples during a PCR run.

The currently used SCA7 diagnostic method, CE, characterizes the size of the isolated SCA7 triplet repeat tract fragments to determine the number of CAG repeats present. This analysis does not assess the sequence of this fragment, in order to identify the presence of any sequence variations in the triplet expansion. Although the CE method is effective in providing the length of the triplet repeat tract in SCA7, false negatives can occur if the triplet expansion is very long and goes beyond the spectrum of the standards being used in the CE analysis. A result showing apparent homozygosity raises the suspicion of a false negative and a confirmatory test is required in the form of a TP PCR.

This project explores an alternative diagnostic method using ONT. Long read sequencing reads the entire length of the amplified SCA7 triplet expansion fragments, however long the fragment may be. Sequencing also allows for analysis of sequence variations within the SCA7 triplet repeat tract. Optimization of this alternative diagnostic method, could provide a sole diagnostic method that provides more analysis information than fragment sizing alone and does not require any confirmation tests. This project also assessed the cost implications of Oxford Nanopore sequencing, in comparison to the currently used CE method.

4.1. PCR Optimization

The primer design process was specifically done to cater for the African samples that were used in this project. The coordinates for the primer binding sites of the designed primer sets were analysed in genome sequence data from the 1000 Genomes Project. These data sets all showed no variations in the designed primer binding sites and provided enough confidence in the effectiveness of the primers. Virtual PCRs done using the UCSC genome browser, showed that the primers were specific for the *ATXN7* gene in chromosome 3.

The PCR optimization step was a key component in this project. There were four laboratory controls used for the optimization step. These controls were the Coriell trio NA12891 (father), NA12892 (mother), NA12878 (daughter) and the Promega control. The first *ATXN7* PCR trial was adapted (Smith *et al.*, 2012) and the agarose gel image from a 2% agarose gel electrophoresis, showed the amplification of the expected region at annealing temperatures 57ºC and 59ºC. The gel image also showed bright bands that are larger than 1kb, at these optimum temperatures. Suggesting non-specific binding and the need for further optimization.

The second ATXN7 PCR optimization trial was adapted from (Cagnoli *et al.*, 2018). This method made use of a combination of DMSO and betaine as additional components to the mastermix. The *ATXN7* PCR protocol by (Cagnoli *et al.*, 2018), made use of the Go-Taq Long polymerase. The second optimization trial in this project also made use of the same polymerase, in order to fuly replicate the conditions. After a successful *ATXN7* amplification, further alterations were made to the protocol in order to increased specificity to the GC rich triplet repeat tract of SCA7. A comparison of polymerase enzymes most ideal for amplification of long range, G-C rich fragments concluded that the most ideal polymerase to use is the Prime Star GXL TaKaRa polymerase (Jia *et al.*, 2014). The Prime star GLX TaKaRa has been used to amplify large fragments of GC rich

DNA sequences with minimal PCR stutter. The Prime Star GXL TaKaRa was used in the third and final PCR optimization trial in this project.

The ProNex size selective chemistry purification kit was used to purify the PCR amplicons. The percentage yield for each of the amplicons was calculated by assessing the concentration of DNA in each of the PCR products before purification, using a Qubit4 fluorometer, before purification and after purification. The ProNex size selective chemistry purification kit gave DNA yield percentages higher than 50%, making it an optimal amplicon purification method.

4.2. Sanger sequencing approach

The purified amplicons of the 4 laboratory controls were sent to Inqaba Biotech for Sanger sequencing. The purpose of this step was to confirm that the correct genomic region had been amplified by the optimized PCR protocol, as it could not be strictly based on amplicon size alone. Sanger sequencing results were provided for both the forward and reverse sequences for the laboratory controls used. These results did not only confirm the identity of the amplicons, but also showed that it was difficult to assess the exact number of triplet repeats using this approach. It was, however, adequate to provide assurance the correct genomic region had been amplified.

The Sanger sequencing results for both forward and reverse sequences for each laboratory control sample were mapped onto the GRCh38 reference sequence and in a multiple sequence alignment using the SnapGene software. The chromatogram representations of the forward and reverse sequences were manually visualized and examined. Both the forward and reverse sequences were accurately mapped onto the SCA7 triplet repeat region of the GRCh38 reference sequence, but the reverse sequence reads mapped more accurately, with fewer uncertain peaks on the mapped read chromatogram. The reverse reads seem to be much more accurate in correctly assigning bases in the sequencing process.

The Sanger sequencing approach managed to give a good estimate of the CAG repeat sequence size, but it might work for smaller lengths but not for longer DNA fragments with large CAG repeats. SCA7 is characterised by triplet repeats between 36 and 460 CAGs. It would not be ideal to use Sanger sequencing as a sole diagnostic method for sequencing SCA7 due to the long fragment lengths that positive SCA7 patient DNA can reach. Sanger sequencing is characterised

by 99.9% accuracy for up to approximately 900bp, as reviewed by (McCombie, McPherson and Mardis, 2019). The range of allele sizes expected in early onset positive SCA7 cases is beyond the size range of Sanger sequencing accuracy. Although the majority of adult onset SCA7 cases have fewer than 60 CAG repeats, the objective of the project was to design a protocol that could be used for the identification of the entire range of possible expansion sizes. As a proof of concept, this project aimed to develop a protocol that could be used to assess other TRDs that may have even larger expansions and larger fragment sizes. For this reason, the project proposes a method that would not be limited by what is possible when using Sanger sequencing.

4.3. Oxford Nanopore MinION Sequencing

The main aim of the study was to develop a protocol for the diagnosis of SCA7 through nanopore sequencing. To do this the optimized *ATXN7* PCR protocol was performed on the control and patient sample DNA. After amplification, the amplicons were purified and the library preparation and Oxford Nanopore sequencing procedure performed. It took approximately 2 hours to complete the library preparation and to be ready for sequencing. A few changes were made to the documented SQK-LSK109 protocol and these were outlined in the methods section. Given the portability of the Nanopore MinION sequencing device, one would consider how practical it is to do library preparation and sequencing in the field. The only drawback one would have is how every step is dependent on the availability of ice. Apart from this aspect, it is very practical to use lithium batteries to power the heating block and to perform the library preparation and sequencing, while in the field.

The most challenging and critical part of the sequencing procedure is loading the library onto the flow cell. One should ensure that there are no air bubbles created when loading and that enough pores are ready to sequence, before and after the loading of the library. The flow cell check, which is done before loading the library, will quantify the number of available pores in the flow cell in use. The careful and fast loading of the library onto the flow cell, will ensure that air bubbles do not form, which cause pores to become inactive during the loading process. The number of available active pores when sequencing begins, will determine the number of passed sequences produced, and the speed of production of these reads.

In this project, the number of pores that were active for each sequencing run varied and this was well represented by the varied amounts of time it took each sequencing run to produce 100 000 passed reads. Sequencing runs that had more active pores at the beginning of sequencing required a much shorter sequencing run time to produce 100 000 passed reads. The Flongle flow cell was used in this project and has 126 pores. The longer the flow cells are kept before use, the fewer the number of active pores that are available for the run. The manufacturer recommends a minimum of 60 active pores for a successful sequencing run (Grädel *et al.*, 2019).

Given the small size of DNA fragments that were sequenced in this project, successful sequencing was obtained by flow cells with active pores which were below the minimum threshold number. Sequencing runs with 40 active pores only required a run time of an average of 2 hours 30 mins to produce 100 000 passed reads of q-score 7 and above. The low number of active pores required to produce such a high depth of sequencing, would suggest that the use of barcoding would be a feasible option for running multiple samples in a single run. With a larger number of active pores, one can perform multiplex sequencing by barcoding the individual samples and lowering the time it takes to sequence several samples.

The cost implications can be reduced significantly by multiplex sequencing, since a single Flongle flow cell can be used to sequence several samples. The MinION flow cell can be alternatively used, as it has 512 pores. With the use of multiplex sequencing, 10 samples can be sequenced in the same sequencing run, using barcoding. This would reduce the duration of the analysis of multiple samples, as well as lowering the cost.

The data from the sequencing run is collected in real time. The advantage of this cloud based data collection and storage is that sequencing can be performed in a different site and the data can be collected at a central laboratory. This opens up the possibility of mobile laboratories generating and collecting data, which is then sent to a central laboratory where the data is analysed. Cloud based data collection also allows for data sharing and collaborative research opportunities. The disadvantage of this model of data collection and accessing is that it is dependent on the presence of an internet connection. Without an internet connection, one cannot access the sequencing data.

4.4. Data Analysis

From the sequencing report provided by the Nanopore MinION's cloud-based software (Metrichor), it was observed that there was a small percentage of reads which were much longer than the expected size and those that were much shorter than the expected length. These anomalies were attributed to library preparation artefacts. The longer reads were ligated strands which joined end to end during the adaptor ligation steps of the library preparation. The shorter reads were likely from strands of DNA which fragmented during the vortexing steps of the library preparation and still underwent adaptor ligations and were sequenced.

ONT provides a number of recommended data analysis tools for nanopore sequencing data. However, the tools were developed for analysing long reads and the program rejected the shorter reads from my study, deeming them implausibly short. During the course of my project, another research group published a study where the Oxford Nanopore MinION was used to sequence the SCA7 triplet repeat region in cloned plasmids. The study constructed semi artificial plasmid/human DNA sequences of SCA7 that had an average length of 16 565.3 bp (Mitsuhashi *et al.*, 2018). The data analysis in that project used very long reads as their initial data was also rejected after using shorter reads with sequences that began at positions that were too close to the CAG repeat tract (Mitsuhashi *et al.*, 2019). The reads generated in my project after sequencing the SCA7 triplet repeat region, were less that 1Kb and were rejected by the manufacturer's recommended data analysis tools. In order to analyse the data, an alternative data analysis pipeline was designed to determine the size of the triplet repeat region in the control and patient samples.

The alternative data analysis pipeline designed for this study, was unable to characterise the exact sequences of the triplet repeat tract, therefore the alternative was to calculating the size of the triplet repeat tract. This was not as stated in the original objective of the project, but still managed to provide vital information about the possibility of variations within the triplet repeat tract and in the sequences before and after the triplet repeat sequence.

For a better understanding of the data analysis process, terms used are defined below:

Filtered reads – Reads that have been filtered by size and q-score

Trimmed reads – Reads that begin at the start of the CAG repeat sequence and everything before the CAG repeat has been cut off

Isolated reads – Reads that begin at the start of the CAG repeat sequence and end at the CCG sequence that flanks the end of the triplet repeat tract. These reads only contain a conserved CAG repeat sequence and a CCG sequence at the end of the CAG repeat sequence

The first step in data analysis was to filter the raw reads by size and q-score. Although this process was done in the basic QC step, some samples, like control 1C, lost too many reads as the filter of minimum q-score 12, was too stringent. Control sample 1C went from 31 258 reads at min q-score 7, to 90 reads at min q-score 12. The loss necessitated changing the q score limit to q score 10 which produced an output file with 29 973 reads for control sample 1C. The filter for the samples then became min q-score 10 for both control and patient samples. A size filter of min 250bp: max 350bp was used for control samples and min 250bp: max 800bp for patient samples. This process produced what we are referring to as the filtered reads.

The filtered reads were taken through a trimming process. This process separated the forward and reverse reads, then trimmed off sequences that came before the triplet repeat tract. There was a very notable difference in the number of forward reads and reverse reads. The reverse reads that passed the filtration by size and q-score 10, were much more abundant than the forward reads. In sample data from control samples 1A, 1B, 1D, 1H and 2A, 90% of the filtered reads were reverse reads. This gave a higher confidence in the reverse sequences, and so the reverse sequences were the only ones used for the data analysis process. A designed python script was used to separate the forward and reverse reads, then trim everything that came before the beginning of the triplet repeat tract. This trimming process produced what we are referring to as the trimmed reads.

The data analysis was done only on the reverse sequences, where the CAG repeat sequence presented as a CTG repeat sequence. The reference sequence used in my analysis is the GRCh38 and presents with a CGG sequence flanking the end of the CAG repeat sequence. Since the data analysis in my project used the reverse sequence data, the sequence following the triplet repeat tract on the GRCh38 reference reverse sequence, presents as CCG. This made it possible to design a python script to chop off the front of the trimmed reads and remove everything which follows the CCG sequence at the end of the triplet repeat tract. The chopping script gave what we are referring to as isolated reads, which consist of only the triplet repeat tract and CCG sequence at the end of each of the reads.

The next python script was designed to count all the characters in the isolated reads. The distribution of the number of counts in the isolated reads was plotted onto a scatter plot and produced a single peak in most of the sample data. This read count data only represented the reads that have no variations within the triplet repeat sequence and have a CCG at the end of the triplet repeat tract. Assuming that all the reads have uninterrupted CAG repeats and a conserved CCG flanking the end of the triplet repeat tract would result in the loss of relevant data.

The python script used to count the characters in the isolated reads was used to count the characters in the trimmed reads. The trimmed reads only discriminated based on direction of sequencing and not based on reads with particular conserved sequences. If there was no variations within the triplet repeat sequences in the reads or the sequences flanking the reads, then the same distribution pattern would be seen in both the isolated read counts and the trimmed read counts. The presence of variations within the CAG triplet repeat sequence of SCA7 have not been explored before in the South African population. The presence of two distinct peaks on the read count scatter plot of the trimmed reads suggests that one peak represents the isolated reads and the second peak represents read lengths of sequences that have variations from the GRCh38 reference.

In the analysis of the control samples, the isolated read counts managed to give a specific number of CAG repeats in the isolated reads. The trimmed read counts gave the total number of bases in the entire sequences of the reads. Trimmed read counts each presented with two peaks. All the control samples that showed 9 CAGs in the isolated read count plot, showed a peak length at 105 bases, in the trimmed read counts. All the control samples that showed 8 CAGs in the isolated read count plot, showed a peak length at 102 bases.

What this observation proves is that the adaptor sequences were cleaved off at the same positions in all the reads from the different sequencing runs. The observation also proves that the isolated reads are represented by the first peak in the trimmed read count scatter plot. Seeing that all the reads were cleaved at the same coordinates by the data collection software, the difference in number bases of the two peaks on the trimmed read plots represented the difference in the number of bases in the triplet repeat tract of the two alleles in each of the sample read data.

The patient samples' sequencing data files were analysed in the same way as the control samples' data. For both the patient samples, the isolated read counts were plotted on scatter plots and showed a single prominent peak and the number of CAG within the conserved isolated read length was

determined for each of the patient samples. From the control sample analysis, we had determined that the data collection software cleaved the reads at the same positions in all the sequencing runs and this was also seen in the patient samples.

In patient sample T1, the isolated reads showed a peak read count with 9 CAGs. The trimmed read counts plot showed the first peak at 105 bases. This further cemented the conclusion that the start and end points of each of the reads were on similar coordinates. The difference in number of bases in each of the trimmed read count peaks is the measure of the difference in the number of bases in the triplet repeat tracts of each of the alleles in the sample being analysed. Patient sample T2, however, had a slight deviation to the trend. The isolated read count plot for sample T2 showed a peak with 9 CAGs, but the trimmed read count plot presented the first peak to be at 102 bases. The theory formulated, would predict the size of the first peak to be 105 bases, but sample T2 data showed a deviation of 3 bases from the expected peak read-length.

The patient samples T1 and T2 had trimmed read plots that showed 3 peaks. This was not the expected result but can be explained as an error. The size filtration done in the basic QC was to remove uncharacteristically larger or smaller read lengths which were carried forward from library preparation errors. The patient samples each contained long fragments of expanded allele reads that could have been disrupted during the vortexing steps of the library preparation. Adaptors still bind onto the fragmented DNA and produce multiple reads which can be identified as a significant peak. Each of the patient samples' trimmed read scatter plots presented with three peaks. The peak that represented the expanded allele reads was expected, so were considered it to be accurate. On the two peaks of unexpanded allele reads, the larger peak was considered to be the more accurate. The library preparation error may produce a significant number of reads, but the read count would be notably less than the accurately prepared library of unexpanded allele sequences, as seen on the trimmed read scatter plot for each of the patient samples.

An allele distribution study showed that some genome sequencing data can suggest heterozygosity where there is homozygosity at a specific position, due to library errors. The study showed that these errors cannot be avoided or circumvented by increasing the sequencing depth, but rather through technical replicates (Heinrich *et al.*, 2012). Nanopore sequencing has a high depth of sequencing, but the sequences are produced from the same library. Technical replicates could address such errors, as different libraries could result in two peaks in the patient sample data, rather

than three peaks. A comparison can be made between the results obtained using CE and the results obtained in this project using nanopore sequencing, as shown in Table 4.1.

Table 4.1. A comparison of the number of CAG repeats calculated using capillary electrophoresis, compared to the number of CAG repeats calculated using Nanopore sequencing

Patient sample	Capillary Electrophoresis	Nanopore Sequencing
T1	7/58	9/67
T2	17/37	17/46

For patient sample T1, the un-expanded allele length had a difference of 6 bases (2 CAGs). Given that the nanopore sequencing was counting actual bases, rather than just relying on the sizing of a fragment using an internal length standard on capillary electrophoresis, it is likely that the nanopore sequence lengths are more accurate that the capillary electrophoresis lengths.

The expanded allele length difference between the capillary electrophoresis result of T1 and the nanopore sequencing result for T1, is 58 CAGs compared 67 CAGs. Although both estimates are clearly in the disease range, this is a large difference. If the triplet expansion numbers were close to the border between unaffected and affected, this would affect the interpretation and the outcome of the diagnosis. However, the reads used to determine the fragment length in the nanopore data analysis contained a lot of apparent stutter fragments and therefore it is difficult to confidently say that the result is more accurate than the capillary electrophoresis result. Both the methods have a PCR step, which is where stutter is likely to occur and thus both would be affected by this, but the nanopore relied on sequencing of the long triplet repeats and not just gel sizing. Since this is a preliminary study, with the nanopore sequencing not yet validated, we need to interpret it with caution.

For patient sample T2, the highest peak in the nanopore sequencing result showed a length of 17 CAG repeats for the non-expanded allele, which is the same result found on the capillary electrophoresis. Although there was a smaller peak in the position indicting 9 CAG repeats, it was not the highest peak and not the length with the most read length frequency in all the read data produced by nanopore sequencing for this sample.

The patient sample T2 result for the expanded allele using nanopore sequencing is 46 CAGs, compared to the 37 repeats from capillary electrophoresis method, a difference of 9 repeats. Interestingly, the triplet number differences between the two methods in T1 and T2 are both 9 repeats. Although the size difference is the same, the sample size (n=2) is too small to conclude that there is a trend in the difference between the two methods. The peak in the scatterplot of the expanded allele reads of the nanopore sequence reads showed a lot of stuttering and very small size differences in the reads clustered near the peak with the most frequency. This gives low confidence in the concluded length of the expanded allele reads and the method will need to be developed further.

4.5. Cost Comparison

The assessment of the duration of the laboratory and analysis work for the assay for each of the methods was an important element of the cost comparison. The times taken to perform different steps for each of the methods, were split into categories of hands-on duration and machine-use duration. The machine-use duration has no bias because the machinery functioning at full capacity will perform in the same time span regardless of the person conducting the analysis. The hands-on duration contains an element of bias because it involves performance durations of different individuals performing the analysis. The hands-on time duration for CE was based on information provided by scientists at the NHLS in Cape Town who regularly perform the CE analysis. A more accurate assessment of the duration would be a comparison of the time taken by the same investigator to perform both the CE and the nanopore analysis methods.

Data analysis of the nanopore sequencing method involves the alternative pipeline which was designed because the manufacturer recommended tools rejected the input data. With more appropriate data analysis pipelines, the data analysis step of the nanopore sequencing methods would be greatly reduced. CE data analysis is automated because the method has been optimized specifically for analysis of this nature. The nanopore sequencing method's data analysis is not automated and requires more hands-on time.

The machine-use duration is the amount of time each machine is used during the analysis. This assessment is essential as it provides informed time spans of the usage of each of the machines. Each machine that is purchased specifically for this method of assessment is costed by calculating

the amount of time it is in use during the analysis method, based on a finite machine life-time of 5 years.

The CE method analyses 10 samples in a single run, while the nanopore sequencing method used in this project analyses one sample at a time. The cost comparison done in this project could give a more accurate picture if multiplex sequencing was used for the nanopore technology and 10 samples were being sequenced in one run. Duration cost (i.e. time) of the nanopore MinION during sequencing of 10 samples was not the major contributor to the large difference in cost comparison. The use of 10 flow cells was the major contributor. This therefore constituted a major misrepresentation of the potential costs involved in the use of nanopore technology. The costs were inflated by 10 times what they could have been and this was because of the project design.

More effective comparisons should be made, where the same project is conducted with a single flow cell and all the samples are multiplexed. With the use of barcoding, several DNA samples can be barcoded before library preparation and the flongle flow cell can be used to sequence all the 10 samples in one sequencing run. This would make use of a single library preparation kit and a single flow cell. The use of multiplexing reduces the costs speculated in Table 3.13 from R37 134.53 to R4 124.55. This is nearly a 10 fold reduction in cost. The cost reduction of using barcoding is also outlined on https://store.nanoporetech.com/pcr-barcoding-kit.html (accessed February 2022). The difference in cost show on the nanopore website is approximately the same ratio as the one calculated in this project, as shown in figure 4.1.

	No barcodes	6 barcodes	12 barcodes
Flow cell price	$500	$500	$500
Library price	$99	$99	$99
Barcode price	-	$25	$50
Price per sample	$599	$104	$54

Figure 4.1. An extract of the from the nanopore website, of comparison of the cost per sample of nanopore sequencing in the absence of barcodes, with the use of 6 barcodes, and with the use of 12 barcodes. https://store.nanoporetech.com/pcr-barcoding-kit.html (accessed February 2022)

As shown in Figure 4.1, the cost per sample, with use of 12 barcodes, is reduced by more than 10 times. The cost of library preparation in each of the sequencing costs being analysed remains the same, as there is only one library prepared in all the different quantities of barcodes used. This method was, however not the one implemented in this project, but offers suggested recommendations for improving this protocol and giving a better comparison for introducing this method in a clinical setting.

Table 4.2: A representation of the cost implications of multiplex nanopore sequencing of 10 samples using a PCR barcoding kit

ITEM	**Unit**		**Run time cost (10 samples)**	
	Qty	**Cost**	**Qty**	**Cost**
Oxford Nanopore MinION	1 (5 years - 0.33/hr)	14500	25hrs	8.27
PCR Barcoding kit	12	725	10	604.16
Library Prep Kit	6	8985	1	1 497
NEB Companion Library Module	24 reactions	16474	1 reaction	486.42
SPRI Beads	5000µL	1435.50	100 µL	28.70
Flongle Flow Cell	1	1350	1	1350
Control DNA expansion Kit (6rxns)	6	900	1	150
TOTAL			**R 4 124.55**	

The analysis costs compared in this project are significantly less for the CE method than for nanopore sequencing because the components considered in the costing analysis are running costs and the set-up costs. Setting up a nanopore sequencing work station has much lower costs than setting up a CE workstation. The nanopore MinION sequencing device costs very little compared to the ABI gene analyser, or any other sequencing device. The reagents and consumables of the nanopore sequencing method are, however, much more expensive than the reagents of the capillary electrophoresis, significantly raising the costs on the nanopore sequencing method. The cost of nanopore sequencing assessed in this project is inflated due to the absence of multiplexing and barcoding, and can be significantly reduced if the project is to be redesigned.

The comparison of the two analysis methods was done considering that the methods were both done in an academic molecular laboratory. The costing procedure in a commercial lab would include everything that the samples touch, which accounts for storage, gloves, pipettes, and other expenses that were all included as overheads. The overheads were a cumulative percentage of the total base costs of the machinery and reagents specific to each of the analysis methods, and the expertise and time required to conduct each of the analyses. All these aspects considered, it costs a lot more to sequence the samples using Oxford Nanopore than it costs to do size analysis using the capillary electrophoresis technique.

4.6. Appropriateness of the project protocol

The appropriateness of a protocol is measured by its strengths, limitations and ability to reach the project objectives. The main aim of the project was to develop an analysis protocol for the diagnosis of SCA7 using Oxford nanopore sequencing and to assess the cost implications associated with the developed analysis protocol. The intention was to explore possibilities of nanopore sequencing becoming the alternative diagnosis method for SCA7, so the effectiveness, cost implications and appropriateness of the nanopore sequencing analysis method was compared to the currently used CE method.

Primer sets were designed and the PCR optimization process resulted in the development of a robust PCR protocol. The choice of polymerase used in the PCR optimization, Prime Star GXL TaKaRa, was based on literature of comparative studies of different polymerases. This polymerase is most ideal in amplifying large fragments with G-C rich sequences. SCA7 can reach a CAG triplet expansion length of up to 460 CAG repeats and the optimization procedure took this into account. The PCR products were sent for Sanger sequencing and confirmed the appropriateness of the primers and the PCR optimization.

Nanopore sequencing involves a fairly long library preparation process. Although the library preparation has a detailed, stepwise checklist, there are several disadvantages to it. The many steps of the library preparation protocol increases the chances of errors that may lead to failure to sequence or low quality sequencing data. There is no way of telling how good the library is before sequencing. The quality of the sequencing data is the only indicator. The high costs of the library preparation reagents make it difficult for technical replications to be done. Sequencing errors and

library preparation errors can only be identified and avoided by making several library replicates and sequencing them. The vortexing steps in the library preparation have a good chance of causing DNA fragmentation during library preparation, especially when handling samples of expanded fragment lengths.

The sequencing protocol used in this project made use of the Flongle flow cell. The 126 well flow cell requires a minimum of 60 active pores for a guaranteed high depth of sequencing of long reads. The SCA7 fragments sequenced in this project were approximately 300bp and required an average of 40 active pores to sequence 100 000 passed reads in two and a half hours. The use of barcoding in this protocol would have allowed for multiplex sequencing of up to three samples at a time. This would have cut down on the data collection duration, lowered the cost of the flow cells used in this project and narrowed down the overall project cost comparison difference between the nanopore sequencing and the CE done at the NHLS.

A variety of analysis tools are recommended by ONT. Data analysis pipelines and automated analysis notebooks are offered by the MinION sequencing device manufacturer. Reviews of these analysis tools and what they are capable of, were on the top of the list of reasons for proposing nanopore sequencing as an alternative to the currently used CE size analysis. It was unfortunate that my data was rejected by the recommended data analysis tools, on account of read length being 350bp and implausibly small. An alternative data analysis pipeline was developed, but the pipeline was aimed at calculating the size of the triplet repeat tract in the control and patient samples, without analysing the exact sequence of the triplet repeat tract. The currently used method of analysis CE uses this assessment model of characterising the triplet repeat tract by fragment size.

Although few, there were advantages of the developed sizing pipeline for nanopore data. Unlike the CE analysis, the nanopore analysis method is not dependent on sizing standards to determine SCA7 allele sizes in a provided sample. Since the developed analysis method does what CE does, it is appropriate to measure the limitations of the nanopore data analysis pipeline against the qualities of the CE method.

Size analysis by CE involves a simple, cost effective and quick sample preparation step. Using nanopore sequencing for size analysis involves longer and more complex sample preparation. The expenses of the library preparation and the flow cell do not justify the use of nanopore sequencing for size characterization of the SCA7 triplet expansion analysis. CE is the commercially used size

analysis method and has been validated. Using nanopore sequencing as an alternative would require validation. The patient samples used in this project were previously analysed using CE. The nanopore sequencing size characterization results of the analysis of patient samples T1 and T2 did not perfectly agree with the CE size characterization of the same patient samples. A method comparison on the 10 control samples would give a more statistically sound comparison of the two size characterization methods when the results of the two methods are measured for agreeability using the Cohen's Kappa (Warrens, 2015).

The nanopore sequencing size characterization data analysis method made some bold assumptions which may have compromised the quality of the conclusions in the allele size assignments. The isolated reads that were used to assign the number of repeats in one of the alleles of each sample, were produced by a using a python script that would trim reads ending with CCG, since the reference sequence showed that the end of the triplet expansion region has a CCG sequence. The limitation of this method is that it assumes that there are no interruptions within the triplet repeat sequence. Triplet repeat sequences with a CCG interruption in the triplet repeat tract would be cut off at the point a CCG interruption occurs and be assigned an inaccurate triplet repeat count. The occurrence of a CCG interruption in the triplet repeat sequence would completely invalidate the analysis method.

The appropriateness of the nanopore sequencing analysis protocol method can be assessed in two parts, where we assess the data collection protocol separately from the data analysis protocol. The designed primers managed to amplify all the patient and control samples. Regardless of the high frequency of variations in the African genome, the primer binding sites showed no variations and the designed primers were appropriate. The PCR optimization was robust and optimized to be effective in amplifying long stretches of CAG repeats. The optimization included the addition of the TaKaRa polymerase. This was to accommodate SCA7 expansions which can be far beyond those assessed in this project.

The sequencing procedure was appropriate for high depth of sequencing, although the expenses of the library preparation kit and the flow cell made it very expensive to have technical replicates and confirm the effectiveness of the protocols. Given the parameters of the data analysis outcome, the costs do not justify the use of nanopore sequencing. It is difficult to give an informed conclusion on the effectiveness of the sequencing procedure or the error rate accompanied by sequencing long

stretches of triplet repeats in the method used, since the sequence analysis was not done in the data analysis, and only the size of the reads was considered.

The data analysis pipeline was not appropriate for assessing sequence variation to suggest it as an alternative to the currently used CE method. The objective of the project was to characterise the sequences in the triplet repeat tract and to assess the possibility of variations, as well as to accurately identify the number of triplet repeats in the expanded and unexpanded SCA7 alleles. The nanopore data analysis pipeline concluded that all the control samples were heterozygous. It is concerning that there was no homozygous result, although this could be attributed to the small sample size, and brings questions to the validity of the specific analysis pipeline used.

5. Conclusion and Recommendations

During the course of this project, advancements were made to improve the protocols for nanopore sequencing analysis. Improvements to the methods and the data analysis were done and researcher communities outlined the challenges they faced, as well as the successful actions they have taken to overcome these challenges. The analysis protocol used met several challenges which made it inappropriate to satisfy the objectives of the project. This chapter explores the advances that have been made in the past two and half years and how they can improve the method used in this project. The manufacturer's improvements to the MinION sequencing protocol also provide recommendations for future adjustments and use of protocols more appropriate to meet the objectives of this study.

5.1. Data collection

Although the PCR optimization was successful and appropriate, the length of the amplicons produced in the PCR amplification for both the control samples and the patient samples was below 1kb. Other studies that have recently sequenced the SCA7 triplet repeat sequence have faced difficulties when analysing reads that are short. The tools designed for analysis of nanopore data are suitable for long read data. Recent studies showed that the SCA7 region was successfully sequenced using the MinION sequencing device, but the procedure used semi artificial rel3 plasmids which were an average length of 16 565.3 bases (Mitsuhashi *et al.*, 2019). Another study used nanopore sequencing to identify and characterise short tandem repeats for forensic analysis.

Amplicon ligation protocols were used to concatenate the short amplicons and create libraries with DNA fragments of a median length between 1000 and 2000 bp (Cornelis *et al.*, 2017). The data analysis tools that are available, are designed to analyse long reads, so a recommendation to improve the protocol used in this project is to design primers that amplify a longer fragment of the *ATXN7* gene in order to meet the read length requirements of the available data analysis tools.

Library preparation of long fragments resulted in mechanical shearing of fragments which led to the rupture of long DNA fragments. The recommendations to improve the library preparation, which are relevant to the analysis used in this project, are to use wide tip pipettes and to minimize vortexing (Dippenaar *et al.*, 2021). The patient samples used in this project presented with three read length peaks that associated with possible rupture of DNA fragments during library preparation. Recent developments in ONT have introduced and perfected the VolTRAX automated library preparation device. This device automatically prepared sequencing libraries and prevented sample handling errors. The VolTRAX can perform multiplex library preparation for up to 10 samples (https://nanoporetech.com/products/voltrax).

5.2. Data analysis

Improvements in the data analysis tools for nanopore data have continued over the past two and half years. The current flow cell pore advancement has developed from the R9.4 which was used in this project to the R10, and more recently the R10.3 (Dippenaar *et al.*, 2021). These advancements may be very relevant in the base calling accuracy in homopolymers and reduce the need for validating nanopore sequence accuracy with SRS data (Giesselmann *et al.*, 2019), but hold no relevance to the analysis used in this project, since the data analysis tools used to analyse nanopore data, rejected the data produced in this project on account of read length.

It is also important to note that improvements have been made to nanopore sequencing, where the accuracy has increased to Q30 coverage. This means that the probability of an error in mis-assignment of a base is 1 in 1000, bringing the accuracy levels to 99.9% and on par with the accuracy of Illumina sequencing. (Riaz *et al.*, 2021). As a recommendation for improvement to the current protocol, it is important to identify whether the latest advancements are more suitable in terms of analysis pipelines. A small adjustment in the PCR amplification will result in longer

fragments and analysis protocols will be developed from the most appropriate analysis tools being developed now.

The designed protocol used in this project was a size analysis. As there were no size standards, the size analysis was dependent on the actual sequence, based on the GRCh38 reference sequence. The SCA7 triplet repeat sequence is characterised as having a CCG following the last CAG repeat. A python script was developed to trim the reads on the occurrence of a CCG after a sequence of at least four continuous CAGs. These parameters were set this way because the minimum number of CAG repeats that are present in the human genome is four. A limitation to this was the possibility of a variation within the triplet repeat tract which is characterised by the occurrence of a CCG in the middle of the CAG triplet repeat sequence. The python script would simply trim the triplet repeat sequence on the position of the CCG variation occurrence, and misrepresent the read length during data analysis. A recommendation to improve this potential data analysis bias is to align the trimmed reads to the isolated reads and observe to see if there is the occurrence of more CAGs after the CCG.

5.3. Cost Implications

One of the objectives of the study was to perform a cost analysis of the SCA7 triplet repeat sequence. The sequencing device chosen for this is the Oxford Nanopore MinION which, at the start of this project, was considered the most affordable sequencing device on the market and remains the most affordable sequencing device to date (Dippenaar *et al.*, 2021). Due to the alternative data analysis pipeline used in this project, the analysis results gave a size characterization, rather than a sequence analysis characterization. Due to the nature of the data analysis pipeline designed, it is not relevant to compare the costs of this analysis procedure to other sequencing analysis costs. The currently used CE method performs size analysis, and so a direct cost comparison of the two procedures is justified. If the data analysis pipeline had provided a sequence analysis, the cost difference would be justified as this protocol would have provided more information than provided by CE.

From the cost comparison done, it is apparent that the initial costs of setting up an analysis laboratory for capillary electrophoresis is more expensive that setting up for nanopore sequencing. The genetic analyser used in CE has other functions in the laboratory and can be used in Sanger sequencing as well. Although this is not relevant to my project, it is still worth noting that the

higher set up costs are justified by the multiple uses of genetic analyser. The running costs for the CE analysis are significantly lower that the running costs for Oxford Nanopore sequencing analysis method because CE sample preparation is simple and reagent costs are less than the cost of the nanopore sequencing library kit.

Recommendations to improve the cost of the nanopore sequencing in the current study are to adjust the sequencing protocol and use multiplex sequencing. The cost comparison was conducted in the context of analysis of 10 samples. The currently used method sequences one sample in each sequencing run. The use of a larger flow cell, such and the MinION flow cell which has 512 available pores can allow 10 samples to be sequenced in one run, through the use of barcoding. Flow cell costs would be decreased, as well as hands on and machine use time for analysing 10 samples. Although this adjustment may improve the cost implications, a better approach would be to develop pipelines that characterise the sequences of the samples, as the initial objective of the study states. The greater value of the obtained results would justify the cost implications and allow for a cost comparison to be conducted against other sequencing methods, rather than against CE analysis.

From a laboratory perspective, the nanopore, once you have done the library preparation, is more expensive and takes more time than just preparing the sample for CE, but with an improved protocol, nanopore sequencing offers more comprehensive information. The Oxford Nanopore MinION is portable, fast and the most affordable sequencing device, although the consumables for the assays are expensive. A genetic analyser is expensive and takes up permanent laboratory work space. From a clinical perspective, adjusting the nanopore analysis pipeline would afford the analysis to provide more information than the CE size analysis. Possibilities of sequence variations and investigating them can provide more understanding of the progression, prognosis and clinical implications based in the exact sequence variations in SCA7 patients. From a patient perspective, very little is known about the disease and relevant genetic counselling would be required for a patient to value a more detailed diagnosis. The currently used method has significantly lower costs and efficiently provides results in the form of a negative or positive diagnosis for SCA7.

5.4. Future recommendations

The intention of this project was to design an analysis protocol for the sequencing of the triplet expansion that causes SCA7 using nanopore sequencing in order to implement this method as an alternative to the currently used CE analysis. Although the final outcome and analysis did not meet the original objectives of the project (i.e. produce sequence data), the different limitations which made the analysis protocol inappropriate for satisfying all the project objectives can be addressed by recommendations from other similar studies.

The fast rate of development of analysis tools is improving the data analysis pipelines of nanopore data. The use of these new nanopore data analysis pipelines was not possible in this project due to the small size of the reads produced through sequencing. Although it is possible to create data analysis pipelines that accommodate the size of reads produced in this project, it is beyond the scope of my expertise and beyond the scope of this project. An easier and more appropriate recommendation is to design primers that amplify a larger fragment of the *ATXN7* gene that can produce amplicons which are between 1000 and 2000 bp. Successful attempts at sequencing tandem repeats using nanopore sequencing, have used libraries made of DNA fragments longer than 1kb.

Longer fragments bring about the concern of DNA fragment rupture during library reparation. The use of the VolTRAX automated library preparation device comes with a huge cost implication. Other researchers have recommended the use of pipette tipped with wider openings and avoiding prolonged vortexing where possible. Technical replications can be done by preparing several libraries of the same samples and comparing the data, to ensure that the errors in the library preparation are reduced.

To propose the use of nanopore sequencing as an alternative analysis method to CE, a validation process is required. Developing more suitable protocol, according to the protocol adjustment recommendations above, a larger sample size is required to validate the sequencing results. With effective sequencing protocols and data analysis methods, the cost implications may be justifiable, due to the value added by the sequencing analysis. Improvements to the costs of nanopore sequencing could be done through sequence multiplexing by barcoding. Although the nanopore sequencing running costs would remain higher than CE analysis running costs, the value added has the potential of being more beneficial to the clinician and to the patient, if it is shown that the

sequence of the triplet repeat region could provide better estimates of the prognosis in a patient and to meiotic stability when couples are making reproductive choices.

6. References

Adams, D. and Wasserstein, M. (2021) “Free Sialic Acid Storage Disorders Summary Genetic counseling GeneReview Scope,” pp. 1–18.

Almeida, B. *et al.* (2013) “Trinucleotide repeats: A structural perspective,” *Frontiers in Neurology*, 4 JUN(June), pp. 1–24. doi: 10.3389/fneur.2013.00076.

Ampe, F. (2019) “THE USE OF NANOPORE SEQUENCING IN ECOTOXICOLOGY .,” pp. 2018–2019.

Babar., U. (2017) “Monogenic Disorders: an Overview.,” *International Journal of Advanced Research*, 5(2), pp. 1398–1424. doi: 10.21474/ijar01/3294.

Cagnoli, C. *et al.* (2018) “Spinocerebellar Ataxia Tethering PCR: A Rapid Genetic Test for the Diagnosis of Spinocerebellar Ataxia Types 1, 2, 3, 6, and 7 by PCR and Capillary Electrophoresis,” *Journal of Molecular Diagnostics*. American Society for Investigative Pathology and the Association for Molecular Pathology, 20(3), pp. 289–297. doi: 10.1016/j.jmoldx.2017.12.006.

Chintalaphani, S. R. *et al.* (2021) “An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics,” *Acta Neuropathologica Communications*. BioMed Central, 9(1). doi: 10.1186/s40478-021-01201-x.

Choudhry, S. (2001) “CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms,” *Human Molecular Genetics*, 10(21), pp. 2437–2446. doi: 10.1093/hmg/10.21.2437.

Cornelis, S. *et al.* (2017) “Forensic SNP Genotyping using Nanopore MinION Sequencing,” *Scientific Reports*. Nature Publishing Group, 7, pp. 1–5. doi: 10.1038/srep41759.

Dieffenbach, C. W., Lowe, T. M. J. and Dveksler, G. S. (1993) “General concepts for PCR primer design,” *Genome Research*, 3(3). doi: 10.1101/gr.3.3.S30.

Dippenaar, A. *et al.* (2021) “ Nanopore sequencing for Mycobacterium tuberculosis : a critical review of the literature, new developments and future opportunities ,” *Journal of Clinical*

Microbiology, (June). doi: 10.1128/jcm.00646-21.

Dorschner, M. O., Barden, D. and Stephens, K. (2002) "Diagnosis of five spinocerebellar ataxia disorders by multiplex amplification and capillary electrophoresis," *Journal of Molecular Diagnostics*. American Society for Investigative Pathology and Association for Molecular Pathology, 4(2), pp. 108–113. doi: 10.1016/S1525-1578(10)60689-7.

Etter, P. D. *et al.* (2011) "Local de novo assembly of rad paired-end contigs using short sequencing reads," *PLoS ONE*, 6(4). doi: 10.1371/journal.pone.0018561.

Gardiner, S. L. *et al.* (2019) "Prevalence of carriers of intermediate and pathological polyglutamine disease-Associated alleles among large population-based cohorts," *JAMA Neurology*, 76(6), pp. 650–656. doi: 10.1001/jamaneurol.2019.0423.

Giesselmann, P. *et al.* (2019) "Nanopype: A modular and scalable nanopore data processing pipeline," *Bioinformatics*, 35(22), pp. 4770–4772. doi: 10.1093/bioinformatics/btz461.

Grada, A. and Weinbrecht, K. (2013) "Next-generation sequencing: methodology and application," *The Journal of investigative dermatology*, 133(8), p. e11. doi: 10.1038/jid.2013.248.

Grädel, C. *et al.* (2019) "Rapid and Cost-Efficient Enterovirus Genotyping from Clinical Samples Using Flongle Flow Cells," *Genes*, 10(659), pp. 1–12.

Heinrich, V. *et al.* (2012) "The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process," *Nucleic Acids Research*, 40(6), pp. 2426–2431. doi: 10.1093/nar/gkr1073.

Huang, T., Shu, Y. and Cai, Y. D. (2015) "Genetic differences among ethnic groups," *BMC Genomics*. BMC Genomics, 16(1), pp. 1–10. doi: 10.1186/s12864-015-2328-0.

Huntington, E. D. (2001) "New problems in testing for Huntington ' s disease : the issue of intermediate and reduced penetrance alleles," 38(39), pp. 1–5.

Jain, M. *et al.* (2018) "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature Biotechnology*. Nature Publishing Group, 36(4), pp. 338–345. doi: 10.1038/nbt.4060.

Jia, H. *et al.* (2014) "Long-range PCR in next-generation sequencing: Comparison of six enzymes and evaluation on the MiSeq sequencer," *Scientific Reports*, 4, pp. 1–8. doi: 10.1038/srep05737.

Kraus-Perrotta, C. and Lagalwar, S. (2016) "Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1," *Cerebellum & Ataxias*. Cerebellum & Ataxias, 3(1), pp. 1–11. doi: 10.1186/s40673-016-0058-y.

Krause, A., Seymour, H. and Ramsay, M. (2018) "Common and Founder Mutations for Monogenic Traits in Sub-Saharan African Populations.," *Annual review of genomics and human genetics*. United States, 19, pp. 149–175. doi: 10.1146/annurev-genom-083117-021256.

Loomis, E. W. *et al.* (2013) "Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene," *Genome Research*, 23(1), pp. 121–128. doi: 10.1101/gr.141705.112.

Lu, H., Giordano, F. and Ning, Z. (2016) "Oxford Nanopore MinION Sequencing and Genome Assembly," *Genomics, Proteomics and Bioinformatics*. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 14(5), pp. 265–279. doi: 10.1016/j.gpb.2016.05.004.

Lutz, R. E. (2007) "Trinucleotide Repeat Disorders," *Seminars in Pediatric Neurology*, 14(1), pp. 26–33. doi: 10.1016/j.spen.2006.11.006.

Mardis, E. R. (2011) "A decade's perspective on DNA sequencing technology," *Nature*, 470(7333), pp. 198–203. doi: 10.1038/nature09796.

Martindale, J. E. (2017) "Diagnosis of spinocerebellar ataxias caused by trinucleotide repeat expansions," *Current Protocols in Human Genetics*, 2017(January), pp. 9.30.1-9.30.22. doi: 10.1002/cphg.30.

McCombie, W. R., McPherson, J. D. and Mardis, E. R. (2019) "Next-generation sequencing technologies," *Cold Spring Harbor Perspectives in Medicine*, 9(11). doi: 10.1101/cshperspect.a036798.

Michalik, A., Martin, J. and Broeckhoven, C. Van (2004) "Spinocerebellar ataxia type 7 associated with pigmentary retinal dystrophy," 7, pp. 2–15. doi: 10.1038/sj.ejhg.5201108.

Min, Y. G. (2021) “Characteristics of very late-onset dentatorubro-pallidoluysian atrophy : the oldest onset case and review of literature.” Neurological Sciences, pp. 377–380.

Mitsuhashi, S. *et al.* (2018) “Robust detection of tandem repeat expansions from long DNA reads,” *bioRxiv*, p. 356931. doi: 10.1101/356931.

Mitsuhashi, S. *et al.* (2019) “Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads,” *Genome Biology*. Genome Biology, 20(1), pp. 1–17. doi: 10.1186/s13059-019-1667-6.

Morey, M. *et al.* (2013) “A glimpse into past, present, and future DNA sequencing,” *Molecular Genetics and Metabolism*. Elsevier Inc., 110(1–2), pp. 3–24. doi: 10.1016/j.ymgme.2013.04.024.

Morey, M. *et al.* (2015) “Corrigendum to ‘A glimpse into past, present, and future DNA sequencing’ [Mol. Genet. Metab. 110 (2013) 3-24],” *Molecular Genetics and Metabolism*. Elsevier Inc., 114(3), p. 484. doi: 10.1016/j.ymgme.2015.01.009.

Mulder, N. (2017) “Development to enable precision medicine in Africa,” *Personalized Medicine*, 14(6), pp. 467–470. doi: 10.2217/pme-2017-0055.

Musova, Z. *et al.* (2009) “Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene,” *American Journal of Medical Genetics, Part A*, 149(7), pp. 1365–1374. doi: 10.1002/ajmg.a.32987.

Nalavade, R. *et al.* (2013) “Mechanisms of RNA-induced toxicity in CAG repeat disorders,” *Cell Death and Disease*. Nature Publishing Group, 4(8), pp. e752-11. doi: 10.1038/cddis.2013.276.

nanoporetech.com (2017) “a Oxford Nanopore Technologies | the Application and Advantages of Long-Read Nanopore Sequencing To Structural Variation Analysis.”

Number, F. C. *et al.* (2019) “Sequence-specific cDNA-PCR Sequencing (SQK-PCS109) Before start checklist Sequence-specific cDNA-PCR Sequencing (SQK-PCS109),” pp. 1–8.

Paszkiewicz, K. and Studholme, D. J. (2010) “De novo assembly of short sequence reads,” *Briefings in Bioinformatics*, 11(5), pp. 457–472. doi: 10.1093/bib/bbq020.

Pereira, L. *et al.* (2021) “African genetic diversity and adaptation inform a precision medicine agenda.,” *Nature reviews. Genetics*. England, 22(5), pp. 284–306. doi: 10.1038/s41576-020-

00306-8.

Prior, T. W. and Committee, A. (2009) “Technical standards and guidelines for myotonic dystrophy type 1 testing BACKGROUND ON MYOTONIC DYSTROPHY,” 11(7). doi: 10.1097/GIM.0b013e3181abce0f.

Rhoads, A. and Au, K. F. (2015) “PacBio Sequencing and Its Applications,” *Genomics, Proteomics & Bioinformatics*. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 13(5), pp. 278–289. doi: 10.1016/j.gpb.2015.08.002.

Riaz, N. *et al.* (2021) “Adaptation of Oxford Nanopore technology for hepatitis C whole genome sequencing and identification of within-host viral variants,” *BMC Genomics*. doi: 10.1186/s12864-021-07460-1.

Salas-Vargas, J. *et al.* (2015) “Spinocerebellar ataxia type 7: A neurodegenerative disorder with peripheral neuropathy,” *European Neurology*, 73(3–4), pp. 173–178. doi: 10.1159/000370239.

Sandberg, G. and Schalling, M. (1997) “Effect of in vitro promoter methylation and CGG repeat expansion on FMR-1 expression,” *Nucleic Acids Research*, 25(14), pp. 2883–2887. doi: 10.1093/nar/25.14.2883.

Smith, D. C. *et al.* (2012) “Inherited polyglutamine spinocerebellar ataxias in South Africa,” *South African Medical Journal*, 102(8), pp. 683–686. doi: 10.7196/SAMJ.5521.

Smith, D. C. (2014) “Spinocerebellar ataxia type 7 in southern africa: an epidemiological, molecular and cellular study,” (August).

Smith, D. C., Esterhuizen, A. and Greenberg, J. (2013) “Caution regarding the interpretation of homoallelism in polyglutamine multiplex assays: A recommendation for confirmatory testing of homozygous alleles,” *Journal of Molecular Diagnostics*. American Society for Investigative Pathology, 15(5), pp. 706–709. doi: 10.1016/j.jmoldx.2013.05.009.

Squitieri, F. and Jankovic, J. (2012) “Huntington ’ s Disease : How Intermediate are Intermediate Repeat Lengths ? CAG Repeats and Penetrance Reported Cases and Series With IA and HD Phenotype,” 27(14), pp. 1714–1717. doi: 10.1002/mds.25172.

Sullivan, R. *et al.* (2019) “Spinocerebellar ataxia: an update,” *Journal of Neurology*. Springer

Berlin Heidelberg, 266(2), pp. 533–544. doi: 10.1007/s00415-018-9076-4.

Technologies, O. N. (2017) “Oxford Nanopore Technologies | the Application and Advantages of Long-Read Nanopore Sequencing To Structural Variation Analysis 3 2 1,” (September).

Tipu, H. N. and Shabbir, A. (2015) “Evolution of DNA Sequencing Technologies,” *Next-Generation Sequencing and Sequence Data Analysis*, 25(3), pp. 17–24. doi: 10.2174/978168108092511501006.

Warner, J. P. *et al.* (1996) “A general method for the detection of large GAG repeat expansions by fluorescent PCR,” *Journal of Medical Genetics*, 33(12), pp. 1022–1026. doi: 10.1136/jmg.33.12.1022.

Warrens, M. J. (2015) “Five Ways to Look at CohenÃ¢Â□Â™s Kappa,” *Journal of Psychology & Psychotherapy*, 05(04), pp. 8–11. doi: 10.4172/2161-0487.1000197.

Wei, S., Weiss, Z. R. and Williams, Z. (2018) “Rapid multiplex small dna sequencing on the MinION nanopore sequencing platform,” *G3: Genes, Genomes, Genetics*, 8(5), pp. 1649–1657. doi: 10.1534/g3.118.200087.

Weinstock, M. A. (2013) “Epidemiology and UV exposure.,” *The Journal of investigative dermatology*. Nature Publishing Group, 133(E1), pp. e11-4. doi: 10.1038/jid.2013.248.

Wilson, A. B. D., Eisenstein, M. and Soh, H. T. (2019) “High-Fidelity Nanopore Sequencing of Ultra-Short DNA Sequences.” doi: 10.1101/552224.

Zanetti, A. *et al.* (2019) “Molecular diagnosis of patients affected by mucopolysaccharidosis: a multicenter study,” *European Journal of Pediatrics*. European Journal of Pediatrics, pp. 739–753. doi: 10.1007/s00431-019-03341-8.

Appendix 1

Calculations

Control sample DNA was diluted to 50ng/µL using -

$C_1V_1 = C_2V_2$

E.g., Sample 1A: 138.155 x 10 = 50 x (TE+DNA)

TE + DNA = 27.63 µL

TE added = 17.63 µL

Final Concentration (Qubit) = 51.4 ng/µL

Patient sample DNA template concentrations calculated using:

$C_1V_1 = C_2V_2$

T1: 2.12µL DNA stock + 47.88µL TE buffer = 50µL T1 working Stock

T2: 5.43µL DNA stock + 44.57µL TE Buffer = 50µL T2 working Stock

Appendix 2

Python script: Trim.py

```
#!/usr/bin/env python3

import os
import sys

pattern1 = sys.argv[1]
pattern2 = sys.argv[2]

infile = sys.argv[3]
out1 = sys.argv[4]
out2 = sys.argv[5]

f = open(infile, "r")
g = open(out1, "w")
m = open(out2, "w")

data = []

for line in f:
        line = line.strip()
        data.append(line)

for i in range(len(data)-1):

        if data[i][0].startswith("@"):
                read_seq = data[i+1]

                if pattern1 in read_seq and (pattern2 not in read_seq):

                        ind1 = read_seq.index(pattern1)

                        if 90 < ind1 < 125:
                                g.write(data[i])
                                m.write(data[i].replace("@", ">"))
                                g.write(data[i+1][ind1:])
                                m.write(data[i+1][ind1:])
                                g.write(data[i+2])
                                g.write(data[i+3][ind1:])

f.close()
g.close()
m.close()
```

Appendix 3

Python script: Chop.py

```
#!/usr/bin/env python3

import os
import sys

pattern1 = sys.argv[1]

infile = sys.argv[3]
out1 = sys.argv[4]
out2 = sys.argv[5]

f = open(infile, "r")
g = open(out1, "w")
m = open(out2, "w")

data = []

for line in f:
	line = line.strip()
	data.append(line)

for in range(len(data)-1):

	if data[i][0].startswith("@"):
		read_seq = data[i+1]

		if pattern1 in read_seq:

			ind1 = read_seq.index(pattern1)

			g.write(data[i]+'\n')
			m.write(data[i].replace("@", ">")+'\n')
			g.write(data[i+1][:(ind1+len(pattern1))]+'\n')
			m.write(data[i+1][:(ind1+len(pattern1))]+'\n')
			g.write(data[i+2]+'\n')
			g.write(data[i+3][:(ind1+len(pattern1))]+'\n')

f.close()
g.close()
m.close()
```

Appendix 4

Python Script: Count.py

```
#!/usr/bin/env python3

import os
import sys


infile = sys.argv[1]

f = open(infile, "r")

test = []

for line in f:

        line = line.strip()
        if line.startswith(">"):
                pass
        else:

                a = len(line)-3

                if a % 3 == 0:
                        test.append(str(a))

                elif a % 3 == 1:
                        test.append(str(a-1))

                elif a % 3 == 2:
                        test.append(str(a-2))


data2 = ['Repeat_length']
data3 = ['Read_count']


for elem in test:

        if elem not in data2:
                data2.append(elem)
                data3.append(str(test.count(elem)))

        else:
                pass


for in range(len(data2)):
        print(data2[i] + "\t" + data3[i])


f.close()
```