Nanopore Sequencing of RNA from Breast Cancer Genes

A thesis submitted for the degree of Bachelor of Biomedical Science with Honours

Department of Pathology

University of Otago, Christchurch

Lucy de Jong

October 2016
Abstract

Abnormal mRNA splicing can disrupt gene function and influence the course of disease. Analysis of abnormal splicing is an important part of determining whether a particular genetic variant found in the population is pathogenic or not. However, to correctly identify abnormal splicing, we must first understand what is normal. This project assessed the isoforms of the genes \textit{BRCA1} and \textit{BARD1}, which are particularly relevant to the onset of breast cancer.

\textit{BRCA1} is a tumour suppressor gene implicated in breast cancer onset. \textit{BARD1} codes for a protein that interacts with \textit{BRCA1} and produces a smaller mRNA transcript. Normal exon skipping events have been identified for both \textit{BRCA1} and \textit{BARD1}, however, current methods are unable to reliably identify full transcripts. This has resulted in knowledge of individual exon skipping events but often does not tell us whether multiple events occur in the same transcript. The MinION nanopore sequencer (Oxford Nanopore Technologies), uses a nanopore to produce long-read, single molecule sequences. This has great potential for identifying multiple long isoforms, which is not practical using current technologies. The aim of this project was to examine the ability of the MinION to identify mRNA splicing patterns of transcripts derived from \textit{BRCA1} and \textit{BARD1}.

All mRNA from a normal lymphoblastoid cell line was converted to cDNA and targeted genes of interest were amplified by polymerase chain reaction (PCR). All potential isoforms generated from \textit{BRCA1} and \textit{BARD1} were then pooled and analysed using the MinION sequencer. After trialling many different analysis methods, the read data was analysed using the BLAST-like Alignment Tool (BLAT) with two outputs, a tabular and a graphical format. The tabular format grouped reads into potential isoforms, while the graphical format allowed visualisation of these isoforms and identified the exon/intron boundaries. Using both these formats 34 \textit{BRCA1} isoforms and 39 \textit{BARD1} isoforms were identified, 24 and 17 of which were potential novel isoforms respectively. Two of these novel isoforms from the \textit{BRCA1} dataset (Δ10-17 and Δ11q ▼ 21) were further verified using Sanger sequencing.

This was a proof of principle research project that demonstrated the potential use of the MinION nanopore sequencer for successful characterisation of multiple mRNA isoforms. This research has successfully identified a number of novel isoforms from the \textit{BRCA1} and \textit{BARD1} genes using the MinION sequencing device.
Acknowledgments

First and foremost, tremendous gratitude to my supervisors Prof. Martin Kennedy and Dr. Logan Walker. Between the two of you I have managed to (mostly) keep my head above water this year, due to your support and enthusiasm for this project even through the moments where I had lost faith in it. The long chats about the tiniest details were invaluable to my learning and understanding, and my deepest gratitude goes to you for reminding me how awesome this research is. Thanks also to Dr. Simone Cree for helping me through the MinION side of things, and fixing up my references multiple times trying to get alignments working, and the effort you have put in to teaching me how parameters work in all these programs from scratch. My sincere gratitude goes to the Cancer Society for funding this work, and to Dr. Amanda Spurdle and kConFab for donating the cell line I worked with this year. Thanks also to the Gene Structure and Function Lab and the Mackenzie Cancer Research Group for making this a great place to spend my Honours year, I have been so lucky to be a part of two fantastic groups.

My deepest thanks to Allison Miller and Dr. Vanessa Lattimore for putting up with all my questions and complaints throughout the days and nights respectively, I don’t think I would have stayed sane if you two weren’t always happy to listen to the latest issue I had and give advice on how to solve it. Thank you also to George Wiggins for being the person I ran to when I was struggling to learn how to use computers for science, and for giving me the line of R code that got this project moving - I don’t think you know how important that one line was!

Huge thanks go to my fiancé Cade Picard, who wrote the Python code needed to make the Samcat.py program, I could not have done this project so well or so fast without you. Additional thanks for putting up with my constant questions about computer commands and programs, for translating conversations on the help pages, and for keeping me laughing when I was too tired to function.

Thank you to the other Christchurch Honours students for being a great support and fun time this year, also to Prof. Madhav Bhatia, Prof. Mark Hampton, and Alice Milnes for keeping the program running so well with so many changes. Lastly, special thanks to my family and friends, who have kept me grounded in the real world throughout my honours year, and have kept me entertained in the off moments with normal life.
Table of Contents

Abstract........................................................................................................................................... ii

Acknowledgments........................................................................................................................... iii

Table of Contents........................................................................................................................... iv

List of Tables ....................................................................................................................................... vii

List of Figures ...................................................................................................................................... viii

List of Abbreviations ........................................................................................................................ ix

1 Chapter 1: Introduction ................................................................................................................. 1
   1.1 Overview ................................................................................................................................ 1
   1.2 BRCA1 and BARD1 .................................................................................................................. 1
       1.2.1 Gene structure ................................................................................................................... 2
       1.2.2 Cellular functions of BRCA1 and BARD1 proteins .......................................................... 2
   1.3 Pre-mRNA processing and final mRNA structure ..................................................................... 5
       1.3.1 Alternative Splicing .......................................................................................................... 6
       1.3.2 BRCA1 and BARD1 isoforms and the ENIGMA Splicing Working Group ....................... 10
       1.3.2.1 Splice event nomenclature .......................................................................................... 12
       1.3.3 Other mRNA modifications ............................................................................................ 12
   1.4 Methods of detecting exon structure ....................................................................................... 13
       1.4.1 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) ........................................ 14
       1.4.2 Molecular Cloning ........................................................................................................... 14
       1.4.3 Short read RNA-Sequencing (RNA-seq) ......................................................................... 15
       1.4.4 Other methods ................................................................................................................ 15
       1.4.5 Limitations of current methods ....................................................................................... 16
   1.5 Single molecule long-read DNA sequencing .......................................................................... 16
       1.5.1 PacBio Sequencing .......................................................................................................... 17
       1.5.2 MinION sequencing ........................................................................................................ 17
       1.5.3 Comparison of PacBio and MinION Platforms ................................................................. 18
   1.6 Aims and Hypotheses ............................................................................................................. 19

2 Chapter 2: Materials and Methods .............................................................................................. 20
   2.1 General stocks ........................................................................................................................ 20
       2.1.1 TAE buffer (Tris/Acetic acid/EDTA) ................................................................................. 20
       2.1.2 TE buffer (Tris/EDTA) .................................................................................................... 20
   2.2 RNA source ............................................................................................................................ 20
   2.3 Primer Design ......................................................................................................................... 20
   2.4 Preparation of cDNA ............................................................................................................. 24
   2.5 Polymerase Chain Reaction (PCR) ........................................................................................ 24
       2.5.1 Standard PCR for a product size <3kb .............................................................................. 24
       2.5.2 Standard PCR for full length BRCA1 mRNA (product size 5.8kb) .................................. 25
3 Chapter 3: Results .................................................................................................30

3.1 Introduction........................................................................................................30
3.2 PCR and identification of BRCA1 and BARD1 from cDNA.................................30
3.2.1 BRCA1 3' end and 5' end PCR products .........................................................30
3.2.1.1 Sanger sequencing .....................................................................................31
3.2.2 BARD1 PCR products ....................................................................................34
3.2.2.1 Sanger sequencing ....................................................................................34
3.2.3 Titration to determine optimal amount of cDNA ............................................36
3.2.4 Full length BRCA1 PCR ................................................................................36
3.2.4.1 Sanger sequencing ....................................................................................37
3.3 MinION ...............................................................................................................37
3.3.1 Samples sequenced ........................................................................................37
3.3.2 Statistics from the 2 runs ...............................................................................40
3.3.3 MinION read data analysis ............................................................................40
3.3.4 BLAT Tabular output ......................................................................................44
3.3.5 BLAT Graphical output and Samcat.py ...........................................................48
3.4 Verification of isoforms .....................................................................................54
3.4.1 PCR and sequencing of predicted isoforms .....................................................54

4 Chapter 4: Discussion ..........................................................................................58

4.1 Summary of results ..........................................................................................58
4.2 Sample preparation ..........................................................................................58
4.2.1 PCR and Sanger sequencing ..........................................................................58
4.2.2 Full length BRCA1 ......................................................................................59
4.2.3 MinION analysis ..........................................................................................59
4.3 Data analysis ...................................................................................................60
4.3.1 Iterative development of analysis pipeline .....................................................60
4.3.2 R and Excel manipulation .............................................................................62
4.3.3 Graphical format output and program development .......................................63
4.4 Isoform Analysis .............................................................................................65
4.4.1 Functional significance ..................................................................................66
4.5 Limitations of this work ..................................................................................67
4.6 Conclusion and implications of this research ..................................................68
Appendix A: Positions of primers described in Table 2.2 ........................................ 78
A.1 Isoform Δ10-17: ........................................................................................................ 78
A.2 Isoform Δ11q▼21 .................................................................................................... 79
A.3 Isoform Δ3,11Δ3110 .............................................................................................. 80
A.4 Isoform Δ11q,14,21 .............................................................................................. 81
A.5 Isoform Δ3,9,10,11q ............................................................................................ 83
A.6 Isoform Δ11q,19 .................................................................................................... 84

Appendix B: Pre-Amplicon protocol bead clean-up ................................................. 85

Appendix C: Amplicon library preparation protocol R9 Version ............................ 86

Appendix D: Data Analysis Methods ........................................................................ 95
D.1 Reference file used ............................................................................................... 95
D.2 Combine reads for full analysis ........................................................................... 97
D.3 BLAT analysis with parameters from UCSC website ......................................... 98
D.4 Separate BRCA1 and BARD1 out into different files ....................................... 99
D.5 R script .................................................................................................................. 99
D.6 Excel methods ...................................................................................................... 99
D.8 Samcat.py Program ............................................................................................. 100

Appendix E: BRCA1 Long Range PCR variations ..................................................... 106
List of Tables

Table 1.1: Alternative splice type identified with BRCA1 transcript analysis [60].*........ 8
Table 1.2: Nomenclature of mRNA isoforms.................................................................12
Table 2.1: Primers for cDNA synthesis, initial PCR and MinION sequencing, see Figure 2.1 for position on genes...............................................................................................21
Table 2.2: Primers for confirmation of isoforms. See Appendix A for positions on isoforms........................................................................................................................................23
Table 3.1: Number of bands per estimated size across all BRCA1 long range PCRs. ......37
Table 3.2: Iterative development of analysis pipeline (unsuccessful)................................42
Table 3.3: Iterative development of analysis pipeline (successful)..................................43
Table 3.4: BRCA1 isoforms found in BLAT tabular output.............................................45
Table 3.5: BARD1 isoforms found in BLAT tabular output *.........................................46
Table 3.6: BRCA1 isoform splice acceptor/donor shifts identification.............................53
Table 3.7: Isoform PCR and sequencing............................................................................55
List of Figures

Figure 1.1 BRCA1 (a) and BARD1 (b) exon structure and areas coding for major functional domains.................................................................4
Figure 1.2: Alternative Splicing.................................................................6
Figure 1.3: Cis- and trans- acting factors involved in alternative splicing of exons........7
Figure 2.1: Primer positions for those described in Table 2.1..............................22
Figure 2.2: Flow chart describing MinION data analysis final pipeline................28
Figure 3.1: BRCA1 3' end Amplification and Sequencing...............................32
Figure 3.2: BRCA1 5' end Amplification and Sequencing...............................33
Figure 3.3: BARD1 Amplification and Sequencing......................................35
Figure 3.4: Titration of cDNA amount.......................................................36
Figure 3.5: Pooled and purified BRCA1 and BARD1 samples..........................37
Figure 3.6: Varying long range BRCA1 PCR results....................................39
Figure 3.7: LAST alignment example result................................................40
Figure 3.8: Iterative development of final data analysis pipeline.......................41
Figure 3.9: Isoform Δ11q, ▽21 BLAT final graphical result.............................49
Figure 3.10: Isoform Δ10-17 BLAT final graphical result..............................50
Figure 3.11: Isoform Δ3,11Δ3110 BLAT final graphical result..........................50
Figure 3.12: Isoform Δ11q,14,21 BLAT final graphical result...........................51
Figure 3.13: Isoform Δ3,9,10,11q BLAT final graphical result..........................51
Figure 3.14: Exon variations in 22 reads from the full length group graphical result..52
Figure 3.15: PCRs to confirm isoform Δ10-17............................................56
Figure 3.16: PCRs to confirm parts of isoform Δ1Aq,2p,11q,16p, ▽21...................57
List of Abbreviations

°C Degrees Celsius
µl microliter
µM micromolar
ANK Ankyrin repeats (functional protein domain)
BLAST Basic Local Alignment Search Tool
BLAT BLAST-like Alignment Tool
bp base pairs (per nucleotide)
BRCT BRCA1 C-Terminus (functional protein domain)
cDNA complementary DNA synthesised from an RNA template
DNA Deoxyribonucleic acid
kb kilobases (unit of measurement of 1000 base pairs)
LAST Local Alignment Search Tool
MAF Multiple Alignment Format
MAP MinION Access Program
mRNA Messenger ribonucleic acid
MPW Millipore grade water
NES Nuclear export sequence
NLS Nuclear localisation sequence
NMD Nonsense-mediated decay
ONT Oxford Nanopore Technologies
Oligo(dT) Oligonucleotides composed of a string of Thymines that are complementary to the Poly(A) tail on an mRNA
PCR Polymerase Chain Reaction
Poly(A) String of Adenine nucleotides (usually at the tail of an mRNA)
RING Really Interesting New Gene (functional protein domain)
RNA Ribonucleic acid
RNAP RNA polymerase
RT-PCR Reverse Transcriptase - Polymerase Chain Reaction
UTR Untranslated region
UV Ultraviolet
Chapter 1: Introduction

1.1 Overview
An important aspect of understanding the genetic basis of many diseases is the underlying genetic variation that contributes to the disease phenotype. In cancer this is particularly important due to the nature of its development in the human body. Genetic variation can contribute greatly to the differential expression of mRNA alternative splicing patterns (splice isoforms), resulting in variations in protein expression. This can have vastly varying effects on cell phenotype, and in some cases can contribute to disease development. Currently, the most popular methods of analysing alternative splicing can only detect individual or adjacent exon inclusion/exclusion, and only in very small genes has it been possible to look easily at an entire transcript. Long-read sequencing devices such as the MinION nanopore sequencer, developed by Oxford Nanopore Technologies (ONT), have the potential to read through an entire transcript of a typical gene. The \textit{BRCA1} gene associated with breast cancer has been very well studied and is currently used in clinical decision making. It has multiple splice isoforms, which have yet to be fully characterised because of technological limitations. The \textit{BARD1} protein interacts with \textit{BRCA1} and is important for the role \textit{BRCA1} has in DNA repair [1]. \textit{BARD1} also has many splice isoforms. The isoforms of both \textit{BRCA1} and \textit{BARD1} have a pathological significance in the development and progression of breast and other cancers. This research aimed to characterise full length transcripts, and therefore splice isoforms, of these two breast cancer associated genes using the MinION nanopore sequencer.

1.2 \textit{BRCA1} and \textit{BARD1}
The gene \textit{BRCA1} (breast cancer susceptibility gene 1) is a well-known tumour suppressor discovered in 1994 [2] and has a wide variety of cellular functions [3, 4]. \textit{BARD1} (BRCA1-associated ring domain 1) was discovered in 1996 [5] and is another tumour suppressor, not quite so well known as its partner \textit{BRCA1} but arguably just as important. The complex formed by the resulting two proteins has ubiquitin ligase activity [6], and cellular functions involving both \textit{BRCA1} and \textit{BARD1} include DNA repair [7, 8], cell cycle progression by means of centrosome regulation [9, 10], and general genetic stability [1, 8, 10]. \textit{BARD1} has an important role in controlling \textit{BRCA1} function and localisation [1, 8, 11, 12], and is independently involved with inhibition of mRNA maturation [13, 14], contact inhibition of cell growth [15], apoptosis by means of p53 stability [16], and repression of cellular oncogenic features [15].
1.2.1 Gene structure

*BRCA1* is located on chromosome 17 (17q21), and is 81 kilobases (kb) long [17]. *BRCA1* has 23 exons with exon 11 being one of the longest coding exons in the genome (3kb) [18, 19]. The full length transcript is 7.2 kb [20]. *BARD1* is located on chromosome 2 (2q35), and is 84kb long [17]. It has 11 exons, and the length of the full length transcript is 2.6 kb [20]. Figure 1.1 shows a schematic of these genes with functional areas annotated.

BRCA1 has three main domains that have been identified with specific functions (Figure 1.1a). At the N-terminal of the protein, encoded by exons 2 to 7, is the RING (Really Interesting New Gene) domain, a finger-like domain essential for BARD1 binding [3, 5, 21]. In the middle lies the region encoded by exons 11-13 that contain binding sites for γ-tubulin [22, 23] and many proteins involved in cellular pathways associated with cancer [3, 24]. This region also contains nuclear localisation signals (NLS) [25, 26] and nuclear export signals (NES) [27, 28]. At the C-terminal lies the BRCT (BRCA1 C-Terminus) domain, encoded in exons 16 to 24, a mainly phosphoprotein binding region that also has DNA binding function and some non-phosphoprotein interactions [29]. RING and BRCT domains are also present in BARD1 (Figure 1.1b), encoded by exons 1 to 3 and exons 9 to 11 respectively [1]. BARD1 also has Ankyrin (ANK) repeats encoded by exons 4 to 7, which are involved with protein binding, particularly to the prominent tumour suppressor p53 [1]. BARD1 has fewer mutations than *BRCA1*, lending to the theory that BARD1 is of higher functional importance than *BRCA1* because it has less tolerance for mutation [1].

1.2.2 Cellular functions of BRCA1 and BARD1 proteins

**Ubiquitin Ligase activity:** The BRCA1-BARD1 complex acts as an E3 Ubiquitin ligase, which catalyses Lys-6-linked polyubiquitin chains on itself (auto-ubiquitination) and on other proteins [6, 30]. Lys-6 linkage is not a degradation signal like more common ubiquitin linkages, and though some research points to it having some involvement with DNA repair, [30, 31] this theory has been extrapolated from the repair functions of BRCA1/BARD1, and has not been explicitly proven.

**DNA damage response:** The BRCA1-BARD1 complex functions in DNA damage response, particularly evidenced by its actions after UV radiation damage [31, 32]. BARD1

---

1 *BRCA1*’s exons are numbered 1 – 24 because exon 4 is no longer classified as an exon and is omitted.
instigates localisation to the nucleus, enabling BRCA1 to facilitate the repair of double stranded breaks, thereby stabilising chromatin structure [7, 11, 32].

**Cell cycle regulation:** The BRCA1-BARD1 complex interacts with the tumour suppressor Retinoblastoma protein (RB) by means of regions encoded within BRCA1 exon 11, causing cell cycle arrest through promotion of RB activity [3, 33]. BRCA1 also halts cell cycle progression through binding to and inhibiting c-Myc, an oncogene which promotes transcription of a large portion of the genome [34, 35]. The complex has also been associated with regulation of centrosomes in a dividing cell, as division is no longer controlled when the BRCA1-BARD1 complex is absent or non-functional [36-39]; however, this fact by itself has been postulated to have nothing to do with chromosome instability [38]. Each protein appears to individually provoke an apoptotic response when localised to the cytoplasm, and a cell survival response when localised to the nucleus [4, 40, 41].

Each of these cellular functions demonstrates the importance of these two genes in maintaining cellular integrity. The consequences of the breakdown of these processes can in time lead to the development and progression of a tumour, through lack of DNA repair activity or promotion of uncontrollable cell growth and division.
Figure 1.1 BRCA1 (a) and BARD1 (b) exon structure and areas coding for major functional domains. BRCA1 exon 4 is not shown, as it is no longer classified as an exon. Nuclear localisation and export sequences are dispersed throughout each gene, usually close to the functional domains indicated. RING domain refers to the finger-like domain typical of protein binding regions. BRCT domain refers to a region typical of cell cycle regulatory proteins, named for the BRCA1 C-terminus. This original figure is not to scale.
1.3 Pre-mRNA processing and final mRNA structure

Messenger RNA (mRNA) is copied from the DNA template and is translated into protein by the ribosome [42]. Non-coding types of RNA, such as transfer RNA (tRNA), and ribosomal RNA (rRNA), assist in this process by matching each mRNA codon to the appropriate amino acid (tRNA) and forming a crucial part of the ribosomal machinery (rRNA) [43-45]. Still more types of non-coding RNA include microRNA (miRNA), small nuclear RNA (snRNA) and many others that contribute to many biological systems [46].

The precursor mRNA (pre-mRNA) transcript goes through many modifications before it is translated into a protein. Most modifications occur co-transcriptionally, while the strand is still being produced by RNA polymerase (RNAP) [47-52]. To prevent degradation a 7-methylguanosine nucleotide is added to the 5’ end (a process called “capping”) immediately upon exit of the beginning of the transcript from the polymerase, and a 3’ poly(A) tail is added to the other end once the full molecule has been produced [53]. Both the 5’ cap and 3’ poly(A) tail are also involved with nuclear export, translation, transcription termination (poly(A) tail only), and splicing out of introns (5’ cap only) [54-57].

Degradation of mRNA occurs by a process called nonsense-mediated decay (NMD) [58]. This pathway’s main purpose is to initiate the degradation of mRNA transcripts that have a premature stop codon. NMD has also been identified as playing a role in the regulation of gene expression [58]. Inhibitors of NMD, such as cycloheximide, have been used to track mRNA transcript expression when variants are likely to result in mRNA degradation [59]. Inhibiting NMD is an important step during analysis, as transcripts that would normally need to be degraded by NMD may be upregulated in disease if they are pathogenic, and characterising these transcripts is important to further our understanding of disease mechanisms.
1.3.1 Alternative Splicing

The structure of genes on mammalian DNA necessitates the splicing out of introns, non-coding stretches of gene, from the mRNA transcript. This leaves only exons, the coding portions of the gene that remain to be translated. However, every exon is not always included in the final mRNA, (exon exclusion) and occasionally parts of introns have been found to be included in the final transcript (intron inclusion) [60, 61]. This phenomenon is known as alternative splicing, and it means that many mRNA isoforms can be present in a cell or tissue, generating different protein isoforms (Figure 1.2). Variation in factors that influence splicing is therefore integral to the flexibility of protein function in cells and tissues [53].

Alternative splicing is facilitated by both cis-acting factors (consensus sequences around splice sites within exons and introns), and trans-acting factors (splicing proteins) [53, 56]. Cis-acting exon splice enhancers/silencers (ESE/S) and intron splice enhancers/silencers (ISE/S) [53], along with the sequence of bases at the exon/intron boundary, influence where splicing occurs and are demonstrated in Figure 1.3. In particular, NAGNAG repeats at the 5’ end of an exon (acceptor site) give the option of a second splice site located three bases away from the initial one, i.e. after the first “NAG” or the second "NAG". Sequences at the 3’ end of the exon (donor site) may also have characteristic motifs. Most of the proteins involved in splicing typically form a complex
called the spliceosome, which recognises the splice sites and enables splicing (Figure 1.3) [42, 53, 62]. The splice site consensus sequence is recognised by small nuclear ribonucleoproteins (snRNPs) that compose the part of the spliceosome complex that recognises the sequence [51, 63]. The consensus sequence may be strong or weak, and as such has differing affinities for the recognition snRNPs, contributing in part to the decision as to which parts of the RNA transcript are spliced out [64, 65]. The complexity of alternative splicing, demonstrated in Table 1.1, allows for numerous possibilities for protein variation within and between tissues [60], and draws attention to the considerable impact alterations in splicing mechanisms have on cell and tissue function.

Splicing is subject to many factors, from environmental changes like exposure to ultraviolet radiation [66, 67], to inherent variation in sequence and protein expression, to chromatin structure due to transcription coupled splicing [47, 49-51, 68-70]. The formation of secondary structures by the RNA strand during transcription also influences access of splicing factors to the consensus sequence [71, 72], and all these processes are subject to the speed of the RNA polymerase as it transcribes the gene. Studies have shown that the slower the gene is transcribed, the more likely certain exons will be included in the final transcript [73-77]. With so many factors contributing to splicing, it is extremely likely that variations in sequence causing aberrant alternative splicing or regulation of splicing will contribute to the development and progression of a significant number of cancers.

Figure 1.3: Cis- and trans- acting factors involved in alternative splicing of exons. Green: factor that facilitates splicing, Red: factor that prevents splicing. Dashed arrows show where trans factors interact with cis factors. NAGNAG splice sequence on 5’ end of exon, and an indicator of the position of the 3’ splice sequence are shown. Original figure.
Table 1.1: Alternative splice type identified with *BRCA1* transcript analysis [60].*

<table>
<thead>
<tr>
<th>Type of splicing event</th>
<th>Description</th>
<th>Diagram</th>
<th>No. of <em>BRCA1</em> events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cassettes</td>
<td>Exon excluded or non-constitutive exon included (usually intronic sequences becoming part of the mature transcript)</td>
<td><img src="image1" alt="Diagram" /></td>
<td>17</td>
</tr>
<tr>
<td>Multi-cassettes</td>
<td>More than one exon excluded or non-constitutive exon included</td>
<td><img src="image2" alt="Diagram" /></td>
<td>20</td>
</tr>
<tr>
<td>Splice donor shifts</td>
<td>Alternative use of NAGNAG(^b) sites or (rarely) use of another site</td>
<td><img src="image3" alt="Diagram" /></td>
<td>4</td>
</tr>
<tr>
<td>Splice acceptor shifts</td>
<td>Alternative use of acceptor sites</td>
<td><img src="image4" alt="Diagram" /></td>
<td>4</td>
</tr>
<tr>
<td>Terminal modifications</td>
<td>Premature termination codon</td>
<td><img src="image5" alt="Diagram" /></td>
<td>2</td>
</tr>
<tr>
<td>Intronisations</td>
<td>An internal part of an exon is spliced out</td>
<td><img src="image6" alt="Diagram" /></td>
<td>2</td>
</tr>
<tr>
<td>Donor shift + multi-cassette</td>
<td>Both splice donor shift and multi-cassette event are present</td>
<td><img src="image7" alt="Diagram" /></td>
<td>4</td>
</tr>
</tbody>
</table>

\(^a\) Diagrams modified from Colombo *et al.* (2014) [60].

\(^b\) NAGNAG sites are tandem exon acceptor sites that allow splicing either after the initial "NAG" or after the whole "NAGNAG" sequence.

*Table continued on following page
<table>
<thead>
<tr>
<th>Type of splicing event</th>
<th>Description</th>
<th>Diagram</th>
<th>No. of (BRCA1) events</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multi-cassette + acceptor shift</strong></td>
<td>Both a multi-cassette and acceptor shift event are present</td>
<td><img src="image1" alt="Diagram" /></td>
<td>4</td>
</tr>
<tr>
<td><strong>Terminal + multi-cassette</strong></td>
<td>Both terminal modification and multi-cassette event are present</td>
<td><img src="image2" alt="Diagram" /></td>
<td>3</td>
</tr>
<tr>
<td><strong>Cassette + multi-cassette</strong></td>
<td>Both a single cassette event and a multi-cassette event are present</td>
<td><img src="image3" alt="Diagram" /></td>
<td>1</td>
</tr>
<tr>
<td><strong>Donor shift + acceptor shift</strong></td>
<td>Both a donor shift and acceptor shift are present</td>
<td><img src="image4" alt="Diagram" /></td>
<td>1</td>
</tr>
<tr>
<td><strong>Donor shift + cassette + multi-cassette</strong></td>
<td>Donor shift, cassette event and multi-cassette event are all present</td>
<td><img src="image5" alt="Diagram" /></td>
<td>1</td>
</tr>
</tbody>
</table>

\(^a\) Diagrams modified from Colombo *et al.* (2014) [60].
1.3.2 BRCA1 and BARD1 isoforms and the ENIGMA Splicing Working Group
The ENIGMA Consortium (Evidence-based Network for the Interpretation of Germline Mutant Alleles) is a multinational group dedicated to the study of breast and ovarian cancer associated genes [78]. ENIGMA members have published numerous papers related to guidelines associated with identifying sequence variants and associated splicing changes to aid clinical classification of such variants [79-83], the identification of the alternative splicing of BRCA1 and BRCA2 [60, 84, 85], and compilations of variant and splice data [80, 81, 86]. Some of the recommendations on splicing analysis that have arisen from these publications include: sequencing all products, using data from at least 10 controls, and utilising comparisons between data and reference datasets [80, 81]. Mini-gene assays are a popular way of determining the influence of variations in splice sites on alternative splicing of a particular area of a gene [18, 19, 87-89]; however, they are only able to functionally assay parts of the gene and not the full isoform (mini-gene methods of analysis are discussed in Section 1.4). This is not a viable method of determining exon connectivity across the whole transcript in a biological system, and at present, there is an inability to measure exon connectivity over longer distances, particularly with features like the 3kb exon 11 in BRCA1.

Colombo et al. (2014) [60] in collaboration with the ENIGMA consortium has published a comprehensive study on naturally occurring BRCA1 splice variants and many other studies have included analysis of BARD1 isoforms [16, 90-101]. Colombo et al. identified 63 alternative splicing events, defined by the splice junctions identified, only 28 of which had been previously published. Interestingly, the study found that every internal exon was alternatively spliced at least once, suggesting that BRCA1 does not have constitutive exons. This study also demonstrated the extent of transcript variation, demonstrated in Table 1.1 by the number of each type of BRCA1 splicing event detected. The authors did note that though this is the most in-depth analysis of BRCA1 alternative splicing events to date, they were not able to identify full length mRNA isoforms present (they did not perform cloning analysis), nor their level of expression. This means that linked splicing events (where multiple exons far apart are always spliced out together), that generate a particular isoform in high abundance, are uncertain in BRCA1 [60]. There are currently 46 published isoforms of BARD1 [16, 90-102], some of which have been validated by sequencing.
For both \textit{BRCA1} and \textit{BARD1} evidence points to full length transcripts being protective, and various isoforms tending to have oncogenic-like activity [16, 60, 90-96, 98, 103], such as BRCA1-IRIS, a product encoded by an open reading frame extending from codon 1 to a termination point a short way into intron 11 [104-106]. The oncogenic activity is apparently due to the lack of functional domains encoded by particular exons that are spliced out, such as the RING domains, nuclear localisation and export sequences, ANK repeats and BRCT domains (Figure 1.1) Many of the \textit{BARD1} isoforms are missing exon 4, which constitutes a large portion of the full length transcript but has no associated annotated functional regions of protein. Some more well-known isoforms of \textit{BARD1} have been shown to have functional importance, such as BARD1β. This isoform is missing exons 2 and 3, which encode a RING domain, and as such is known for having a strong oncogenic effect when highly expressed [90, 95], mainly due to the lack of interaction with BRCA1. There are very few isoforms that have been characterised that have not been detected in normal tissue [16, 90-101].

Splice isoforms play an important role in the pathogenesis of diseases, especially cancer, and should be treated as clinically relevant. It has already been suggested that knowledge of splice isoforms can help develop treatments [103], and this is a key reason for characterising and classifying isoforms of high risk genes such as \textit{BRCA1} and \textit{BARD1}. Sequence variation such as point mutations in regulatory regions such as enhancers and silencers, as well as those in splice sites themselves, play a significant role in determining the splicing pattern of individual genes [18, 19], and as such have functional and pathological significance [79, 107]. The classification of this genetic variation in high risk genes is therefore important in clinical decision making, as these variants, through their influence on splicing, can have a significant effect on disease development and progression [82]. Inclusion of mRNA splicing assays into the multifactorial risk assessment of variants is therefore essential for classifying many variants of unknown clinical significance in the \textit{BRCA1} gene [79]. Correctly identifying these pathogenic variants is essential for the clinical evaluation and risk assessment of a patient with breast/ovarian cancer [79, 84]. In fact, loss of function in \textit{BARD1} is predominantly through splicing changes, not mutation of coding sequence [82]. Determining the extent of naturally occurring isoforms in \textit{BRCA1} and \textit{BARD1} is essential to understand the pathological implications of alterations in the splicing landscape of these genes. Successful variant classification, however, is difficult due to the length of time it takes to perform sufficient functional assays on an individual variant to
determine its pathogenicity [108]. The consistency of annotation across research groups and databases needs to improve in order to provide adequate clinical care across all disease types [86], though ENIGMA displays expertise in this area.

1.3.2.1 Splice event nomenclature
Isoforms have a notation style associated with their naming which is followed in this report, detailed in Table 1.2.

Table 1.2: Nomenclature of mRNA isoforms

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
</table>
| Δ      | mRNA is missing a part of the full length transcript | Δ9 - Exon 9 is skipped  
Δ9,12 – exons 9 and 12 are skipped  
Δ9-12 – exons 9, 10, 11 and 12 are skipped  
11Δ3 110 – 3110 bases from inside exon 11 are spliced out (intronisation) |
| ▼     | mRNA retains a part that is not normally included in the full length transcript | ▼21 – retention of part of intron 21. |
| Δ_p   | Bases are missing from the 5’ end (acceptor site) of an exon | Δ5p – bases are missing from the 5’ end of exon 5.  
Could be any number of bases, identified by A or B etc., as described where discovered. |
| Δ_q   | Bases are missing from the 3’ end (donor site) of an exon | Δ5q – bases are missing from the 3’ end of exon 5.  
Could be any number of bases, identified by A or B etc., as described where discovered. |

1.3.3 Other mRNA modifications
There are other types of post-transcriptional modifications that have been found to occur in mammalian systems, and so there is the possibility that some might be applicable to BRCA1 or BARD1. Exon shuffling, including the production of circular RNA, is a variation of alternative splicing (and an instigator of genomic variation [109]), that is generally accepted to be found in human cells, particularly neurons [110, 111]. Exons in transcripts are not always in the order they appear on the DNA strand, indicating a splicing event that results in a 3’ donor splice site interacting with a relatively 5’ acceptor splice site upstream of the donor site. This can result in circular RNA, with intervening introns either removed or retained, but this pattern can also be found in linear RNA. Other methods of exon shuffling involve trans-splicing [110, 112], where exons from different transcripts, genes or even chromosomes are spliced together. So far no definitive instances of BRCA1 or BARD1 exon shuffling have been identified.

Micro-exons have also been characterised in genes involved in neuronal and cardiac pathways [113-115], and have functional consequences; however, none have yet been found in BRCA1 or BARD1. Micro-exons are typically defined as coding regions ranging
from 3-25 nucleotides in length, though some literature states the upper size limit as 51 nucleotides [114-117]. Micro-exons are hard to detect using most alignment programs, though two groups have succeeded in developing reliable algorithms [113, 117]. Given this information, and the depth of high-throughput short-read mRNA sequencing data analysis (RNA-seq) on the genes in question, it is unlikely that *BRCA1* or *BARD1* contain undiscovered micro-exons.

Another type of modification process is that of RNA editing [118, 119]. These include several types of base modifications. Adenosine and cytosine methylation involve a methyl group addition to these respective bases. Pseudouridine is formed when pseudouridine synthase “flips” uridine into an isomeric form attached to the ribose [120]. Pseudouridine is abundant in mammals, and has been identified as particularly prevalent *in vitro* under a stressed state [119-121]. RNA editing also includes the deamination of cytosine to uracil and of adenine to inosine, which occurs mainly in the brain [119-122]. Some RNA editing has been detected in *BRCA1* transcripts in the intronic regions [123].

These types of modification seem to be relatively rare in comparison to splicing variation. However, this could be because of a lack of reliable technology for this type of investigation.

### 1.4 Methods of detecting exon structure

The need for a reliable detection method of full length isoforms has grown out of the large body of evidence that differing isoforms from a single gene can have many different functions within a cell or tissue [53, 84, 111, 115, 124-126]. For example, transcripts of *BARD1* have been identified that lack the exons encoding the RING domain, or the ANK repeat domains [16, 90-101]. Protein products of these transcripts have been identified, and are associated with poor outcomes in breast cancer patients [16, 90]. It is essential that there is knowledge of naturally occurring isoforms and their associated expression levels to allow successful clinical decision-making on the basis of a patient’s genetic profile. At present the studies surrounding normal *BARD1* isoform expression have come together slowly [16, 90-101], and are less comprehensive than those for *BRCA1*, as explained in Section 1.3.2 [60].

One of the first methods of transcript analysis was the Northern blotting technique, which gives a size estimation and a rough quantification; more recent methods, though,
are easier to use and give more accurate information. Typical methods for detecting transcripts presently include cloning of cDNA into bacterial plasmids and Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR) coupled with electrophoresis or sequencing [80]. High-throughput RNA-sequencing (RNA-seq) is a relatively new method of isoform analysis that is excellent for quick and accurate identification of exon splicing events present in a sample. In their analysis of BRCA1 splicing events, Colombo et al. (2014) used all of the above analysis techniques to characterise these isoforms [60].

1.4.1 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)

RT-PCR is used to amplify complementary DNA (cDNA), produced from mRNA, to create a product requiring further analysis [127]. Reverse transcriptase converts RNA into cDNA, which is more stable and easier to work with. PCR is performed on the resulting cDNA to generate a product that is further analysed. The PCR product is initially applied in agarose gel electrophoresis or capillary electrophoresis for size analysis. This can give valuable information about the length of the transcript that can be used to support evidence of a particular isoform being present. Further analysis can be done by Sanger sequencing, in which a single band is cut out of the agarose gel, or the PCR reaction is purified, before sequencing. A major limitation of analysis by size is the possibility of two different transcripts of similar size being present in the same band; this also presents problems for sequencing from cut-out gel bands, as a single product is needed for reliable Sanger sequencing. This method was used extensively during this research project, coupled with agarose gel electrophoresis and both Sanger sequencing and MinION sequencing.

1.4.2 Molecular Cloning

The RT-PCR products can also be cloned into a plasmid and cultured in bacteria. The resulting colonies on a plate can then be picked for sequencing (typically Sanger sequencing), and will usually contain just one cDNA each. It is recommended to pick at least 40 clones off a plate to be individually sequenced [80], in order to gain a reasonably good representation of the isoforms present in a sample. This procedure is very laborious and time consuming, and also relies on chance (that colonies containing all the isoforms between them were picked). There may also be unforeseen bias in the cDNA transcripts that are successfully transformed into plasmids.
1.4.3 Short read RNA-Sequencing (RNA-seq)

RNA-seq is a high throughput sequencing method that does not always require PCR [128]. It relies on a fragmentation method in which cDNA is produced then fragmented to 200-500 base pair lengths. Sequencing ranges from 30 to 400bp at a time, either using single-end or paired-end reads (reading a section at each end of a single molecule). The depth and speed of analysis is excellent, and can give reliable data on exon junctions and connectivity of a few exons at a time using paired-end reads. It is very accurate, and as such can detect sequence variations. However, the nature of short sequence reads does not allow for analysis of a full transcript at once. There is also bias depending on the type of fragmentation, either towards the transcript body or towards the 3’ end. RNA-Seq still lacks the ability to identify splicing events between distant exons, and Wang, Z. et al. [128] make a point that longer reads are required for full characterisation of complex genomes.

1.4.4 Other methods

For splicing analysis, reporter mini-gene assays can be used to analyse individual genes, in both wild-type and pathologically mutated forms. This method was used extensively when characterising the splicing of exon 11 of BRCA1 [18, 19], and has also been used for other portions of the gene [87]. A mini-gene construct is a small section of a gene, usually containing several exons cloned into a vector such as a plasmid or bacterial artificial chromosome. This assay enables detailed analysis of splicing regions, regulatory elements and alternative splicing patterns of that section of the gene. This method does not look at splicing events across the whole gene at once.

Synthetic long read sequencing is a process that has been developed to resolve the problem generated by high-throughput sequencing of multiple short sequence reads that are difficult to piece together [129]. The sample is essentially made so dilute when it is distributed across a multi-well plate that only one cDNA per well is sequenced. Though this eliminates the initial problem, it is a relatively inefficient, expensive and cumbersome technique.

CAGE and RACE are methods for characterising 5’ ends, which are often difficult to characterise due to reverse transcriptase failing to extend the full length of the mRNA template because of mRNA length and secondary structure. CAGE is Cap Analysis Gene Expression [130], RACE is Rapid Amplification of cDNA Ends [131], and both focus on amplifying the 5’ end of transcripts (and 3’ as well in the case of RACE), using PCR and
restriction enzymes. These techniques are still limited by inherent PCR limitations, such as high GC content and polymerase processivity.

1.4.5 Limitations of current methods
RT-PCR combined with gel or capillary electrophoresis is not sufficient by itself to distinguish between transcripts of similar size. PCR also relies on primer design, which usually spans junctions and possible junctions to detect exon skipping, not whole transcripts. Sanger sequencing performed on cut-outs from an agarose gel do not give a good result if there is more than one transcript present of a similar size; for good Sanger sequencing the template DNA must be reasonably pure, i.e. one transcript per band. As for Sanger sequencing on cloned products, every bacterial colony needs to be sequenced individually. The whole process of transformation, transfection and bacterial culture is laborious, time consuming, and relies on chance to detect all transcripts, both at the transfecting stage and at selection of colonies. Sanger sequencing of cloned products is currently the gold standard method of isoform validation [60], and it is relatively expensive for the amount of data it produces. RNA-seq is excellent at detecting individual splice events but cannot accurately determine the whole transcript because of the fragmentation steps. RNA-seq methods based on short read sequencing are unreliable for full length isoform identification [132]. Though transcripts combining multiple splicing events can be statistically imputed through analysis of RT-PCR data, and indeed this was utilised in the BRCA1 isoform study by Colombo et al. (2014) [60], validation of these transcripts is still required for confirmation of these isoforms.

1.5 Single molecule long-read DNA sequencing
Currently, long read platforms such as the Pacific Biosciences system (PacBio) and the Oxford Nanopore Technologies (ONT) MinION have been employed for a small number of papers looking at cDNA, and even fewer on isoform detection. Neural synapse formation, known for its complexity, has been examined using PacBio and this study was successful in identifying novel transcripts of Neurexins in mice [111]. The isoforms of the most complicated spliced gene currently known, Dscam in Drosophila, were partially characterised in the MinION platform, proving that it is possible to characterise transcript isoforms of particular genes. These types of analysis have yet to be done on human genes.
1.5.1 PacBio Sequencing
The PacBio system utilises Single Molecule Real-Time (SMRT) Sequencing to generate long sequence reads. The current version has an average read length over 10kb and maximum read lengths over 60kb [133, 134]. SMRT sequencing involves tracking the activity of DNA polymerase as it synthesises a complementary strand with fluorescently labelled nucleotides. The polymerase is immobilised to the base of a tiny well that is smaller than the wavelength of light, allowing it to detect the nucleotide’s fluorescence with no background. Adapters are attached to either end of the DNA molecule to be sequenced which enables it to be circularised. This means the DNA polymerase can keep sequencing the same strand many times, generating a circular consensus sequence (CCS), which gives higher accuracy than a single read of the strand [133, 134]. Consequently, this means that the longer the DNA strand is, the less accurate the sequence is.

The PacBio system has so far been used successfully for completing many bacterial genomes, its long reads solving some of the problems short read sequencers have not been able to address. The PacBio system has been used for RNA transcript analysis in conjunction with short sequencing reads, a process called hybrid sequencing. This utilises the PacBio sequence as a scaffold for aligning the more accurate short sequence reads, and has been successful at identifying full splice isoforms of the three neurexin genes [111], as well as characterising the human embryonic stem cell transcriptome [132]. The authors noted that their results on the transcriptome indicate that human gene identification is probably incomplete.

1.5.2 MinION sequencing
The Oxford Nanopore MinION is a nanopore sequencing device that is about the size of a cellphone. It works by passing DNA through a tiny pore (nanopore) set into a membrane with an electrical current across it. As each base passes through the pore it causes a unique disruption in the electrical current, generating a signal characteristic of each individual nucleotide. Before application to the MinION flow cell, adapters are ligated to the ends of the DNA strand, with a hairpin adapter linking the two strands together. The lead adapter guides the DNA to the nanopore, where a helicase enzyme separates the two strands [135]. The hairpin connects the template and complement strands so that after the first strand (template) has passed through the pore, the second strand (complement) follows. Though each strand can be analysed individually (1D
analysis), ideally both strands are analysed together, generating a “2D” read [136]. This is generally more accurate than the single sequence of either the template or the complement alone. The electrical signal generated is converted into a .fast5 file (HDF5 file format) by the MinKNOW program that runs the MinION [137], and these files are typically sent to Metrichor over the internet, a software program based at Oxford Nanopore Technologies. Metrichor basecalls the sequence files and sends it back, and the entire process can occur within minutes. The MinION has been known to produce reads over 100kb in length, though an average read tends to be 10-20kb. This length is largely limited by fragmentation of strands during library preparation, and not by the device itself.

The MinION is a relatively new sequencing technology, and has been available only through an early access program (the MinION Access Program, known as the MAP community) while in development. This means there is a limited number of publications related to MinION use. The MinION has been used to sequence pathogens such as the Ebola virus on the ground in West Africa during the Ebola outbreak [138], and also in sequencing bacterial pathogens such as Staphylococcus aureus and Mycobacterium tuberculosis in order to determine differing strains and predict antibacterial resistance [139]. It has been valuable in completing genomes, being able to sequence through repetitive regions [140]. It is particularly useful in long range connectivity within the genome and transcriptome, being able to sequence long haplotypes [141] and RNA transcripts [142, 143]. Given the broad range of applications, there is an even wider range of possible analysis methods. Research groups across the MAP community, including ONT itself, are working on software and algorithms that are suitable for long read sequence data produced with the unique .fast5 file format and that can handle the relatively high error rate of the system.

1.5.3 Comparison of PacBio and MinION Platforms

PacBio sequencing and MinION sequencing are based on two completely different methods of long-read sequencing. At present, the average read lengths of both these systems are more or less equivalent, but in the future this may change; the length of PacBio sequencing is dependent on the life of the polymerase, and can only sequence up to ~60kb at the time of writing, whereas theoretically the MinION has no such biological limitation. Due to the design of the MinION, the DNA strand can be sequenced a maximum of twice (the template, then the complement strands), whereas the
circularisation of DNA in the PacBio system allows it to be sequenced as many times as the lifetime of the polymerase allows. As a result, the PacBio system is typically more accurate than the MinION system, though the longer the read, the more the difference is reduced. The capital cost of the PacBio system is huge at $1 million, while in comparison the MinION cost is extremely low (at present the cost is only for the consumable flow cells at ~NZ$1400 and sequencing kits). This low cost and portability makes the MinION a far more realistic sequencing option for smaller laboratories.

1.6 Aims and Hypotheses
The overall aim of this research was to examine the ability of a novel DNA sequencing device, the MinION, to identify mRNA splicing patterns of full length exon transcripts derived from the human genes BRCA1 and BARD1. The splicing patterns in BRCA1 are very complex and are clinically relevant, so the purpose of this project was to test whether the MinION could identify known and novel isoforms. BARD1 has a shorter transcript length than BRCA1, has less complex splicing patterns and is functionally related, so was chosen as a more tractable target in case BRCA1 analysis was unattainable.

There were two main hypotheses:

1. That the MinION can accurately and comprehensively generate whole transcript data to determine isoform structure.
2. That MinION data will uncover novel splicing patterns in the BRCA1 transcript and the BARD1 transcript.

This study explores these hypotheses by using RNA obtained from a lymphoblastoid cell line. The cDNA was generated from the total RNA, amplified by PCR, and run on the MinION. Data were analysed using specific bioinformatic and computational methods.
Chapter 2: Materials and Methods

2.1 General stocks

2.1.1 TAE buffer (Tris/Acetic acid/EDTA)
A stock solution of 10x TAE buffer was prepared with MPW to 1L. This was made with 48.4g of Tris (Roche Diagnostics, Germany), 3.7g of EDTA (Ajax Finechem, ThermoFisher Scientific, Auckland, NZ), and 11.4mL of glacial acetic acid (Fisher Scientific, UK).

2.1.2 TE buffer (Tris/EDTA)
TE buffer was prepared by adding 1mL of 1M Tris Base (Roche Diagnostics, Germany), and 200µl of 0.5M EDTA (Ajax Finechem, ThermoFisher Scientific, Auckland, NZ), to a final volume of 100ml with MPW. The final concentrations of each were 10mM of Tris Base, and 1mM EDTA.

2.2 RNA source
RNA had already been extracted from a normal human lymphoblastoid cell line treated with the NMD inhibitor cycloheximide prior to the beginning of this project [102]. The concentration of extracted RNA was measured as 620ng/µl on a Nanodrop 8000 spectrophotometer (Bioline, ThermoFisher Scientific).

2.3 Primer Design
The web program Primer3 (version 0.4.0) was used for designing all primers. The design parameters consisted of a minimum primer Tm of 58.5°C, optimal Tm of 60.0°C, and maximum Tm of 61.5°C. The minimum GC% was 20%, optimal was 50% and maximum was 80%. The ideal product length range was adjusted according to the target region for each pair of primers, and all other parameters remained as the default value. Table 2.1 and Table 2.2 describes these primers and their purpose. Figure 2.1 displays where Table 2.1 primers are positioned on the genes. Figures demonstrating placement of primers in Table 2.2 can be found in Appendix A. Primers (Integrated DNA Technologies, Singapore) were reconstituted in TE buffer to a concentration of 100µM. Working solutions of 10µM and 5µM (Table 2.1 primers and Table 2.2 primers respectively) were made from these stocks using Millipore water (MPW). Working solutions were stored at 4°C while stock solutions were stored at -20°C.
Table 2.1: Primers for cDNA synthesis, initial PCR and MinION sequencing, see Figure 2.1 for position on genes.

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer Sequence</th>
<th>Tm (°C)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1_1F</td>
<td>GCCGGGAATTACAGATAAAA</td>
<td>68.1</td>
<td>BRCA1 Forward primer for full length PCR and for 5’ end confirmation.</td>
</tr>
<tr>
<td>BRCA1_5R</td>
<td>GAAGGCCCTTTCTCTGTT</td>
<td>69.8</td>
<td>BRCA1 Reverse primer for 5’end confirmation.</td>
</tr>
<tr>
<td>BRCA1_20F</td>
<td>GAAGAAACCACAAAGTCCA</td>
<td>69.6</td>
<td>BRCA1 Forward primer for 3’ end confirmation</td>
</tr>
<tr>
<td>BRCA1_24pR</td>
<td>AAGCTCATTCTGGGTCCT</td>
<td>71.0</td>
<td>BRCA1 Reverse primer for full length PCR and for 3’ end confirmation.</td>
</tr>
<tr>
<td>BRCA1_24qR</td>
<td>AGTCTTCACTGCCCTGCAC</td>
<td>72.1</td>
<td>Synthesis of cDNA, at 3’ end of exon 24, close to polyA tail.</td>
</tr>
<tr>
<td>BARD1_1F</td>
<td>CTGACCAGCTGGAGAAG</td>
<td>70.8</td>
<td>BARD1 Forward primer</td>
</tr>
<tr>
<td>BARD1_11R</td>
<td>CTGGCTTGGCTTTTACTG</td>
<td>70.0</td>
<td>BARD1 Reverse primer</td>
</tr>
</tbody>
</table>

\(^a\) Naming refers to the gene name, exon number, and orientation.

\(^b\) Tm was calculated using the conditions detailed in Standard PCR for short products, using the Integrated DNA Technologies (IDT) tool OligoAnalyzer 3.1.  
[https://sg.idtdna.com/calc/analyzer](https://sg.idtdna.com/calc/analyzer)
Figure 2.1: Primer positions for those described in Table 2.1. Primers are shown to be designed to regions within the exons indicated in *BRCA1* (A) and *BARD1* (B). This original figure is not to scale.
Table 2.2: Primers for confirmation of isoforms. See Appendix A for positions on isoforms.

<table>
<thead>
<tr>
<th>Name</th>
<th>Primer Sequence</th>
<th>Tm (°C)</th>
<th>Targeted Isoform</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Junction primers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA1_9-18FJunc</td>
<td>AGGCAAATTTATTCAGATGC</td>
<td>67.2</td>
<td>Δ10-17</td>
</tr>
<tr>
<td>BRCA1_23RJunc</td>
<td>ATGGAAGCCAATTGTCCCTTG</td>
<td>69.6</td>
<td>Δ10-17</td>
</tr>
<tr>
<td>BRCA1_11α-12FJunc</td>
<td>ATCCAGAAAATATCGGTGAAGCC</td>
<td>70.4</td>
<td>Δ11q,19, Δ11q, ▼21, and Δ11q,14,21</td>
</tr>
<tr>
<td>BRCA1_18-20RJunc</td>
<td>CTCTGACTTCAAATATCAGAATAG</td>
<td>68.3</td>
<td>Δ11q,19</td>
</tr>
<tr>
<td>BRCA1_2-5FJunc</td>
<td>GAGTGTCCTCCATCTGATTTTGC</td>
<td>68.7</td>
<td>Δ3,11b, and Δ3,9,10,11q</td>
</tr>
<tr>
<td>BRCA1_11α-11cRJunc</td>
<td>GTCAGGTCTTCCAATTCACCTGAT</td>
<td>70.5</td>
<td>Δ3,11b</td>
</tr>
<tr>
<td>BRCA1_ins21RJunc</td>
<td>CTCTGACTTCAAATTCACCTGAT</td>
<td>70.8</td>
<td>Δ11q, ▼21</td>
</tr>
<tr>
<td>BRCA1_20-22RJunc</td>
<td>CATCCAGTTGATCTTTCTGTC</td>
<td>68.8</td>
<td>Δ11q,14,21</td>
</tr>
<tr>
<td>BRCA1_11α-12RJunc</td>
<td>CAGATGTCCTTCCAATTCACCTGATAC</td>
<td>70.3</td>
<td>Δ3,9,10,11q</td>
</tr>
<tr>
<td><strong>Flanking primers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>del10-17FFlank</td>
<td>TCAGCTTGACACAGTTTG</td>
<td>70.1</td>
<td>Δ10-17</td>
</tr>
<tr>
<td>del10-17RFlank</td>
<td>TGGACCTTTGCTGTTTCTTC</td>
<td>69.6</td>
<td>Δ10-17</td>
</tr>
<tr>
<td>del11qF_Flank</td>
<td>CTCAAGGAACCCAGGGATGAA</td>
<td>68.9</td>
<td>Δ11q,19, Δ11q, ▼21, and Δ11q,14,21</td>
</tr>
<tr>
<td>del19R_Flank</td>
<td>GTTGATCTGTGGGCATGTTG</td>
<td>68.9</td>
<td>Δ11q,19</td>
</tr>
<tr>
<td>del3F_Flank</td>
<td>GCCCGGGAATTACAGATAAA</td>
<td>68.1</td>
<td>Δ3,11b, and Δ3,9,10,11q</td>
</tr>
<tr>
<td>del11bR_Flank</td>
<td>TCACCTCTCACACCCAGATGC</td>
<td>71</td>
<td>Δ3,11b</td>
</tr>
<tr>
<td>del21R_Flank</td>
<td>ATGGAAGCCAATTGTCCTCTG</td>
<td>69.6</td>
<td>Δ11q, ▼21</td>
</tr>
<tr>
<td><strong>Sequencing primers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>del11qR_Seq</td>
<td>AGTTCAAGCCATTTCCTGCTG</td>
<td>70.3</td>
<td>Δ3,9,10,11q, and Δ11q</td>
</tr>
<tr>
<td>del19F_Seq</td>
<td>GGGAAGACCCAGAATGACA</td>
<td>69.5</td>
<td>Δ19</td>
</tr>
<tr>
<td>del3R_Seq</td>
<td>TCCAAACTCTGTCAAGCTG</td>
<td>70.1</td>
<td>Δ3</td>
</tr>
<tr>
<td>del11bF_Seq</td>
<td>CTCAAGGAACCCAGGGATGAA</td>
<td>68.9</td>
<td>Intronisation 11Δ3110</td>
</tr>
<tr>
<td>del11q+14R_Seq</td>
<td>TTCTGAGAACCTCCAGAGCA</td>
<td>70.7</td>
<td>Δ11q,14</td>
</tr>
<tr>
<td>del21F_Seq</td>
<td>GTTTGCAGAAAAACACCCA</td>
<td>69.3</td>
<td>Δ21</td>
</tr>
<tr>
<td>del9-11qF_Seq</td>
<td>CAGCTTGACACAGGGTTGGA</td>
<td>70.1</td>
<td>Δ3,9,10,11q (9-11q part)</td>
</tr>
<tr>
<td>ins21_middleFwd</td>
<td>TCTTGATCAGGCGAGTAC</td>
<td>70.8</td>
<td>▼21</td>
</tr>
</tbody>
</table>

* Tm was calculated using the conditions detailed in Standard PCR for short products, using the Integrated DNA Technologies (IDT) tool OligoAnalyzer 3.1. [https://sg.idtdna.com/calc/analyzer](https://sg.idtdna.com/calc/analyzer)

*b These primers were designed if the PCR product was over 500bp in order to sequence parts of isoforms reliably. Otherwise the relevant PCR primer was used in sequencing reactions.
2.4 Preparation of cDNA
The enzyme used for cDNA synthesis was Superscript® III Reverse Transcriptase (Invitrogen, ThermoFisher Scientific Inc., Waltham, USA). The cDNA synthesis was performed according to the manufacturer’s instructions with the RNA sample described in Section 2.1. The final reaction volume was 20µl. All preparation was carried out in sterile 200µl strip tubes on ice at a pre-PCR workbench treated with RNaseZAP (Ambion, USA) to avoid contamination with RNase and other products. Filter pipette tips were used until the first heating step. Products were made up to 100µl with MPW and stored at -20°C.

Three sets of primers were tested to determine which was the most effective for synthesising BRCA1 and BARD1 cDNA: oligo(dT), random hexamers (both sourced from Invitrogen, ThermoFisher Scientific Inc., Waltham, USA), and a gene specific primer for BRCA1. The design of the gene specific primer is described in Section 2.3, and is specific for the 3’ end region of BRCA1 mRNA.

2.5 Polymerase Chain Reaction (PCR)
All PCR reactions had a total reaction volume of 10µl, contained 3µl of the final 100µl cDNA solution unless otherwise stated, and were performed on an Eppendorf Mastercycler thermal cycler (GlobalScience, Germany). Pre-PCR preparation was carried out in a separate pre-PCR room to avoid risk of contamination from post-PCR products. Equipment and consumables (pipettes and tips, tubes and strip tubes) used were sterile. After 2µl was removed for gel electrophoresis, all reactions were stored at -20°C for Sanger and MinION sequencing. Where thermal cycling variables have been given a range of values in the PCR methods detailed below, a specific value is given separately for each result.

2.5.1 Standard PCR for a product size <3kb
For PCR products with length under 3kb each reaction contained 0.5µM of each forward and reverse primer (Integrated DNA Technologies, Singapore), 1.5mM of MgCl₂, 200µM of each dNTP, 1x PCR reaction buffer and 0.25 units of Fisher Taq-Ti polymerase (all sourced from Fisher Biotec, Wembley WA, Australia). Unless otherwise stated, thermal cycling conditions consisted of a denaturation step of 94°C for 2 minutes, 35 cycles of 94°C for 30 seconds, an annealing temperature between 64°C and 66°C (3-5°C below the Tm of primers in the reaction) for 30 seconds, an extension step at 72°C for a time
period according to product size (~1000bp per minute), with a final extension of 72°C for 5 minutes.

2.5.2 Standard PCR for full length *BRCA1* mRNA (product size 5.8kb)
For PCR products with length over 3kb, final reactions contained 1.3M betaine (Sigma-Aldrich, Steinheim, Germany), 1x KAPA long range buffer, 1.75mM of MgCl₂ (both from KAPA Biosystems, Wilmington, Massachusetts), 0.5µM of each forward and reverse primer (BRCA1_1F and BRCA1_24pR respectively), 300µM of KAPA 10mM dNTP mix, and 0.5 units of KAPA Long Range HotStart (both from KAPA Biosystems, Wilmington, Massachusetts), added in the order given. Unless otherwise stated, thermal cycling conditions consisted of a denaturation step of 94°C for 4 minutes, then 35 cycles of 94°C for 30 seconds, a 59°C annealing temperature for 30 seconds, and an extension step at 68°C for 12 minutes, with a final extension of 72°C for 12 minutes.

2.5.3 Alternative PCR for full length *BRCA1* mRNA
Final reactions contained 1M betaine (Sigma-Aldrich, Steinheim, Germany), 1x KAPA long range buffer, 2mM of MgCl₂ (both from KAPA Biosystems, Wilmington, Massachusetts), 0.7µM of each forward and reverse primer (BRCA1_1F and BRCA1_24pR respectively), 200µM of KAPA 10mM dNTP mix, and 0.5 units of KAPA Long Range HotStart (both from KAPA Biosystems, Wilmington, Massachusetts), added in the order given. Unless otherwise stated, thermal cycling conditions consisted of a denaturation step of 94°C for 2 minutes, then 35 cycles of 94°C for 30 seconds, a 59°C annealing temperature for 30 seconds, and an extension step at 68°C for 7 minutes, with a final extension of 72°C for 7 minutes.

2.5.4 Other PCR methods used for full length *BRCA1* mRNA
Reaction components were identical to either Section 2.5.2 or Section 2.5.3, and are specified for each experiment in the results sections. Touchdown PCR cycling conditions differed from the standard or alternative PCR cycling conditions by starting with 8 cycles in which the annealing temperature decreased by 1°C each cycle, then 30 cycles using the lowest annealing temperature reached. The starting temperatures ranged from 66°C to 62°C. Temperature gradient PCR cycling conditions differed by changing the annealing temperature across a plate so each sample has a different temperature at the annealing step, which ranged from 55°C to 63°C, and 35 cycles were used.
2.6 Agarose Gel Electrophoresis

A 1% agarose gel was used for resolving PCR products >1kb, while a 2% agarose gel was used for resolving products <1kb. 2µl of SYBR® Safe DNA Gel stain (Invitrogen, ThermoFisher Scientific, Auckland) was added to heat-dissolved agarose powder (Duchefa Biochemie Agarose SPI, Total Lab Systems Ltd, Haanlen, NL) in 1x TAE buffer. This solution was then allowed to set in an EASY-CAST™ Electrophoresis System (Owl Separation Systems, Portsmouth, NH, USA). The gel was then submerged in 1x TAE buffer.

1µl of 6x KAPA loading dye (KAPA Biosystems, Wilmington, Massachusetts) was added to 2µl of PCR product and 3µl of MPW, and loaded into wells. 2µl of KAPA Universal DNA ladder (KAPA Biosystems, Wilmington, Massachusetts) was loaded into at least one well in each gel and run alongside the products. This system was electrophoresed (using the PowerPac Basic, BioRad, NSW, Australia) at 80V for 30-60 minutes until the dye had visibly migrated at least two thirds the length of the gel. For higher separation of bands of similar size, the leading edge of the dye was occasionally run off the end of the gel.

Visualisation of the DNA bands was achieved using the gel documentation system Alliance 4.7, UVItec (Cambridge).

2.7 Extraction of DNA from agarose gel

Obtaining DNA from agarose gels was sometimes required for accurate Sanger sequencing. There were two techniques used for this purpose, detailed below. Both DNA extraction procedures were performed using a white/UV Transilluminator box (Alpha Innotech, San Leandro, CA).

2.7.1 Band stab
Sterile 6 gauge syringe needles were inserted into DNA bands of interest, swirled in 50ul of MPW and the needles discarded. 1µl of the resulting sample was then used in a repeat PCR with the same protocol used to produce the original band.

2.7.2 Full band extraction
In some cases, bands of interest were re-amplified using band stab PCR as described above. When such bands were still not pure enough for sequencing, they were excised from the gel using a sterile scalpel in order to isolate the single band of interest. This was purified from the agarose gel using the MEGAquick-spin Total fragment DNA
Purification kit, (iNtRON Biotechnology Inc., Korea). When required, recovered DNA was concentrated using an Eppendorf Concentrator 5301 (Hamburg, Germany).

2.8 Sanger Sequencing

Sanger sequencing using the Applied Biosystems Big Dye Terminator version 3.1 (BDT) was used to confirm PCR products were from BRCA1 and BARD1 mRNA prior to MinION nanopore sequencing. PCR products were diluted 1:3 in MPW and 1µl of this was added to a 10µl reaction volume containing 0.5µM primer, 2µl of BDT 5x sequencing buffer, and 0.5µl of BDT terminators. These reactions were cleaned using the illustra Sephadex G50 spin column (GE Healthcare, Sigma-Aldrich, Auckland) and eluted in 30µl of MPW. This final volume was electrophoresed using a 50cm array on the Hitachi 3130XL Genetic Analyzer (Applied BioSystems). Preparation was carried out according to the manufacturer’s instructions.

The program Geneious® version 9.1.5 was used (Biomatters, Ltd) [144], to analyse the raw sequence data, in conjunction with the BLAT tool [145] on UCSC genome browser [146] (https://genome.ucsc.edu/cgi-bin/hgBlat?command=start) to identify the gene each sequence matched. In the case of confirming isoforms, the Geneious®: Multiple Sequence Aligner tool was used to match the Sanger sequence of the sample with the predicted isoform as a reference sequence.

2.9 MinION sequencing (R9 model)

PCR samples to be sequenced required a clean-up step before beginning the MinION library preparation. The clean-up protocol was based on the bead clean-up step in the End-prep section of the Amplicon protocol, and can be found separately in Appendix B. Once the initial clean-up was done the samples were quantified using the Qubit® Fluorometer HS Assay Kit (Life Technologies, ThermoFisher Scientific, Auckland) then used in the full Amplicon protocol (R9 version) which can be found in Appendix C. The Nanopore Sequencing Kit SQK-NSK007 (R9 Version) was used in library preparation. The flow cell was of the code FLO-MIN104. The MinKNOW program used for running the MinION was NC_48Hr_Sequencing_Run_FLO_MIN104.py. The flow cell was topped up when the number of pores being used was less than 20, until the entire sample was used. The raw electrical signal was uploaded to Metrichor (version 1.107), a proprietary program run on servers at ONT, using the 2D Basecalling RNN for SQK-NSK007 which basecalled and returned the file in .fast5 file format.
2.10 MinION Sequence Analysis

Figure 2.2 describes the steps derived in Chapter 3 to analyse the basecalled data files. The software package Poretools [137] was used to convert .fast5 into .fasta files. All computer commands used and Excel methods can be found in Appendix D. The multi-fasta reference file (Appendix D.1) contains BRCA1 exon 11 split into three segments, labelled Exon11a, Exon11b, and Exon11c. These segments corresponded to published data on the common splice sites of the isoform Δ11q [60], where the 3’ end of exon 11 is spliced out, and the isoform 11Δ3110, where an intronisation occurs in the middle of exon 11 (see Table 1.1 for an explanation of intronisations).

![Flowchart](https://example.com/flowchart.png)

**Figure 2.2:** Flow chart describing MinION data analysis final pipeline.
2.10.1 Local BLAT parameters

The standalone BLAT program (version 36) was downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/ and used on a local Linux computer rather than on the UCSC Genome Browser website. Details about computer commands used can be found in Appendix D.2. The tabular format produced by BLAT is modelled on the BLAST tabular output. The graphical format is a .psl file, which was converted to a .sam file using the psl2sam.pl program (obtained June 2016). This creates .sam file of a format which uses M to denote both matches and mismatches in the CIGAR string. Samtools (version 0.1.18) was used to adjust .sam files in order to be read by visual programs.

2.10.2 Tabular format data analysis using R-Studio and Excel

The tabular output from BLAT was imported into R-Studio. The first two columns of read name and reference were used to generate a list of the references (in this case, each exon and intron) which matched to each read in the statistical program R (version 3.2.2) using R-Studio. This table was then opened in Microsoft Office Excel 2013, which was used to group the reads by references matched and then to summarise the results into standard isoform notation. The R script and exact methods of Excel analysis can be found in Appendix D.5 and D.6.

2.10.3 Graphical format data analysis using Samcat.py

Samcat.py (Appendix D.7) is a program developed in collaboration with a Software Engineer during the course of this project to concatenate all the .sam files associated with a single read, when mapped to different reference sequences, by joining the CIGAR strings together. The result is a .sam file which contains just one line of information per read. Due to the nature of this process, Samcat.py also eliminates spurious alignments, and therefore acts as a filtering step as well. The .sam files were viewed in Integrated Genomics Viewer (IGV) version 2.3.82. More information on the development of Samcat.py can be found in Chapter 4.
Chapter 3: Results

3.1 Introduction

There were two main parts to this project. The first involved the RT-PCR amplification of the 5.8kb long \textit{BRCA1} amplicon (includes all of the coding region) and the 2.0kb long \textit{BARD1} amplicon. If these could be produced it was assumed that other, alternatively spliced isoforms would be produced as well.

MinION data analysis of cDNA was the second part of this project. This posed a major challenge because the technology is so new that methods of analysis for full length mRNA data, (not data on individual splice junctions), is at a very early stage and established methods of analysis are not available. As a consequence, my project required development of methods to properly analyse this type of MinION data.

3.2 PCR and identification of \textit{BRCA1} and \textit{BARD1} from cDNA

Oligo(dT) primers, random hexamers and gene specific primers were used to synthesise cDNA in three separate reaction tubes. For each reaction there was also a control reaction with no reverse transcriptase enzyme added to ensure that products were the result of cDNA synthesis and not of any genomic products that may be present. Relatively short PCR reactions (under 500kb) were then used to check that the synthesis of cDNA was successful. These reactions used primers designed to amplify the 3’ end and the 5’ end of the \textit{BRCA1} transcript separately. Primers designed for the \textit{BARD1} full transcript were also used to test cDNA synthesis from oligo(dT) primers.

3.2.1 \textit{BRCA1} 3’ end and 5’ end PCR products

Products were detected from all cDNA samples. PCR of the 3’ end of \textit{BRCA1} resulted in the same two visible bands on the agarose gel for each sample, at about 500bp and 400bp respectively (Figure 3.1A). The lane containing cDNA made with oligo(dT) primers showed an intense, smeared band, while gene specific priming was less intense, and that from random hexamer priming was comparatively faint. Figure 3.2A shows the PCR of the 5’ end of \textit{BRCA1} that resulted in six visible bands from oligo(dT) priming, one bright and two faint bands from gene specific priming, and two barely visible bands from the random hexamer primed sample, all migrating between a 200bp and 600bp fragment size. There was significantly more product produced in oligo(dT) primed samples than there was in the random hexamer and gene specific primed samples for both the 3’ end and 5’ end. These results indicated that the most efficient priming method of \textit{BRCA1} cDNA synthesis from this RNA sample was with oligo(dT) primers.
3.2.1.1 Sanger sequencing

In order to confirm these products were produced from BRCA1 transcript a number of bands from each of the 3’ and 5’ PCR reactions were extracted for Sanger sequencing. Initially this was carried out using the band stab technique. Both bands produced in 3’ end PCR by gene specific priming in Figure 3.1A were extracted using the band stab method and underwent a second PCR to generate sufficient material for sequencing. Figure 3.1B shows the PCR result of the product indicated on Figure 3.1A, which was pure enough to be successfully Sanger sequenced (Figure 3.1C); however, the other band was not successfully sequenced (data not shown). A BLAT alignment verified that the sequence obtained from the pure band in Figure 3.1B matched to the BRCA1 gene (Figure 3.1D). Of the 6 bands produced in 5’ PCR (Figure 3.2) by Oligo(dT) priming, only one band successfully produced usable sequence (indicated on Figure 3.2A). This was extracted using the band stab method, and the second PCR (Figure 3.2B) produced 5 bands in total. The band of interest was then fully extracted and purified before Sanger sequencing (Figure 3.2C). The resulting sequence was confirmed to belong to the BRCA1 gene using a BLAT alignment (Figure 3.2D). These two results enabled us to move forward with long range PCR of the full length BRCA1 transcript.
Figure 3.1: *BRCA1* 3' end Amplification and Sequencing.
A: PCR check of 3' end *BRCA1* cDNA synthesis. 5µl of cDNA was used. DNA extracted by band stab is indicated on the figure. RT+ refers to reactions containing reverse transcriptase, RT- refers to negative control reactions without reverse transcriptase. B: PCR result of the band stab in A. For both A and B the standard PCR protocol for product size <3kb was used, with an annealing temperature of 65°C and extension time of 30 seconds. Primers were BRCA1_20F and BRCA1_24pR. 5µl of each PCR reaction was run on 1% agarose gels. Size in base pairs (bp) of relevant bands in the marker lane (Ladder) is indicated. C: Sanger sequencing result of this band primed with BRCA1_20F visualised in Geneious™. D: BLAT alignment tool showing the sequence match using the UCSC Genome Browser. Enlarged pictures are shown of individual exon alignments. The black bar at the top of each picture represents the sequence searched.
Figure 3.2: BRCA1 5’ end Amplification and Sequencing.  
A: PCR check of 5’ end BRCA1 cDNA synthesis. 5µl of cDNA was used. DNA extracted by band stab is indicated on the figure. 5µl of each PCR reaction was run on a 1% agarose gel. RT+ refers to reactions containing reverse transcriptase, RT- refers to negative control reactions without reverse transcriptase. B: PCR result of the band stab in A. 5µl of each PCR reaction was run on a 2% agarose gel. The band indicated was cut out and purified before sequencing. For both A and B the standard PCR protocol for product size <3kb was used, with an annealing temperature of 64°C and extension time of 32 seconds. Primers were BRCA1_1F and BRCA1_5R. Size in bp of relevant bands in the marker lane (Ladder) is indicated. C: Sanger sequencing result of this band primed with BRCA1_5R visualised in Geneious™. D: BLAT alignment tool showing the sequence match using the UCSC Genome Browser. Enlarged pictures are shown of individual exon alignments. The black bar at the top of each picture represents the sequence searched.
3.2.2  *BARD1* PCR products

On the basis of previous results, oligo(dT) primed cDNA was also used to generate *BARD1* PCR products. Primers were designed so the length of the amplicon was reduced to 2.0kb from the full length 2.6kb transcript (see Figure 2.1B). Multiple bands were generated by *BARD1* PCR, indicative of multiple isoforms (Figure 3.3A). One of these bands was ~2000bp, and seemed likely to represent the full length transcript. This experiment was replicated a number of times throughout the course of this project in order to gain enough product for MinION sequencing, and each replicate produced consistent results (data not shown).

3.2.2.1 Sanger sequencing

In order to confirm these products were produced from *BARD1* transcript a number of bands from the initial PCR reaction were extracted from the gel for Sanger sequencing. Initially this was carried out using the band stab technique. Unfortunately all resulting samples had contamination from other bands, and could not be directly used in Sanger sequencing. These bands sometimes required two consecutive band stabs and a full band extraction before sequencing. The last corresponding PCR reaction was also sequenced in an attempt to ensure success.

All bands in Figure 3.3 eventually produced at least two sequences that were of good quality, as demonstrated by Figure 3.3C. BLAT analysis indicated that all bands matched *BARD1* demonstrated in Figure 3.3D. Some exon skipping was also detected, confirming that the multiple bands visible represented different isoforms (data not shown). This result confirmed that this PCR protocol produces *BARD1* PCR products suitable for use on the MinION.
Figure 3.3: BARD1 Amplification and Sequencing.

A: PCR check of BARD1 cDNA synthesis. 5µl of cDNA was used. DNA extracted by band stab is indicated on the figure. RT+ refers to reactions containing reverse transcriptase, RT- refers to negative control reactions without reverse transcriptase. B: PCR result of the band stab in A. The band indicated was cut out and purified before sequencing. For both A and B the standard PCR protocol for product size <3kb was used, with an annealing temperature of 66°C and extension time of 2 minutes. Primers were BARD1_1F and BARD1_11R. 5µl of each PCR reaction was run on 1% agarose gels. Size in bp of relevant bands in the marker lane (Ladder) is indicated. C: Sanger sequencing result of this band primed with BARD1_11R visualised in Geneious™. D: BLAT alignment tool showing the sequence match using the UCSC Genome Browser. Enlarged pictures are shown of individual exon alignments. The black bar at the top of each picture represents the sequence searched.
3.2.3 Titration to determine optimal amount of cDNA

To optimise the RT-PCR a titration experiment was performed with varying amounts of cDNA. For simplicity, this was performed using BARD1 primers and reaction conditions. This experiment (Figure 3.4) demonstrated that 3µl of cDNA per reaction was sufficient to gain satisfactory results.

3.2.4 Full length BRCA1 PCR

The BRCA1 reverse primer is situated near the 5’ end of exon 24 (Figure 2.1), reducing the size of the 7.2kb full length transcript to an amplicon size of 5.8kb. A different pattern of bands was observed each time a new BRCA1 PCR reaction was performed (Figure 3.6). As a result, a number of variations to the standard PCR for full length BRCA1 mRNA were used to generate the final pool of PCR products used in MinION sequencing. Initially, the aim was to ensure that a band the size of the full length PCR product was present. However, when the results began to show a wide range of variation between samples, a general goal of generating the full length product as well as a wide selection of potential isoforms was established. Stochastic amplification of various BRCA1 isoforms in the initial PCR cycle was decided to be the likely cause of the wide variation.

A summary of the number and approximate size of the bands present across all full length BRCA1 PCR reactions can be found in Table 3.1. The PCR protocols to generate these bands were based on either the standard PCR for full length BRCA1 mRNA, or the alternative PCR for full length BRCA1 mRNA. Along with the normal standard and alternative protocol, temperature gradients and touchdowns for each protocol were used. Some of these results are demonstrated in Figure 3.6; the remaining experiments can be found in Appendix E.
3.2.4.1 Sanger sequencing
No attempt was made to purify long PCR products as done for the short amplicons. As each reaction was generated using the same forward and reverse primer there was enough similarity at the beginning and end of amplicons to generate sufficient sequence data to confirm the presence of BRCA1. A total of 16 samples representative of the range of PCR protocols used were chosen for Sanger sequencing, and 11 of these were of sufficient quality to match to BRCA1 using the BLAT alignment (data not shown).

3.3 MinION
3.3.1 Samples sequenced
Two MinION samples were run during this project. The first comprised both BRCA1 and BARD1 PCR reactions, as indicated in Figure 3.6. The first selection of BRCA1 reactions was pooled together, all BARD1 reactions were pooled together, and the initial clean-up step was performed on each of these two pooled samples. Samples were checked by gel electrophoresis prior to the MinION run (Figure 3.5). A notable feature of this result is the high molecular weight smear present in the BARD1 sample, which did not, however, seem to affect sequencing or downstream analyses. The final concentration of DNA in the cleaned BRCA1 samples was 37.8ng/μl, with a total of 18μl of sample. The final concentration of DNA in the BARD1 sample was 116ng/μl, with a total of 26μl of sample. There was not enough BRCA1 to run on the MinION on its own, so the decision was made to pool it with BARD1, regardless of the smear present in the BARD1 sample. Prior to this decision BRCA1 and BARD1 full

<table>
<thead>
<tr>
<th>Estimated size</th>
<th>Total number of bands</th>
<th>Estimated size</th>
<th>Total number of bands</th>
</tr>
</thead>
<tbody>
<tr>
<td>7000</td>
<td>4</td>
<td>2500</td>
<td>12</td>
</tr>
<tr>
<td>6100</td>
<td>6</td>
<td>2300</td>
<td>6</td>
</tr>
<tr>
<td>6000</td>
<td>9</td>
<td>1700</td>
<td>1</td>
</tr>
<tr>
<td>4000</td>
<td>2</td>
<td>1400</td>
<td>2</td>
</tr>
<tr>
<td>3100</td>
<td>3</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td>3000</td>
<td>2</td>
<td>600</td>
<td>1</td>
</tr>
<tr>
<td>2900</td>
<td>6</td>
<td>470</td>
<td>1</td>
</tr>
<tr>
<td>2700</td>
<td>16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
length cDNA reference sequences were aligned to each other using the proposed method of alignment at the time (LAST aligner [147]), to determine if they could be distinguished if they were pooled. They have similar protein domains and may have shown a relationship at the nucleotide level, which would have the potential to confuse the analysis if the two samples were pooled. However, no alignments were found (data not shown).

All 18µl of the BRCA1 sample and 5.8µl of BARD1 sample was used to get an equal amount of DNA from each, then it was made up to 50µl with MPW to begin the MinION Amplicon protocol. The second sample run on the MinION contained BRCA1 reactions only, as indicated in Figure 3.6 and related figures in Appendix E. These reactions were pooled and cleaned as described above, and the concentration of DNA in the cleaned sample was 54.8ng/µl in a total sample size of 30µl. This was made up to 50µl with MPW to begin the MinION Amplicon protocol. The MinION flow cell was washed according to the manufacturer’s instructions before the second sample was sequenced.
Figure 3.6: Varying long range BRCA1 PCR results. Samples that underwent Sanger sequencing and MinION sequencing are indicated. MinION sequenced samples are further identified by which pooled sample they contributed to. Size in bp of relevant bands in the marker lane (Ladder) is indicated on each figure. 2µl of each reaction was run on 1% agarose gels. A: Temperature gradient based on the standard PCR for full length BRCA1. Annealing temperature is displayed for each sample. B and C: Temperature gradient based on the alternative PCR for full length BRCA1, carried out on two different days. Annealing temperature is displayed for each sample. D: Standard PCR for full length BRCA1.
3.3.2 Statistics from the 2 runs

The first MinION run produced a total of 105,482 reads, of which 23,978 passed the quality control filters for 2D reads by Metrichor analysis and 81,504 failed. Of the total, 4,302 aligned to \textit{BRCA1} in BLAT analysis, and 54,505 aligned to \textit{BARD1}, leaving 46,675 unaligned. The second MinION run produced a total of 12,022 reads, of which 995 passed the quality control filters for 2D reads by Metrichor analysis and 11,027 failed. Of the total, 2115 aligned to \textit{BRCA1} in BLAT analysis, and 837 aligned to \textit{BARD1}, leaving 9070 unaligned. Though the second sample contained \textit{BRCA1} PCR products only, some of the \textit{BARD1} amplicons from the previous run were not able to be fully washed out, accounting for the number of reads in the second run which aligned to \textit{BARD1}.

3.3.3 MinION read data analysis

The flow chart in Figure 3.8 describes the trial-and-error approach to developing the final data analysis pipeline described in Chapter 2. Many alignment programs and reference sequence arrangements were tried before BLAT alignment was selected. Table 3.2 and Table 3.3 summarise these programs, the parameters used and the reason each program was discarded before the final data analysis. Many of these programs produced a similar final result of aligning to just one exon, demonstrated by Figure 3.7, generated using the LAST analysis [147]. For the final analysis using a local BLAT download, the two MinION read data sets were combined.

![Figure 3.7: LAST alignment example result.](image)

Reference is indicated on the figure, located at \textit{BRCA1} Exon 11. Each individual read was assigned a number, and some of the reads that align to \textit{BRCA1} exon 11 are shown. This figure was generated using the program LAST to align the reads to the whole gene reference (Table 3.2), and visualised in the program Geneious® 9.1.5, Biomatters Ltd.
Figure 3.8: Iterative development of final data analysis pipeline. Arrows from the centre indicate a new trial, while arrows between alignments represent trials that influence decisions of the next trial. Each box has a number corresponding to a row in Table 3.2 or Table 3.3. Grey represents trials leading to "dead ends". Blue indicates trials that lead to a successful analysis.
Table 3.2: Iterative development of analysis pipeline (unsuccessful)

<table>
<thead>
<tr>
<th>Flow chart code</th>
<th>Program – reference used</th>
<th>Parameters used</th>
<th>Description of result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LAST (version last-621) [147] – cDNA reference – whole gene reference</td>
<td>lastdb -Q 0 lastal -s2 -T0 -Q0 -a1 last-map-probs (default)</td>
<td>Can identify some full length reads, but not able to see variations well. Only shows best part of alignment (Figure 3.7). No way to group similar reads.</td>
</tr>
<tr>
<td>2</td>
<td>LAST (version last-621) [147] – whole gene reference</td>
<td>As in number 1, with an e-value of 120.</td>
<td>No significant difference to number 1.</td>
</tr>
<tr>
<td>3</td>
<td>BBMap (v. 36.14) [148] – whole gene reference</td>
<td>mapPacBio.sh, recommended parameters for nanopore reads.</td>
<td>No significant difference to number 1.</td>
</tr>
<tr>
<td>4</td>
<td>Bwa-sw (version 0.7.7) [149] – whole gene reference</td>
<td>defaults</td>
<td>No significant difference to number 1.</td>
</tr>
<tr>
<td>5</td>
<td>Bwa-mem (version 0.7.7) [149] – whole gene reference</td>
<td>-x ont2d</td>
<td>No significant difference to number 1, did not split sequences and instead deleted adjacent sequence</td>
</tr>
<tr>
<td>6</td>
<td>STAR alignment (v. 2.2) [150] – whole gene reference</td>
<td>Indexed reference only</td>
<td>Program crashed while indexing reference.</td>
</tr>
<tr>
<td>7</td>
<td>Salmon (v. 0.6.0) [151] – multifa reference with exons only</td>
<td>quant -p 4 -k 5 --sensitive -U</td>
<td>Quantitative, gave a number of reads which aligned to each exon. Not helpful.</td>
</tr>
<tr>
<td>8</td>
<td>Kallisto (v. 0.42.4) [152] – multifa reference with exons only Note: used .fastq files instead of .fasta</td>
<td>quant -i5000 -s 2000 --plaintext --single --pseudobam</td>
<td>Did not split sequences and instead deleted adjacent sequence</td>
</tr>
<tr>
<td>9</td>
<td>LAST (version last-621) [147] – whole gene reference Note: used .fastq files instead of .fasta</td>
<td>As in number 1, using function “last-split -c0,” after last-map-probs.</td>
<td>No significant difference to number 1.</td>
</tr>
<tr>
<td>10</td>
<td>HISAT (v. 2.0.3) [through Galaxy server] [153, 154] – whole gene reference</td>
<td>Defaults, some penalties adjusted for nanopore error-rate.</td>
<td>Output file had no information.</td>
</tr>
<tr>
<td>11</td>
<td>Geneious: Map-to-genome function (v. 9.1.0) [144] – whole gene reference</td>
<td>Defaults</td>
<td>No significant difference to number 1.</td>
</tr>
</tbody>
</table>
## Table 3.3: Iterative development of analysis pipeline (successful)

<table>
<thead>
<tr>
<th>Flow chart code</th>
<th>Program – reference used</th>
<th>Parameters used</th>
<th>Description of result</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>BLAT alignment (v. 36) [145] – hg19 human genome</td>
<td>Defaults</td>
<td>Aligned reads against correct regions of genome, showed connection between exons.</td>
</tr>
<tr>
<td>13</td>
<td>BLAST alignment (v. 2.2.28) [155] – multi-fasta file with exons only</td>
<td>Defaults</td>
<td>Default output: shows each individual alignment. Tabular output: gives information able to be manipulated to give a summary. Stringent parameters missed alignments to exons found in BLAT.</td>
</tr>
<tr>
<td>14</td>
<td>BLAT alignment (v. 36) [145] - whole gene reference</td>
<td>Defaults. Output to four file types: pslx, MAF, standard BLAST, and BLAST tabular.</td>
<td>MAF format: similar to no. 1. Standard BLAST: not helpful. Pslx showed alignments not consistent to exons. BLAST tabular: not helpful as reference was all the same, difficult to determine specific places of alignment.</td>
</tr>
<tr>
<td>15</td>
<td>BLAT alignment (v. 36) [145] with UCSC website parameters – whole gene reference, one .fasta file</td>
<td>-stepSize=5 - repMatch=2253 - minScore=0 - minIdentity=0</td>
<td>Output for one read was identical to output generated by the same read when using the tool on the Genome Browser website.</td>
</tr>
<tr>
<td>18</td>
<td>Tabular output manipulated in R (v. 3.2.2) and Excel (2013)</td>
<td>See Appendix D.5 and D.6 for R script and Excel analysis.</td>
<td>Summary of isoforms ordered by number of reads associated with each isoform.</td>
</tr>
<tr>
<td>19</td>
<td>Pslx converted to .sam file using Perl program psl2sam.pl (obtained June 2016) [156]. Program Samcat.py developed to join individual alignments together</td>
<td>See Appendix D.7 for Samcat.py program.</td>
<td>Able to now look at full isoform alignments. Also filtered out bad alignments.</td>
</tr>
<tr>
<td><strong>Final method implemented</strong></td>
<td>BLAT alignment (v. 36) to multi-fasta reference (Appendix D.1)</td>
<td>UCSC website parameters and two output formats</td>
<td>Using both outputs enabled a comprehensive analysis of the isoform data from the MinION.</td>
</tr>
</tbody>
</table>
3.3.4 BLAT Tabular output
Table 3.4 and Table 3.5 summarise the results from *BRCA1* and *BARD1* analyses respectively. A total of 34 *BRCA1* isoforms were detected with a read depth greater than five reads per isoform. A total of 39 *BARD1* isoforms were detected with a read depth greater than 23 reads per isoform. A greater number of reads was chosen for the read depth threshold because of the significantly greater amount of *BARD1* in the sample. There were 24 *BRCA1* isoforms and 18 *BARD1* isoforms that have not been previously described. An additional seven *BARD1* isoforms have only been identified in disease states.
Table 3.4: BRCA1 isoforms found in BLAT tabular output

<table>
<thead>
<tr>
<th>Official name</th>
<th>No. reads per isoform</th>
<th>Previously described?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ11q</td>
<td>86</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ9,10,11q</td>
<td>79</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>11Δ3110</td>
<td>34</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ10-17</td>
<td>29</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,11q</td>
<td>28</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ9,10,11q,12-15</td>
<td>19</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,10,11q,19</td>
<td>18</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ11q,19</td>
<td>17</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,10,11q, ▼21</td>
<td>16</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9-11</td>
<td>16</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ3,11Δ3110</td>
<td>13</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ3</td>
<td>12</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ19</td>
<td>12</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ11q, ▼21</td>
<td>12</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,10</td>
<td>10</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ9,10,11q,19,21</td>
<td>10</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,9,10,11Δ3110/3240,Δ14,20,22</td>
<td>10</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ11q,14,21</td>
<td>10</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,10,21</td>
<td>9</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ3,9,10,11q</td>
<td>9</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,9,10,11q, ▼21</td>
<td>9</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ21</td>
<td>9</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>Δ11</td>
<td>9</td>
<td>Colombo, et. al. [60]</td>
</tr>
<tr>
<td>11Δ3110/3240, Δ15-17</td>
<td>8</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2-19</td>
<td>8</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,10,11q,21</td>
<td>7</td>
<td>Not found</td>
</tr>
<tr>
<td>11Δ3110/3240,19</td>
<td>7</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2-6</td>
<td>7</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ11q,23</td>
<td>7</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ10-17,21</td>
<td>7</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9,11q, ▼21</td>
<td>6</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ10-17,19</td>
<td>6</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ11q,17</td>
<td>6</td>
<td>Not found</td>
</tr>
<tr>
<td>Name</td>
<td>No. reads per isoform</td>
<td>Previously described?</td>
</tr>
<tr>
<td>------</td>
<td>----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>$\Delta 2,4$</td>
<td>286</td>
<td>Not found</td>
</tr>
<tr>
<td>$\Delta 4$-6</td>
<td>176</td>
<td>Bosse, K. R., et. al., (2012) [95] Identified as pathogenic.</td>
</tr>
<tr>
<td>▼10</td>
<td>99</td>
<td>Lattimore, V., (2016) [102]</td>
</tr>
<tr>
<td>$\Delta 7$</td>
<td>95</td>
<td>Lattimore, V., (2016) [102]</td>
</tr>
</tbody>
</table>

*Table continued on following page*
<table>
<thead>
<tr>
<th>Name</th>
<th>No. reads per isoform</th>
<th>Previously described?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ10</td>
<td>91</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,4-9</td>
<td>81</td>
<td>Not found</td>
</tr>
<tr>
<td>▼4</td>
<td>72</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2-10</td>
<td>71</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ9</td>
<td>60</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2-7</td>
<td>51</td>
<td>Feki, A., (2005) [16] Identified as pathogenic.</td>
</tr>
<tr>
<td>Δ4 ▼ 10</td>
<td>45</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,5</td>
<td>45</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ4,7</td>
<td>44</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ8</td>
<td>43</td>
<td>Sporn, J.C., et.al., (2011) [96]</td>
</tr>
<tr>
<td>Δ4,9</td>
<td>41</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ4,10</td>
<td>39</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,10</td>
<td>38</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ4,5</td>
<td>32</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2,4-6</td>
<td>32</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ3-4,10</td>
<td>30</td>
<td>Not found</td>
</tr>
<tr>
<td>Δ2-5</td>
<td>28</td>
<td>Bosse, K. R., et. al., (2012) [95]</td>
</tr>
<tr>
<td>Δ2-4,8</td>
<td>27</td>
<td>Sporn, J.C., et.al., (2011) [96]</td>
</tr>
<tr>
<td>Δ5-9</td>
<td>26</td>
<td>Not found</td>
</tr>
<tr>
<td>▼2</td>
<td>26</td>
<td>Bosse, K. R., et. al., (2012) [95] Insertion not found independently before, has been found in association with Δ4q,6q.</td>
</tr>
<tr>
<td>Δ3-5</td>
<td>26</td>
<td>Bosse, K. R., et. al., (2012) [95]</td>
</tr>
<tr>
<td>Δ2,8</td>
<td>24</td>
<td>Not found</td>
</tr>
</tbody>
</table>
3.3.5 BLAT Graphical output and Samcat.py

The program Samcat.py that was developed during the course of this project was required to obtain a complete view of the transcript reads spanning the entire gene. Without the program, the reads could only be viewed according to each reference sequence, due to the necessary structure of the multi-fasta file (data not shown). Samcat.py was applied to the entire BLAT graphical output dataset, but at this stage, enough data was obtained for the focus to be turned solely on six BRCA1 isoforms, so only these particular groups, identified from the BRCA1 tabular results, were analysed. The reads associated with each group were transferred into .sam files of isoform groups to be viewed in IGV separately. The six novel BRCA1 isoforms that were chosen to be further analysed were: Δ11q, ▼21 (Figure 3.9), Δ10-17 (Figure 3.10), Δ3,11Δ3110 (Figure 3.11), Δ11q,14,21 (Figure 3.12) Δ3,9,10,11q (Figure 3.13), and Δ11q,19. This format allowed identification of a number of possible splice acceptor and donor shifts. Some of these shifts occurred in all reads in an isoform group, while others only occurred in some. These changes are described in Table 3.6. It will require further analysis to determine if the patterns of multiple splice shifts occur are consistent between molecules, though a cursory inspection suggests that the patterns are not consistent. Discrepancy between the splice shifts seen in the figures and those described in Table 3.6 is a result of the alignment program aligning bases in adjacent exons twice, and will be discussed further in Chapter 4. Figure 3.9A is an example of the entire gene view for isoform Δ11q, ▼21 in IGV, with large portions of introns removed for viewing purposes, and is not shown for other isoforms. Of the full length group, 22 reads were selected for further examination (Figure 3.14). Upon close examination of exons in the full length group variations were apparent in exons 1, 6, 8, 10, and 16. No single read displays a full alignment to every exon due to a consistent absence of three bases in the 3’ end of exon 10 (Figure 3.14D). Other exon alignments showed little variation in splice shifts between reads (data not shown). It is important to note that due to the format of the .sam files used, mismatched base information is not available, and is shown in Figure 3.9 to Figure 3.14 as a match.
Figure 3.9: Isoform Δ11q, ▼21 BLAT final graphical result.
The position of each exon is shown as a blue bar and is indicated on the figure. 
A: Full gene analysis picture from IGV, with large intron stretches removed. Exon 11 and the intron 21 insertion positions are indicated. 
B: Detail view of exon 1 specific reads, deletion of 6 bases indicated by a black box. 
C: Detail of exon 2 specific reads, deletion of 12 bases indicated by a black box. 
D: Detail of exon 16 specific reads, deletion of 2 bases indicated by a black box. 
E: Detail of specific reads retaining intron 21. The black box indicates 5 bases at the 5’ end of the insertion that are identical to the 5 bases at the 3’ end of exon 15, and have aligned to both locations.
Figure 3.10: Isoform Δ10-17 BLAT final graphical result. The position of each exon is shown as a blue bar and is indicated on the figure. A: Detail view of exon 1 specific reads, occasional deletion of 6 bases indicated by a black box. B: Detail of exon 3 specific reads, occasional deletion of 4 bases indicated by a black box. C: Detail of exon 6 specific reads, occasional deletion of 4 bases indicated by a black box.

Figure 3.11: Isoform Δ3,11Δ3110 BLAT final graphical result. The position of exon 5 is shown as a blue bar and is indicated on the figure. Detail view of exon 5 specific reads, occasional deletion of 21 bases indicated by a black box.
Figure 3.12: Isoform Δ11q,14,21 BLAT final graphical result. The position of exon 1 is shown as a blue bar and is indicated on the figure. Detail view of exon 1 specific reads, deletion of 6 bases indicated by a black box.

Figure 3.13: Isoform Δ3,9,10,11q BLAT final graphical result. The position of the exon is shown as a blue bar and is indicated on the figure. A: Detail view of exon 1 specific reads, deletion of 6 bases indicated by a black box. B: Detail view of exon 14 specific reads, possible occasional deletion of 7 bases indicated by black box.
Figure 3.14: Exon variations in 22 reads from the full length group graphical result. The position of the exon is shown as a blue bar and is indicated on the figure. A: Detail view of exon 1 specific reads, occasional deletion of 6 bases indicated by a black box. B: Detail of exon 6 specific reads, occasional deletion of 4 bases indicated by a black box. C: Detail view of exon 8 specific reads, occasional deletion of 3 bases indicated by a black box. D: Detail of exon 10 specific reads, consistent deletion of 3 bases indicated by a black box. E: Detail of exon 16 specific reads, occasional deletion of 2 bases indicated by the black box.
Table 3.6: BRCA1 isoform splice acceptor/donor shifts identification

<table>
<thead>
<tr>
<th>Previous name</th>
<th>No. of reads(^a)</th>
<th>Acceptor/Donor shifts(^b)</th>
<th>Acceptor/Donor shift description</th>
<th>New name</th>
<th>Figure reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ10-17</td>
<td>29</td>
<td>Δ1Aq</td>
<td>Δ1Aq: 6bp shift, occurs in some transcripts</td>
<td>-</td>
<td>Figure 3.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Δ3p(^c)</td>
<td>Δ3p: 4bp shift, occurs in some transcripts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Δ6q(^c)</td>
<td>Δ6q: 4bp shift, occurs in some transcripts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ11q,19</td>
<td>17</td>
<td>None detected</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Δ3,11Δ3,110</td>
<td>13</td>
<td>Δ5q</td>
<td>Δ5q: 21bp shift, occurs in most transcripts</td>
<td>-</td>
<td>Figure 3.11</td>
</tr>
<tr>
<td>Δ11q,▼21</td>
<td>12</td>
<td>Δ1Aq</td>
<td>Δ1Aq: 6bp shift, occurs in all transcripts</td>
<td>Δ1Aq,2p,11q,▼21</td>
<td>Figure 3.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Δ2p</td>
<td>Δ2p: 13bp shift, occurs in all transcripts</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Δ16p(^c)</td>
<td>Δ16p: 3bp shift, occurs in all transcripts but one</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ11q,14,21</td>
<td>10</td>
<td>Δ1Aq</td>
<td>Δ1Aq: 6bp shift, occurs in all transcripts</td>
<td>Δ1Aq,11q,14,21</td>
<td>Figure 3.12</td>
</tr>
<tr>
<td>Δ3,9,10,11q</td>
<td>9</td>
<td>Δ1Aq(^f)</td>
<td>Δ1Aq: 6bp shift, occurs in all transcripts but one</td>
<td>-</td>
<td>Figure 3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Δ14Ap(^f)</td>
<td>Δ14Ap: 7bp shift, occurs in some transcripts</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Number of reads identified by the tabular output. The graphical output revealed reads that were not of the isoform indicated.

\(^b\) For the purposes of this identification, 11q has been treated as a normal exon splice rather than a donor shift.

\(^c\) Not previously described in the literature.

\(^d\) There appears to be one read aligning to the entire exon in Figure 3.9, however this particular read does not belong in this isoform group as it does not contain the intron 21 insertion.

\(^e\) Appears to be a 7bp shift but could be a 9bp shift due to the presence of a polyA tract in adjacent exon 13. Δ14p was described previously as a 3bp deletion 60. Colombo M, Blok MJ, Whiley P, Santamarina M, Gutierrez-Enriquez S, Romero A, et al. Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. Human Molecular Genetics. 2014;23(14):3666-80, so this variation has been distinguished here as Δ14Ap. These changes are further discussed in Chapter 4: Discussion.
3.4 Verification of isoforms
The primers used for isoform verification were designed prior to the graphical analysis, so do not take into account possible splice acceptor or donor shifts. PCR was performed on cDNA with the goal of isolating the variant parts of the six isoforms identified in Table 3.6 for Sanger sequencing. The full length BRCA1 PCR reaction was not performed. Many of the bands that appeared to be the correct size for the targeted isoform were revealed by Sanger sequencing to contain consecutive exons instead of the hypothesised alternatively spliced exons (data not shown).

3.4.1 PCR and sequencing of predicted isoforms
Two sets of primers (four primers in total) were initially designed for PCR reactions of each isoform. Junction primers were designed to span the exon/exon junctions at either side of each target isoform, while for the intron 21 insertion, primers were designed in the middle of the inserted region. Flanking primers were designed to encompass the whole isoform. Target product length was no less than 350bp. Additional primers were designed to increase the chances of successful Sanger sequencing for PCR products that were greater than 500bp. Primers used per isoform in PCR and sequencing are detailed in Table 3.7.

The PCR method used in all cases was based on the standard protocol for products <3kb. For isoforms that were difficult to amplify and sequence additional primers were selected for testing from all primers designed throughout this project. The sequencing reactions were aligned to their predicted isoform references using the Geneious: Map to Reference function. Isoform Δ10-17 (Figure 3.15) and part of isoform Δ1Aq,2p,11q,16p,▼21 (Figure 3.16) had enough sequence to be successfully verified.
Table 3.7: Isoform PCR and sequencing.

<table>
<thead>
<tr>
<th>Name</th>
<th>PCR Forward Primer</th>
<th>PCR Reverse Primer</th>
<th>Sequencing Primer 1</th>
<th>Sequencing Primer 2</th>
<th>PCR target product length</th>
<th>PCR produced bands?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ10-17 (Junction)</td>
<td>BRCA1_9-18FJunc</td>
<td>BRCA1_23RJunc</td>
<td>del3R_Seq</td>
<td>del11bF_Seq</td>
<td>408</td>
<td>YES</td>
</tr>
<tr>
<td>Δ10-17 (Flanking)</td>
<td>del10-17FFlank</td>
<td>del10-17RFlank</td>
<td>-</td>
<td>-</td>
<td>488</td>
<td>YES</td>
</tr>
<tr>
<td>Δ11q19 (Junction)</td>
<td>BRCA1_11a-12FJunc</td>
<td>BRCA1_18-20RJunc</td>
<td>-</td>
<td>-</td>
<td>1092</td>
<td>NO</td>
</tr>
<tr>
<td>Δ11q19 (Flanking)</td>
<td>del11qF_Flank</td>
<td>del19R_Flank</td>
<td>del11qR_Seq</td>
<td>del19F_Seq</td>
<td>1362</td>
<td>YES</td>
</tr>
<tr>
<td>Δ3,11b (Junction)</td>
<td>BRCA1_2-5FJunc</td>
<td>BRCA1_11a-11cRJunc</td>
<td>del3R_Seq</td>
<td>del11bF_Seq</td>
<td>685</td>
<td>YES</td>
</tr>
<tr>
<td>Δ3,11b (Flanking)</td>
<td>del3F_Flank</td>
<td>del11bR_Flank</td>
<td>BRCA1_2-5FJunc</td>
<td>11a-11cRJunc</td>
<td>1117</td>
<td>YES</td>
</tr>
<tr>
<td>Δ11q,Δ21 (Junction)</td>
<td>BRCA1_11a-12FJunc</td>
<td>BRCA1_ins21RJunc</td>
<td>del19F_Seq</td>
<td>del11qR_Seq</td>
<td>1361</td>
<td>YES</td>
</tr>
<tr>
<td>Δ11q,Δ21 (Flanking)</td>
<td>del11qF_Flank</td>
<td>ins21_RFlank</td>
<td>del11qR_Seq</td>
<td>Ins21middleFwd</td>
<td>1666</td>
<td>YES</td>
</tr>
<tr>
<td>Δ11q,14,21 (Junction)</td>
<td>BRCA1_11a-12FJunc</td>
<td>BRCA1_20-22RJunc</td>
<td>-</td>
<td>-</td>
<td>1084</td>
<td>NO</td>
</tr>
<tr>
<td>Δ11q,14,21 (Flanking)</td>
<td>del11qF_Flank</td>
<td>del21R_Flank</td>
<td>del11q+14R_Seq</td>
<td>del21F_Seq</td>
<td>1349</td>
<td>YES</td>
</tr>
<tr>
<td>Δ3,9,10,11q (Junction)</td>
<td>BRCA1_2-5FJunc</td>
<td>BRCA1_11a-12RJunc</td>
<td>-</td>
<td>-</td>
<td>558</td>
<td>YES</td>
</tr>
<tr>
<td>Δ3,9,10,11q (Flanking)</td>
<td>del3F_Flank</td>
<td>del11qR_Seq</td>
<td>del3R_Seq</td>
<td>del9-11qF_Seq</td>
<td>916</td>
<td>YES</td>
</tr>
</tbody>
</table>

* Names of isoforms according to previous name because this experiment was designed prior to the graphical output analysis.
* Primers used for PCR were also used for Sanger sequencing when applicable.
* Sequencing primers were used when the predicted PCR product size was greater than 500bp, or when the verification process proved difficult. Additional or alternative primers were selected from all primers designed during this project to provide a greater chance of isoform confirmation.
* Successfully confirmed.
For all experiments, the standard protocol for products <3kb was used. The forward primer was BRCA1_9-18FJunc and reverse primer was BRCA1_23RJunc. Annealing temperature was 64°C and extension time was 30 seconds. A: 3µl of cDNA was used for the PCR reaction. 2µl of PCR product was loaded on a 1% gel. The band indicated was extracted by band stab. B: 1µl of extracted band solution from A was used in this PCR. 2µl of PCR product was loaded on a 2% gel. The band indicated was extracted by band stab. C: 1µl of extracted band solution from B was used in this PCR. 2µl of PCR product was loaded on a 2% gel. This reaction was Sanger sequenced. D: Geneious Map to Reference result of Sanger sequencing from C. Purple bar denotes position of exons 9 and 18, green bar demonstrates the position of the BRCA1_9-18FJunc primer. The primer used for sequencing is noted beside each sequencing result.
Figure 3.16: PCRs to confirm parts of isoform Δ1Aq,2p,11q,16p, ▼21.
For all experiments, the standard protocol for products <3kb was used. The forward primer was BRCA1_11a-12FJunc and reverse primer was BRCA1_ins21RJunc. Annealing temperature was 64°C and extension time was 1 minute. A: 3µl of cDNA was used for the PCR reaction. 2µl of PCR product was loaded on a 1% gel. The band indicated was extracted by band stab. B: 1µl of extracted band solution from A was used in this PCR. 2µl of PCR product was loaded on a 2% gel. This reaction was Sanger sequenced. C: Geneious: Map to Reference result of Sanger sequencing from B. Purple bar denotes positions of exon 11 (part a), exon 12, and exon 21. Pink bar denotes position of the intron 21 retention selected from published data. The primer used for sequencing is noted beside each sequencing result.
Chapter 4: Discussion

4.1 Summary of results

This project has confirmed that the MinION nanopore sequencer can successfully read through cDNA of the long \textit{BRCA1} and \textit{BARD1} transcripts. In total, 52 full length \textit{BRCA1} transcripts and 4835 full length \textit{BARD1} transcripts were sequenced. This research has also discovered 24 potential novel \textit{BRCA1} isoforms, and 18 potential novel \textit{BARD1} isoforms. Of the six novel \textit{BRCA1} isoforms chosen for further analysis, five of these had acceptor or donor shifts present in some, if not all transcripts. Two of these were successfully Sanger sequenced, and the multiple exon deletions initially discovered were verified for isoforms Δ10-17 and Δ1Aq,2p,11q, ▼21.

During the course of this project, the full length \textit{BRCA1} transcript was amplified successfully through PCR, and a new data analysis pipeline for long transcript reads was developed. This pipeline uses the BLAT alignment program as its main component, and an additional program (Samcat.py) was developed in order to fully analyse each molecule's sequence data.

4.2 Sample preparation

4.2.1 PCR and Sanger sequencing

PCR of cDNA products was generally successful throughout this project. Short PCR for either end of \textit{BRCA1} generated PCR products that were relatively easy to isolate and Sanger sequence. \textit{BARD1} PCR consistently amplified the same set of bands, and was more difficult to isolate for Sanger sequencing. In order to gain a good representation of \textit{BARD1} products selected bands were extracted and sequenced. However, some of the final sample used for sequencing contained bands that had been amplified at least three times (using 35 cycles of amplification each time), so it was fortunate that the sequencing results turned out well enough that the sequence could be identified with ease.

The primers to amplify the full transcript were designed to be within \textit{BRCA1} exons 1 and 24, and \textit{BARD1} exons 1 and 11. This design does not allow the identification of alternative start or end transcription sites, and so will not identify transcripts such as \textit{BRCA1}-IRIS, which has normal full length characteristics until it terminates shortly into intron 11. The focus during this project was whether this type of analysis was achievable with the MinION instead of trying to obtain a comprehensive survey of all
possible \textit{BRCA1} transcripts. Alternative transcription start and end sites is an area for further research now that the base method of analysis with the MinION has been established.

4.2.2 Full length \textit{BRCA1}

Many different PCR conditions were used to amplify \textit{BRCA1} transcripts. These varied by annealing temperature and extension time, in most cases, and no single reaction produced completely consistent results. The stochastic variation in band patterns on electrophoresis of PCR reaction products indicated the possibility of random sampling during the process, particularly as this variation occurred even when carrying out multiple reactions using the same cDNA under the same conditions. Random sampling could have occurred either by the initial transcripts added to the sample, or in selection of available cDNA templates during the PCR reaction itself, particularly for those transcripts of smaller size. If it is the former, using the entire RNA sample from the cell line may ensure full isoform coverage, but if it is the latter, a pull-down approach may be a more reliable method, and may also be a more quantitative approach. It is likely that both possible explanations are contributing to the variation seen. This project was a proof-of-principle, and did not rely on generating the full range of isoforms produced by \textit{BRCA1}, but this is an element that may be important for future work. The variation present in the long range PCR is testament to the many varied isoforms generated from the \textit{BRCA1} gene.

4.2.3 MinION analysis

The high molecular weight smear present in the \textit{BARD1} PCR samples (Figure 3.3A), became more prominent after the purification step (Figure 3.5). Presumably the smear resulted from nucleic acids of some sort. There was concern that it may affect the sequencing reaction, though there was enough data generated and it did not seem to have an effect on analyses. The low concentration of cDNA in the \textit{BRCA1} sample meant that it had to be pooled with \textit{BARD1} in MinION Sample 1, and many more \textit{BRCA1} PCR reactions were generated in order to have sufficient material for MinION Sample 2, which was solely composed of \textit{BRCA1} products. The difference in amount of \textit{BARD1} PCR products versus those of \textit{BRCA1} becomes strikingly apparent when the number of full length transcripts found for each are compared. Only 52 \textit{BRCA1} full length transcripts were identified, in contrast to 4835 full length \textit{BARD1} transcripts. Though the major cause for the large difference would be the amount of sample contributed for each gene,
another contributing factor is the long length of the full transcript from *BRCA1*. In the PCR reactions the smaller isoforms appeared to be more abundant than the full length, most of them had a molecular weight of 2300 to 2900 base pairs (Table 3.1). This is most likely due to the preferential amplification of smaller products during the PCR reaction, and is not a quantitative measure of relative abundance in the cell.

Sample 2 was run on the same flow cell the day after Sample 1. The second run of the flow cell was not as good as the first (there were fewer pores available), but as we were reusing the same flow cell this was not unexpected. The washing step in between runs was performed twice, but as it is difficult to clean out completely there is still some *BARD1* product appearing in the second sample. This contributes to the reasoning behind combining both data sets together for analysis, as there was no advantage in keeping the two separate.

4.3 Data analysis
The data analysis pipeline was developed using the entire dataset. Due to time constraints, a selection of six *BRCA1* isoforms were chosen for further analysis as *BRCA1* was the main focus of this project. These isoforms were selected based on analysis of the BLAT tabular output, and further analysis was performed using the graphical format and Sanger sequencing.

4.3.1 Iterative development of analysis pipeline
Established methods of analysis for long read data have been optimised for genome sequences and unfortunately are not able to account for the long stretches of “deletions” that occur in transcript data. Established analysis methods for mRNA and cDNA data are based around splice junction detection, and are not suitable for detecting multiple exon alignments and potential intron retentions in one read. Though there has been some research carried out on long read transcript data using the PacBio system I was unable to test these methods on my data set. The nanopore research that looked at alternative splice isoforms of the fly gene *Dscam1* [142] seemed promising to inform my data analysis steps, but there were a number of differences between their research and the goal of this study which affected how the information was used. Firstly, the pattern of splicing in *Dscam1* is very different to that of *BRCA1*. *Dscam1* is divided into exon clusters instead of discrete exons, and within these clusters the exons are mutually exclusive. Bolisetty et. al., (2015) developed several LAST databases as references based on this clustering pattern to accommodate this pattern of splicing [142]. Bolisetty et. al,
(2015) also did not attempt to look for intronic variations in mRNA, which may have been due to the extensive splicing pattern already identified in this gene [142]. It was not clear how they had identified the nearly 8000 unique isoforms of the Dscam1 gene, and they do note that 3515 of these had a read depth of one. Three other genes, with simpler splicing patterns than Dscam1 and BRCA1, were looked at in this paper, with the longest isoform being close to 4kb, but the methods used were similar to those used for Dscam1 and were not able to inform the data analysis for BRCA1 and BARD1 in this research.

The aim of this analysis was to establish a method that is able to map the cDNA reads back to their respective genes, and so be able to determine the exons present in each read. Ideally the method would be able to detect partial exon deletions and any intron retentions present as well, so ultimately a reference file was used that had the entire gene present. Many different methods of sequence analysis were trialled before the BLAT program was eventually settled on as the alignment that gave the best all-round results.

Initially I began to work with the program LAST [147] because at the time it was a well-used program within the MinION Access Program (MAP) community to work with MinION data [136, 142, 157-159]. However, it became apparent that although LAST appears to have specialised parameters for spliced reads it does not display the results very well (Figure 3.7). This may be due to the Multiple Alignment Format (MAF) that the program uses. Other programs gave similar results. BLAST [155], was excellent for giving the tabular summary that was needed to begin to group the results by isoform, though its parameters were too stringent for nanopore data. I decided to continue with BLAT’s more relaxed parameters and an output using the same tabular format as BLAST, rather than pursue experimentation with BLAST parameters to get a similar outcome. This way also guaranteed that the result would be the same as the graphical output of the BLAT alignment.

The BLAT program required two different approaches to obtain a good overall view of the data. A summary of the isoforms present in the data set was required, therefore information on which exons were present in each read was needed. It was for this reason that the reference file was created with individual exon and intron sequences. The next step was to group them by which exons (and introns) were present. These
groups form the read depth of each isoform, required in order to identify which isoforms had high chances of being real.

The tabular data analysis produced information as to whether each read aligned to each exon or intron reference sequence in a “yes or no” format, which does not give information about where exactly each read aligned in each reference. Therefore, visualisation of the results was important to determine alternative splice sites occurring in an isoform group, as demonstrated by the multiple acceptor and donor shifts described as present in BRCA1 transcripts in Table 1.1, and found by this research shown in Table 3.6. Upon initial examination of the graphical output, multiple reads were shown for each exon or intron reference sequence, however it was not feasible to view all the references together. It was this issue that prompted the development of the Samcat.py program. In order to properly analyse each isoform, the .sam files belonging to each group identified in the tabular analysis were manually extracted from the output of Samcat.py, and inspected individually in IGV.

Shorter exons were more susceptible to changes in alignment parameters across all programs including BLAT. When more stringent parameters were applied there appeared to be a loss in the number of exons that registered an alignment, and often the shorter exons were the first to be lost. Using looser alignment parameters allowed more exons to be mapped to each read. This is highly likely to be the result of the high error rate of the MinION, as in shorter exons there is less chance of a good alignment occurring, resulting in these matches being discarded by most programs and default parameters. The shortest exon present in BRCA1 is exon 19, and because this was rarely missing in the final data set of isoforms I believe these BLAT alignment parameters are appropriate for this data.

4.3.2 R and Excel manipulation
Most of my tabular analysis was carried out in the program Excel. The statistical program R was necessary for the initial step of obtaining one line of data per read in order to open the smaller file in Excel for subsequent analysis steps. Given time constraints I was not able to develop the entire process in R, though that would have been preferable.

Ideally the reads would not have needed to be forced into a specified alignment the way they were, but it was the only way to easily obtain a summary of isoforms. The exon
references included the known outer spliced areas in general (such as the initial NAG splice site), so this is likely to minimise the risk of missing splice occurrences. This was an advantage to looking at such a well-studied gene, as the risk of missing identification of new splice events is minimal.

During the Excel analysis, filtering steps were applied which significantly reduced the number of reads available for analysis. One of these steps was removing all the reads that did not contain both exons 1 and 24, as without these exons the information for the transcripts amplified in this study design is incomplete. These reads may have been generated by shearing and fragmentation of the DNA during the library preparation.

4.3.3 Graphical format output and program development
Initially a BLAT alignment was performed against a reference file containing the single gene sequence (i.e. not split up into the multi-fasta file). This did produce alignments in the graphical view, though they seemed to align indiscriminately to introns and exons alike. This method also meant that there were two different reference files used for the same alignment, which may introduce other issues. To attempt to rectify these problems, a graphical output was used with an identical reference and parameters as the tabular output, which retained alignment consistency. This graphical format output initially gave an alignment per exon/intron, and the only way to view which alignment belonged to which read was to hover the mouse over the bar showing in IGV. The program Samcat.py was created in response to this problem. The aim of the program was to concatenate the CIGAR string part of the .sam alignment file in order to create one line per read, as previously there were multiple lines per read corresponding to each reference alignment. To the best of my knowledge, there is no existing tool that is able to combine alignment data like this.

Due to the version of psl2sam.pl program that was used for this study, the resulting .sam file did not contain any sequence information, which meant that though the Samcat.py program was therefore easier to make, it also reduces the ability to check that the final alignment output from Samcat.py is correct. This version also uses M to denote matches and mismatches instead of differentiating the two. This resulted in the graphic associated with these .sam files not distinguishing between matches and mismatches.

This program went through iterative development in order to debug outliers in the data. The types of problems occurring during program development were mostly due to
reads that probably were not from BRCA1 or BARD1, and were aligning everywhere and in any direction. In the end the decision was made to remove reads that were causing an excessive number of problems from the final file. This made the program simpler and at the same time it became useful for filtering out undesired alignments. At this stage, the calculations involved in creating the program, and the program as a whole, have not been independently tested. However, there is confidence that the reads that are of good enough quality to give helpful information will make it through the filtering process inherent in the program’s creation.

The major disadvantage of splitting the reference up into multiple sequences is the alignment of bases at each end of the exon (or intron) that are identical to the bases at the end of adjacent exons. The intron 21 graphical alignment demonstrates this problem exceptionally well (Figure 3.9E). According to the graphical alignment, the first five bases of the intron retention are CACAG. When this portion of the graphical alignment was first being examined, the initial conclusion was that in contrast to the published intronic inclusion in this area, the intron 21 alignment was picking up an acceptor shift. On inspection of the exon 21 sequence, it was found that the last five bases of exon 21 are also CACAG (data not shown). This section of the sequence has aligned twice due to the reference sequences being treated as separate alignments, which is important to remember when looking at Samcat.py generated graphical outputs. This phenomenon can also be seen in Δ16p occurring in the same isoform (Figure 3.9D), and is described in Table 3.6.

In each isoform group investigated there was at least one read which had appeared to belong to the relevant group in the tabular data but did not match correctly in the graphical view. This was apparent due to some of the filtering capability of Samcat.py, and it demonstrates that the initial BLAT alignment to the multi-fasta reference file in the tabular format requires further investigation.

Upon initial observation, there seemed to be many small splice acceptor/donor shifts in the graphical view. On closer inspection, however, these tended to occur in runs of bases, or “slippery” sequences, such as “GGGGG” or “AAAA” (data not shown), and appeared to have little to do with the more standard NAGNAG sites of such splice changes in the 5’ end of the exon alignment. Given the extent and type of error the MinION is currently prone to, many of these “shifts” are most likely errors of sequencing rather than real changes. This theory is supported by the observation that deletions
often occurred at runs of the same base as well. This gives rise to some uncertainty for the Δ14Ap alignment in isoform Δ3,9,10,11q (Figure 3.13B). There appears to be a 7bp shift, but as this is out of frame it is unlikely, and could possibly be representing a 9bp shift instead due to the alignment then beginning with "AA". Exon 13 contains “AAAAAG” at the 3’ end, which could be confounding this alignment. This shift could also be the previously described 3 base pair shift at this position, and this area is not being sequenced properly. There is certainly something out of the ordinary occurring at the 5’ end of exon 14, though the read depth is not enough to be confident about what it is exactly, and any conclusions are confounded both by MinION inaccuracy and the fact that the graphical output aligns identical bases at the edges of exons twice.

On close examination of the exon alignment in the full length reads it was discovered that all reads were missing three bases from the 3’ end of exon 10 (Figure 3.14D). This follows a run of six “A” nucleotides which may have affected the electrical signal of the MinION and subsequent basecalling, especially given the 6-mer increments of sequencing. This change is unlikely to be biologically valid, especially considering this particular change has not been seen before. The multiple combinations of splice shifts apparent in reads that contain all exons is interesting, however, and is important to consider in the characterisation of BRCA1 and associated isoforms. If the variations seen occurring in the graphical data are correct, there may be many more isoforms with small variations that would be difficult to identify.

4.4 Isoform Analysis
Six BRCA1 isoforms were chosen for further analysis by Sanger sequencing. This selection was based on the tabular output prior to the completion of the methodology for the graphical analysis, and therefore focuses only on detecting whole exon skipping events. Only two isoforms went through successful PCR and sequencing. As the BRCA1 PCR was so variable, isoform-specific PCR was performed on cDNA sample directly. In order to detect the smaller changes within exons found through graphical analysis, isoform-specific PCR perhaps should be performed as a nested PCR on BRCA1 amplified samples. The parts Δ11q and ▼21 were confirmed from isoform Δ1Aq,2p,11q,▼21, but the rest of the changes described for this particular isoform have not yet been verified.

All the proposed splice shifts were checked using Human Splicing Finder version 3.0 [160]. In most cases the prediction was identical to my data analysis, and in the rest the prediction was in the same area but slightly shifted by one or two bases (data not
shown). This may reflect errors either in the MinION or in the alignment, and will need to be further investigated.

4.4.1 Functional significance

The commonality in all six isoforms further investigated is the lack of either the entirety of exon 11 or a large portion of it. This was a common feature of most of the isoforms found in the initial tabular analysis. It may be partially due to the fact that PCR preferentially amplified isoforms lacking most of exon 11 because they were significantly smaller. The lack of the majority of exon 11 does have functional consequences however, as exon 11 contains nuclear localisation signals, phosphorylation sites, sites of interaction with other proteins, and has also been found to play a role in DNA repair mechanisms and cell cycle arrest. Isoforms lacking this region may produce proteins that will not be efficiently localised to the nucleus, which means they could have other effects in the cytoplasm [161, 162]. However, if all these isoforms begin translation at the normal start site in exon 2, then very soon after the skipping event translation will stop, as all of these skipping events other than Δ11q create stop codons either at the junction (in the case of Δ3), or very soon after the event. The intron retention encodes a stop codon 56 bases into the retained region. It is difficult to predict what effects all these changes will have on BRCA1 isoform function, though without the repair regions and phosphoprotein binding domains encoded in exon 11 and in later exons, it can be hypothesised that many of these isoforms may display oncogenic features similar to BRCA1-IRIS, which stops translation a short way into intron 11 and does not retain the rest of the transcript [60, 105, 106]. On the other hand, these transcripts may have an unknown regulatory function within the cell, or perhaps even undergo ribosomal “slipping” during translation and manage to get back in-frame, due to the many slippery sequences in BRCA1 exons such as those which the MinION also has trouble reading (described in Section 4.3).

All complete exon skipping events of BRCA1 isoforms, the splice donor shifts Δ1Aq and Δ5q, and acceptor shift Δ2p, that have been identified in this work have been found previously [60]. The difference is that none of these changes have been described in the same transcript as those presented here. The fact that these changes have been previously identified increases the confidence that the results of this study are correct. The other splice shifts identified by the graphical analysis will need further verification to be certain they are correct. Finding events such as these was an unexpected result
from the MinION during this study, and the number of these small shifts observed demonstrates the complexity of splicing at the BRCA1 locus. These smaller variations may not cause as great a difference in functional outcome compared to the larger exon skipping events, but they may still have cellular effects which can only be explained by tiny changes in transcripts. Additionally, these multiple splice shifts observed in a single transcript have important implications for determining the clinical relevance of splicing associated variants. For example, a single shift or exon skipping event may be interpreted as an in-frame deletion which is non-deleterious, but other events in the same molecule may have an out-of-frame effect, changing the functional outcome. Alternatively spliced transcripts previously identified as being low risk for disease might actually be high risk, and *vice versa*. These are possibilities that may now need to be considered when assessing splicing changes associated with BRCA1 variants.

4.5 Limitations of this work

One limitation of the transcript analysis in this thesis was that it was not quantitative due to the differential amplification nature of PCR. This research demonstrates that PCR amplifies a random selection of cDNA molecules. Perhaps a more quantitative study would use a pull-down approach to isolating cDNA, such as using gene specific biotinylated oligonucleotides and streptavidin beads to capture the target transcripts before sequencing on the MinION. There would need to be a very large amount of starting RNA to work with to ensure as many BRCA1 isoforms as possible were collected, and a final amount of 1µg of cDNA to go into MinION library preparation.

Nonsense-mediated decay (NMD) plays a biological role in cell processes deciding which mRNA molecules are destroyed. NMD may vary in function between cell types. It is important to be aware that other cells, such as those in the breast tissue, may have NMD that functions differently to that of a lymphoblastoid cell line. Isoforms found in this NMD-treated cell line may not be detected in an untreated cell line, but they may be a normally expressed isoform in the breast. To test this would require analysis to be carried out on normal breast tissue, and this was not possible during this study.

The MinION read depth was enough to gain an idea of the isoforms present, but due to the current error rate of the MinION, ideally a higher read depth would increase the confidence of the results, particularly of those splice shift events that have not been identified before. There was on average about seven reads per isoform group in the graphical output that actually matched everywhere predicted in the tabular output, so
small differences identified, such as 2-3 base acceptor/donor shifts, are much less certain than the larger ones of 9-12 bases, especially if they only occurred in some of the reads. Nevertheless, two of the identified isoforms have had a number of their non-contiguous splicing events confirmed in this study.

Error correction is available for nanopore data, however this is based on an averaging of the reads which is not suitable for this research. Error correction on isoform groups was considered prior to isoform groups being identified, but was not plausible due to the low read depth of each isoform, and may have masked the smaller changes occurring in some of the reads.

In previous work on cDNA with MinION sequencing, the rate of template switching was quantified [142]. This may have occurred during PCR cycles in BRCA1 and BARD1 samples, however, due to limited time and resources I was not able to test whether template switching of strands during PCR was occurring in my samples. This paper found that the rate of template switching in the Drosophila gene Dscam1 was high with 30 cycles of PCR [142], while I needed to use 35 cycles of PCR in order to gain enough BRCA1 molecules to visualise on an agarose gel. There was a large difference in size of our respective amplicons, the size of the Dscam1 amplicon was 1.8kb whereas BRCA1 and BARD1 were 5.8kb and 2.0kb respectively. Template switching is specific to each reaction depending on the molecule and primers, and would have been of concern for Dscam1 in particular because of the immense variability in possible exons per transcript, but is not so crucial for the amplification of the less variable BRCA1 or BARD1 genes.

Improvements to this project could include: where at all possible mixing the samples containing cDNA by inversion rather than pipetting to reduce amount of DNA shearing, generating significantly more BRCA1 PCR product to run on the MinION (enough to be able to run it independently) in order to have greater read depth per isoform, and developing the whole tabular analysis in R instead of the hybrid R/Excel approach used here.
Conclusion and implications of this research

This research aimed to determine the viability of using the long read MinION nanopore sequencer to help fill the gap in knowledge of the alternative splice transcripts generated by *BRCA1* and *BARD1*. During the course of this study, successful long range *BRCA1* PCR was performed to amplify the entire full length transcript, along with many isoforms, and a bioinformatic pipeline using the BLAT aligner has been developed in order to analyse long transcript reads that span multiple exons and introns. This analysis has also been able to detect both large and small splice shifts in these transcript reads. A new program to deal with the multi-reference graphical output of the alignment was also developed. Overall, this project has been successful in demonstrating the capability of the MinION device to identify these isoforms, and in demonstrating how to use the necessary tools to analyse the sequence data.

Improvements to the data analysis pipeline could be made with a tool for aligning whole gene reference sequences but with the ability to assign regions and get a summary of this data, effectively combining the two branches of my data analysis pipeline together.

The complexity of splicing at the *BRCA1* locus is revealed in this research to be much greater than previously speculated, and this may also be true for many other genes with alternative splicing characteristics. This research opens the door for a more complete annotation and understanding of alternative splicing in mammals. More immediately, it will greatly increase the confidence of *BRCA1* and *BARD1* isoform prediction, furthering the understanding of these critical tumour suppressor genes, and may lead to improvements in detection and prediction of breast cancer risk.


122. Paul MS, Bass BL. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. The EMBO journal. 1998;17(4):1120-7.


Appendix A: Positions of primers described in Table 2.2

A.1 Isoform Δ10-17:

Figure A.1: Figure of primer positions used for amplification and sequencing of isoform Δ10-17. Figure generated using Geneious® 9.1.5, Biomatters Ltd.
A.2 Isoform Δ11q ▼ 21

Figure A.2: Figure of primer positions used for amplification and sequencing of isoform Δ11q ▼ 21.
A shows primers designed for Δ11q, while B shows primers designed for ▼ 21. Figure generated using Geneious® 9.1.5, Biomatters Ltd.
A.3 Isoform Δ3,11Δ3110

Figure A.3: Figure of primer positions used for amplification and sequencing of isoform Δ3,11Δ3110. Figure generated using Geneious® 9.1.5, Biomatters Ltd.
A.4 Isoform Δ11q,14,21

*Figure continued on following page.*
Figure A.4: Figure of primer positions used for amplification and sequencing of isoform Δ11q,14,21.

A shows primers designed for Δ11q,14 while B shows primers designed for Δ21. Figure generated using Geneious® 9.1.5, Biomatters Ltd.
A.5  Isoform Δ3,9,10,11q

Figure A.5: Figure of primer positions used for amplification and sequencing of isoform Δ3,9,10,11q.
Figure generated using Geneious® 9.1.5, Biomatters Ltd.

83
A.6 Isoform Δ11q,19

A shows primers designed for Δ11q, while B shows primers designed for Δ19. Figure generated using Geneious® 9.1.5, Biomatters Ltd.
Appendix B: Pre-Amplicon protocol bead clean-up

Materials:
- MagBio Ampclean beads.
- Make fresh 80% ethanol
- Using 4 KingFisher 96-well plates - every second well
  - Plate 1: DNA plate
  - Plate 2: Wash 1
  - Plate 3: Wash 2
  - Plate 4: Eluent (ddH2O)

Note: Label bottom of wells in each plate you are using with a black dot, so you can see where you are

Methods:
1. Measure volume of DNA, and add to the DNA plate
2. Add 1x volume of vortexed beads (e.g. for 80ul of sample use 80ul of beads). Use higher concentrations for smaller sizes.
3. Mix up and down with the empty KingFisher tip
4. Add 200ul 80% ethanol to Wash 1 and Wash 2 in the appropriate wells
5. Add KingFisher magnetic head to DNA plate (in the tip used before), and wait for the supernatant to clear. Should be done within 5 mins.
6. Transfer magnetic head carefully into Wash 1
7. Transfer it again into Wash 2
8. Add appropriate amount of water to Eluent plate, typically 20-25ul. Depends how concentrated you want it.
9. Remove head from Wash 2 and air dry, want to let the ethanol evaporate but not have the beads so dry they crack. Watch so it doesn't get too dry, look for shinyness.
10. Now add the head to the Eluent plate, then remove the magnet, leaving the tip plate in place.
11. Mix the tips up and down and leave for ~5mins to elute.
12. Put the magnetic head back and wait for the solution to clear - want to get all the beads. Remove the head and tips
14. Measure on Nanodrop/Qubit.

Note: At this stage we want to be able to get 1ug of DNA to go into the MinION library preparation steps.
Appendix C: Amplicon library preparation protocol R9 Version

1. **Library Preparation**
   
a. **Overview of the Nanopore Sequencing protocol**

**Introduction to the Nanopore Sequencing Kit**

**IMPORTANT:**
The products in use with this Experiment Companion

This protocol should only be used in combination with:

- Nanopore Sequencing Kit SQK-NSK007 (R9 Version)
- MinION Flow cells with the code FLO-MIN104 (R9 Version)
- MinKNOW scripts with the prefix NC_ and ending in FLO_MIN104
- Workflows for SQK-NSK007

If using Expansion Packs all are compatible with the exception of the Native Barcoding Kit where the product code EXP-NBD02 is required.

**Introduction to the Nanopore Sequencing Kit**

- The Nanopore Sequencing Kit is designed to prepare genomic, amplicon and cDNA, with or without barcoding, for sequencing on the Oxford Nanopore MinION™.
- The kit features an adapter which is compatible with the Oxford Nanopore motor protein and which must be ligated onto end-repaired and A-tailed fragments. The workflow for nanopore sequencing consists of steps for template preparation and then steps required for adapter ligation.

*Figure NSKS: Schematic of the Nanopore Sequencing Kit protocol.*

The steps in the box are carried out using the Nanopore Sequencing Kit reagents. The input DNA can be genomic, amplicon or cDNA and these can be with or without barcoding.
Storage instructions for the Nanopore Sequencing Kit (SQK-NSK007)

On receipt the kit should be stored at -20 °C.

**IMPORTANT:**
To keep freeze/thawing to a minimum, it is advisable to thaw only the barcodes and adapters required for the particular experiment. The kit should be stored at -20 °C when not in use. Keep on ice when thawed.

b. Nanopore sequencing kit contents

Reagents provided in the kit

**Nanopore Sequencing Kit (SQK-NSK007)**
The kit contains the reagents required to prepare a DNA library for the MinION. There are sufficient reagents for preparing six libraries for loading into the MinION Flow Cell.

The reagents are multi-use tubes. Freeze/thaw of the RBF1, Hairpin Adapter, Adapter Mix and Hairpin Tether should be kept to a minimum, so it is advisable to only thaw the tubes required on the particular day. RBF1 can be stored at 4 °C if carrying out multiple loadings. Each RBF1 tube contains enough buffer for priming the flow cell and preparing the library for loading.

<table>
<thead>
<tr>
<th>Contents</th>
<th>Colour</th>
<th>No. of tubes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapter Mix (AMX)</td>
<td>green</td>
<td>3</td>
</tr>
<tr>
<td>HP Adapter (HPA)</td>
<td>purple</td>
<td>1</td>
</tr>
<tr>
<td>HP Tether (HPT)</td>
<td>purple stripes</td>
<td>1</td>
</tr>
<tr>
<td>Lambda DNA (LMD) at 50 ng/µl</td>
<td>yellow</td>
<td>1</td>
</tr>
<tr>
<td>DNA CS (DCS)</td>
<td>yellow stripes</td>
<td>1</td>
</tr>
<tr>
<td>Elution Buffer (ELB)</td>
<td>black</td>
<td>1</td>
</tr>
<tr>
<td>Material</td>
<td>Color</td>
<td>Quantity</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------</td>
<td>----------</td>
</tr>
<tr>
<td>Bead Binding Buffer (BBB)</td>
<td>red stripes</td>
<td>2</td>
</tr>
<tr>
<td>Running Buffer with Fuel1 (RBF1)</td>
<td>red</td>
<td>6</td>
</tr>
<tr>
<td>Primer Mix (PRM)</td>
<td>grey</td>
<td>1</td>
</tr>
<tr>
<td>PCR Adapters (PCA)</td>
<td>pale blue</td>
<td>1</td>
</tr>
</tbody>
</table>

**IMPORTANT:**
Thoroughly mix the contents of each tube by inversion (vortexing or pipetting for Running Buffer), and spin down very briefly before pipetting to ensure the contents of the tube can be aspirated accurately. For the Running Buffer with Fuel 1 (RBF1) this should be carried out prior to use.

c. **End-prep**

End-repair and dA-tail of double-stranded DNA fragments (~50 minutes)

<table>
<thead>
<tr>
<th>Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA CS</td>
</tr>
<tr>
<td>1 µg amplicon DNA in 45 µl nuclease-free water</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consumables</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEBNext End repair / dA-tailing Module (E7546)</td>
</tr>
<tr>
<td>Freshly prepared 70% ethanol in nuclease-free water</td>
</tr>
<tr>
<td>1.5 ml Eppendorf DNA LoBind tubes</td>
</tr>
<tr>
<td>Nuclease-free water</td>
</tr>
<tr>
<td>Agencourt AMPure XP beads</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal cycler at 20 °C and 65 °C</td>
</tr>
<tr>
<td>Magnetic rack</td>
</tr>
<tr>
<td>Hula mixer (gentle rotator mixer)</td>
</tr>
<tr>
<td>Vortex mixer</td>
</tr>
<tr>
<td>Ice bucket with ice</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optional Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuBit fluorimeter (or equivalent for QC check)</td>
</tr>
</tbody>
</table>

1. Perform end repair and dA-tailing of fragmented DNA as follows:

Mix the following reagents in a 1.5 ml Eppendorf DNA LoBind tube:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>~1 µg DNA (fragmented genomic DNA, amplicon or cDNA)</td>
<td>45 µl</td>
</tr>
<tr>
<td>Ultra II End-prep reaction buffer</td>
<td>7 µl</td>
</tr>
<tr>
<td>Ultra II End-prep enzyme mix</td>
<td>3 µl</td>
</tr>
<tr>
<td>DNA CS</td>
<td>5 µl</td>
</tr>
<tr>
<td>Total</td>
<td>60 µl</td>
</tr>
</tbody>
</table>
If using the barcoding approach the pooled input DNA should be ~1 µg in 45 µl, whether genomic, amplicon or cDNA.

2. Mix gently by inversion and spin down.

3. Transfer the sample to a 0.2 ml PCR tube, and incubate for 5 minutes at 20 °C and 5 minutes at 65 °C using the thermal cycler.

*If condensation is observed in the tube after the thermocycling, briefly spin down the tube contents in a microfuge.*

4. Resuspend AMPure XP beads by vortexing.

5. Transfer the sample to a 1.5 ml DNA LoBind Eppendorf tube.

6. Add 60 µl of resuspended AMPure XP beads to the end-prep reaction and mix by pipetting.

7. Incubate on a rotator mixer (Hula mixer) for 5 minutes at room temperature.

8. Prepare 500 µl of fresh 70% ethanol in nuclease-free water.

9. Spin down the sample and pellet on a magnet. Leaving the tube on the magnet, pipette off the supernatant.

10. Keep on magnet, wash beads with 200 µl of freshly prepared 70% ethanol without disturbing the pellet. Remove the 70% ethanol using a pipette and discard. Repeat.

11. Spin down and replace on magnet to collect and pipette off any residual 70% ethanol. Briefly allow to dry.

12. Remove the tube from the magnetic rack and resuspend pellet in 31 µl nuclease-free water.

13. Pellet beads on magnet until the eluate is clear and colourless.

14. Remove and retain 31 µl of eluate into a clean 1.5 ml Eppendorf DNA LoBind tube.

15. Quantify 1 µl of end-prepped DNA using a QuBit fluorimeter - recovery aim > 700 ng.

**End of Step:**

Take forward approximately 700 ng of end-prepped DNA in 30 µl into adapter ligation.

d. **Adapter Ligation**

Addition of adapters and the Tether (~25 minutes)

**Materials**

- Bead Binding Buffer (BBB)
- Elution Buffer (ELB)
- HP Adapter (HPA)
- HP Tether (HPT)
- Adapter Mix (AMX)

**Consumables**

- NEB Blunt/TA Ligase Master Mix (M0367)
- 1.5 ml Eppendorf DNA LoBind tubes
- Nuclease-free water

**Equipment**

- Microfuge
- Ice bucket with ice
**Fragment size and adapter ligation**
The amount of adapter has been optimised for fragment sizes greater or equal to 8 kb. If the fragments are generally smaller than 3 kb adjustments should be made to use 0.2 pmoles of DNA in the adapter ligation step.

1. Check the content of each tube is clear of any precipitate and are thoroughly mixed before setting up the reaction.
   - Mix the contents of each tube by inversion
   - Check that there is no precipitates present (DTT in the Blunt/TA Master Mix can sometimes form a precipitate)
   - Spin down briefly before accurately pipetting the contents in the reaction
2. Taking the end-prepped DNA, perform adapter ligation as follows, mixing by inversion between each sequential addition.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-prepped DNA</td>
<td>30 µl</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>8 µl</td>
</tr>
<tr>
<td>Adapter Mix</td>
<td>10 µl</td>
</tr>
<tr>
<td>HP Adapter</td>
<td>2 µl</td>
</tr>
<tr>
<td>Blunt/TA Ligation Master Mix</td>
<td>50 µl</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100 µl</strong></td>
</tr>
</tbody>
</table>

3. Mix gently by inversion and spin down.
4. Incubate the reaction for 10 minutes at room temperature.
5. Add 1 µl HP Tether.
6. Mix gently by inversion and spin down.
7. Incubate the reaction for 10 minutes at room temperature.
8. Prepare MyOne C1 beads for the purification of the DNA library during this incubation.

**End of Step:**

**Adapted and tethered DNA library.**
The adapted and tethered DNA library is now ready to be purified prior to loading into the MinION Flow Cell. The purification removes any excess proteins, nucleotides and salts etc. from the DNA library.
e. **MyOne C1 bead buffer exchange**

Shipping buffer exchanged for Bead Binding Buffer (~20 minutes)

<table>
<thead>
<tr>
<th>Materials</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bead Binding Buffer (BBB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consumables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MyOne C1 Streptavidin beads</td>
</tr>
<tr>
<td></td>
<td>1.5 ml Eppendorf DNA LoBind tubes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Equipment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Magnetic rack</td>
</tr>
<tr>
<td></td>
<td>Vortex mixer</td>
</tr>
<tr>
<td></td>
<td>Microfuge</td>
</tr>
<tr>
<td></td>
<td>Ice bucket with ice</td>
</tr>
</tbody>
</table>

1. Vortex MyOne beads and transfer 50 µl to a clean 1.5 ml Eppendorf DNA LoBind tube.
2. Pellet beads on magnet until the eluate is clear and colourless.
3. Pipette off the supernatant and discard. Retain the bead pellet.
   - Pellet the beads on the magnet until the supernatant is clear and colourless
   - Avoid removing the supernatant too early as this will cause loss of the beads available for purification. You can check whether all the beads have pelleted by looking inside the tube from the top down.
4. Add 100 µl Bead Binding Buffer to the pelleted beads, vortex until homogeneous, pellet the beads on the magnet. Pipette off and discard the supernatant. Repeat.
   - This is a wash step which needs to be carried out twice
   - Add 100 µl Bead Binding Buffer to the pelleted beads
   - Resuspend beads by vortexing
   - Place the tube on the magnet and allow the pellet
   - Discard supernatant
5. Add 100 µl Bead Binding Buffer to the pelleted washed beads. Resuspend beads by vortexing.

**End of Step:**
These are the washed beads required for the DNA library purification.

f. **Library purification**

Purification of the adapted and tethered DNA library (~10 minutes)

<table>
<thead>
<tr>
<th>Materials</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bead Binding Buffer (BBB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consumables</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MyOne C1 Streptavidin beads in Bead Binding Buffer (washed beads)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Equipment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Magnetic rack</td>
</tr>
<tr>
<td></td>
<td>Hula mixer (gentle rotator mixer)</td>
</tr>
</tbody>
</table>
1. Add 100 µl of washed MyOne C1 beads to the adapted and tethered DNA reaction and carefully mix by pipetting.

Library purification demo

2. Incubate on a rotator mixer (Hula mixer) for 5 minutes at room temperature.

**Mixing washed MyOne C1 beads and DNA library**
The mixing is gentle but enough for the DNA fragments and the beads to come into contact with each other.

3. Place on magnetic rack, allow beads to pellet and pipette off supernatant.

4. Wash the pellet by resuspending in 150 µl Bead Binding Buffer by pipetting. Pellet on magnet, pipette off and discard the supernatant when clear and colourless. Repeat.

5. Close tube lid, spin down, replace on magnet for 1-2 minutes, open lid and pipette off any residual Bead Binding Buffer.

**End of Step:**
The DNA library is bound to the beads and ready for elution.

**g. Elution of library**

Elution of the library and a final QC check (~15 minutes)

<table>
<thead>
<tr>
<th>Materials</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Elution Buffer (ELB)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consumables</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 1.5 ml Eppendorf DNA LoBind tubes</td>
</tr>
</tbody>
</table>

1. Resuspend pellet in 25 µl of Elution Buffer by pipetting up and down. Incubate for 10 minutes at 37 °C.

2. Pellet beads on magnet until the eluate is clear and colourless.

3. Remove and retain 25 µl of eluate into a clean 1.5 ml Eppendorf DNA LoBind tube.
   • Remove and retain the eluate which contains the DNA library in a clean 1.5 ml Eppendorf DNA LoBind tube
   • Dispose of the pelleted beads

4. Place the tube of library on ice until required for library loading.

**Quantify 1 µl of fragmented and repaired DNA using a QuBit fluorimeter - total recovery aim ~ 250 ng.**

**End of Step:**
Aliquots of the prepared library are used for loading into the MinION Flow Cell. This library is called the Pre-sequencing Mix.
2. **Preparing to load a library**  
a. **Priming the MinION Flow Cell**  

Priming of the sensory array in the flow cell (~20 minutes)

<table>
<thead>
<tr>
<th>Materials</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Running Buffer with Fuel Mix 1 (RBF1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consumables</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• MinION Flow Cell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Equipment</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• MinION Mk I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**IMPORTANT:**
Thoroughly mix the contents of the RBF1 tube by vortexing or pipetting, and spin down briefly.

1. Flip back the MinION lid and slide the sample port cover clockwise to that the sample port is visible.
   Priming and loading a library into a flow cell

   ![Sample Port](image1.png)

   *Figure PFOP: Opening the sample port by rotating the sample port cover. Ensure that the sample port cover is fully opened (a 90° clockwise turn).*

2. After opening the sample port, check for small bubble under the cover. Draw back a small volume to remove any bubble (a few µls). Visually check that there is continuous buffer from the sample port across the sensor array.
   This image shows the MinION Mk I Flow Cell.

   ![MinION Mk I Flow Cell](image2.png)

   *Figure PFLM: Labelled view of the MinION Flow Cell.*

3. Prepare the flow cell priming mix in a clean 1.5 ml Eppendorf DNA LoBind tube.
4. Load 500 µl of the priming mix into the flow cell and wait 10 minutes, avoiding the introduction of air bubbles. Repeat.

**End of Step:**
The flow cell is now ready for the library to be loaded.

### 3. Loading a library
Preparation and loading the library into the flow cell (~5 minutes)

#### Materials
- Adapted and tethered DNA library (Pre-sequencing Mix)
- Running Buffer with Fuel Mix 1 (RBF1)

#### Consumables
- Nuclease-free water
- 1.5 ml Eppendorf DNA LoBind tubes

1. Prepare the library for loading as follows:

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF1</td>
<td>75 µl</td>
</tr>
<tr>
<td>Nuclease-free water</td>
<td>63 µl</td>
</tr>
<tr>
<td>Adapted and tethered library</td>
<td>12 µl</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>150 µl</strong></td>
</tr>
</tbody>
</table>

2. Mix gently by inversion and spin down.
3. Using a Gilson P-1000, load 150 µl of the prepared library into the flow cell keeping the pipette vertical.

4. Close the sample port with the cover and replace the MinION lid.
Appendix D: Data Analysis Methods

D.1 Reference file used

```
>BRCA1_Exon01
GTACCTTGATTTGCTATCTCAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAAT
TACAGATAAATTTAATCTGAGCTCGCGGCTGAGACTTCTCTGGACGGGGGACAGGCTGTGGGGTTTCTC
AGATAACCTGGGCCCTGCGCTCAAGAGGCTCCTACCCTCTCCTCCTCAGTGATAG
>BRCA1_Intron01
GTAGTA.............TTAAAG
>BRCA1_Exon02
TTACCTGGAAGGAAATGGATTTATCCTGCTCTCGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAAT
TACAGATAAATTTAATCTGAGCTCGCGGCTGAGACTTCTCTGGACGGGGGACAGGCTGTGGGGTTTCTC
AGATAACCTGGGCCCTGCGCTCAAGAGGCTCCTACCCTCTCCTCCTCAGTGATAG
>BRCA1_Intron02
GTAGTA.............TGCTAG
>BRCA1_Exon03
TCTGAGCTCTGATCAAGGAACTCTGCTCCACAAAAGGTGACCACATATTTCGCAA
>BRCA1_Intron03
GTAGTA.............TTATAG
>BRCA1_Exon05
ATTTCGCTGATGAACTTCTCAACCAGAAGAAAGGGGCCCTCACAGTGCTCTTTATGAAAGATGATAAAACAAAAAG
>BRCA1_Intron05
GTATAT.............TTTCAG
>BRCA1_Exon06
GAGCCTCAACAGAAAGTGACAGATTTAGTCAACTTTGTAAGAGCTAGTTAAATCCTTGTCTTTTCAAGCTTGACACAG
GTTCAGGT
>BRCA1_Intron06
GTAAAG.............TTTAG
>BRCA1_Exon07
CAGAAACAGCTTATAATTTTGAACAAAAAGAAATCTCTCTCTCCAACACATCTAAAGATGAAATTTTTCTATCATCCAAAAGT
ATGAGCTGAGAAGAGCCAACACGGTACCTACAACTATTTAGGG
>BRCA1_Intron07
GTAAAA.............GTGGAG
>BRCA1_Exon08
CAGAAACAGCTTATCAACACGTTTTATAGTGAAGAGCTATTTTTGACAGCTTTTTCAGCTTGACACAG
>BRCA1_Intron08
GTAAAG.............TTTAAAG
>BRCA1_Exon09
CAGCCTGGAAGGTACTTGCTTTTTTAAATTTAAGGCAACCTTTATGAG
>BRCA1_Intron09
GTGAGT.............CTATAG
>BRCA1_Exon10
TTGCGAGATCAAGAATTTTTTCAACTACACCCTTCAAGGAAGGGAACAGGTTTTGATTTTACGTTCAACAGAG
AAAGACAACCTGCTCTTCACTATTTTGG
>BRCA1_Intron10
GTAAAG.............TTTTCAG
>BRCA1_Exon11a
CTGCTGTGTAATTTTCTTTAGAGACGAGATGAACTAATCTGCAACATCTCAACACGTCAACGGTAAACTTCGTTTGCAACTACGAG
AAAGGGCAGCAGCAGGGCATCCAGAAAGATTTGACAGC
>BRCA1_Intron11b
GTGAGT.............AGTGCA
>BRCA1_Exon11c
CTGCTGTGTAATTTTCTTTAGAGACGAGATGAACTAATCTGCAACATCTCAACACGTCAACGGTAAACTTCGTTTGCAACTACGAG
AAAGGGCAGCAGCAGGGCATCCAGAAAGATTTGACAGC
>BRCA1_Intron11
GTATTG.............TTAAAG
>BRCA1_Exon12
```
>BRCA1_Intron12
GTAAA...........TTGAAG

>BRCA1_Exon13
CAGAGGATACATGGAACATACATGAGAAGTCGATACAGAAGCAGGAGTTGACAGACGAGGAGCTGTGTTAGAACAGCATGGGAGCCAGCCTTCTAACAGCTACCCTTCCATCATAAGTGACTCTTCTGCCCTTGAGGACCTGCGAAATCCAGAACAAAGCACTACGAAAAG

>BRCA1_Intron13
GTGTTG...........TTAAAG

>BRCA1_Exon14
GTCACTCCCTTTCAATATGGCCATATCATTAGATGATGGTGTACATGCACAGTTGCTGGAGCTCTCAGAATAGAAAACACCACAGGAGTTGGAAGG

>BRCA1_Intron14
GTAATA...........TTCTAG

>BRCA1_Exon15
ATGCTGAGTTGGTGGAACGGACACTGAATATTTCTAGGAATTTGCGAGGAAATGGGTAGTTAGCTATTTCTGAGAAG

>BRCA1_Intron15
GTAAATA...........TTCCAG

>BRCA1_Exon16
GGGAAACCCCTTACCTTGGAATCTCGACACCTCTTCTCTGATGACCTGGAGCTCTCAGAATAGAAAACACCACAGGAGTTGGAAGG

>BRCA1_Intron16
GTAGT...........TTCCAG

>BRCA1_Exon17
ATGCTGCTGTTCAAGTTTGCCAGAAACACCCACATCATCACTTTAACTAATCTAATTACTGAAGAGACTACTCATGTGTTTTATGAAACAG

>BRCA1_Intron17
GTATAC...........CTGCAG

>BRCA1_Exon18
ATGCTGATGTGTTGTAACCGGACACTGAATATTTCTAGGAATTTGCGAGGAAATGGGTAGTTAGCTATTTCTGAGAAG

>BRCA1_Intron18
GTAAATA...........TTCCAG

>BRCA1_Exon19
GGGTGACCCAGTCTATTAAAGAAAGAAAAATGCTGAATTGAGAAG

>BRCA1_Intron19
GTAGT...........TTCCAG

>BRCA1_Exon20
CATGATTTTGAAGTCAGAGGAGATGTGGTCAATGGAAGAACCCACCAAGTCGAGAGCAAGAATGCTGCTCAGAGAG

>BRCA1_Intron20
GTAAAG...........TTCCAG

>BRCA1_Exon21
ACTCTCAGGGGCTGAAATCTGTGCTATGCGCCCTTCACACATGCCCACAG

>BRCA1_Intron21
GTAAAG...........TTCCAG

>BRCA1_Exon22
ATCAACTGGAATGGATGGTACAGCTGTGTGGTGCTTCTGTGGTGAAGGAGCTTTCATCATTCACCCTTGGCACA

>BRCA1_Intron22
GTAAAG...........TTCCAG

>BRCA1_Exon23
GGGTGACCCAGTCTATTAAAGAAAGAAAAATGCTGAATTGAGAAG

>BRCA1_Intron23
GTAAAG...........CTCCAG

>BRCA1_Exon24
CAATTG---------CTTCCA
>BARD1_ Exon01
CCTCTG---------GCCTG6
>BARD1_Intron01
GTAAGG---------AACAG
>BARD1_Exon02
TACTAACTCCTGAGAGGCTCTGTGTTTTAGAGGATGAGCACTTCTCTGTAG
>BARD1_Intron02
GTAAGT---------TTAAG
>BARD1_Exon03
TAATTGTGAAGTGACTGCATTTGGAACCTGGATGCAGGTGTTACACCCCGGCTGAGATAAGACTGAAGATAAATA
GACAAGCTGGAGCATGATTCAACTTTTGTAGTAAGGACTTGAAATTTGGCTACATGACATGAGCTGTCAG
>BARD1_Intron03
GTAAGAA--------TGTCA
>BARD1_Exon04
ATTTTG---------TTAAG
>BARD1_Intron04
GTAAGG---------TCTTAG
>BARD1_Exon05
GGCCACATACCTTTGTTGAATACCTTTTACAAAAATGGAAGTGATCCAAATGTTAAAGACCATGCTGGGATGACCCATT
G
>BARD1_Intron05
GTAAGG---------AACAG
>BARD1_Exon06
CATGAAGCTTCATGCGACCTGAAAGGTTAGTGAATTTTGTCCACGATACAAGCCACCGGCTGA
TCACAAATGACTCACCCTCACTCCAGATGCCAGCAAGAAATGGGCACTGGATGATAGCTCAAGCTGTTACTTCTATGGAGCCTTCAGCAAAATCTtü
>BARD1_Intron06
GTAAGT---------TACCA
>BARD1_Exon07
TAATTATTTTGGTCTGCGGCCTGTCGATTATACAGATGATGAAATGCTATTGCTGTCATTACGAGGAAAGATGATGAATCGACTTTGCTGACAGTACAG
>BARD1_Intron07
GTAAGG---------TCTTAG
>BARD1_Exon08
ATGACACCTGGGACCGTACGGGATGACCTCTTGTAGTTATAGCAGCATGACAAAAATGCTCAACATGAGG
TCAAGCTTCAAGTTAACATTTGCTGACAGTACAG
>BARD1_Intron08
GTAAGT---------TCTAG
>BARD1_Exon09
TAACCTAGTTGTTGTTTTCTGTTGTAGTACAGTTCAAGATGATGTTGCTTTGGATTCTCAATGAGCTGGATTCTTCAGTAAATGTAAT
>BARD1_Intron09
GTAAGT---------TCTAG
>BARD1_Exon10
GGGGAAAACAAGTCTACGCGAAGGAAATGATGAAATTCCTGAAAGGTCACACGAGAAGCAGC
CTCAACAAGAACAGCTG
>BARD1_Intron10
GTAATT---------TTTCA
>BARD1_Exon11
TTGCCA---------AATCA
D.2 Combine reads for full analysis

**Objective:** To combine all BRCARD1 reads, 2D only, and 1D reads from the fail folder, and also combine with the same of the BRCA1 folder.

Want a copy of each of BRCARD1D.fasta, BRCARD2D.fasta, BRCA11D.fasta and BRCA12D.fasta in /media/projects/Lucy/Reads/final_group

```bash
# Make the fasta files from MinION reads
poretools fasta --type 2D /media/gsfl/MinION/BRCARD_1/pass > BRCARD2D.fasta
poretools fasta --type all /media/gsfl/MinION/BRCARD_1/fail > BRCARD1D.fasta
poretools fasta --type 2D /media/gsfl/MinION/BRCA1/pass > BRCA12D.fasta
poretools fasta --type all /media/gsfl/MinION/BRCA1/fail > BRCA11D.fasta

# Concatenate the files to create a file with all the reads contained
cat BRCARD2D.fasta BRCARD1D.fasta BRCA12D.fasta BRCA11D.fasta > all.fasta

# Number the files
awk '/>/ {i=i+1} 1' all.fasta > allnum.fasta
awk 'NR % 2 != 0' all.fasta > alltrackd.txt

# Changed names manually at this stage: all.fasta to all_orig.fasta to retain the original file, then allnum.fasta to all.fasta for general use.
```

D.3 BLAT analysis with parameters from UCSC website

**Objective:** To align the reads to my BRCARD reference with both exons and introns in separate sequence lines, including exon 11 split into 3 parts. This reference is called BRCARD1genetranscript.fasta and is found in the Y drive.

BLAT commands `<file://Y:\Lucy\Data\scripts\BLAT.sh>`

```bash
# Do BLAT alignment
# Usage: /media/projects/Lucy/programs/blatSuite/blat [ref.fasta] [reads.fasta] [options] [name_of_file]
cd /media/projects/Lucy/Alignments/Blat/final
cp /media/gsfl/Lucy/Data/RefSeqs/BRCARD1genetranscript.fasta BRCARD1genetranscript.fasta
cmp /media/gsfl/Lucy/Data/Reads/all.fasta all.fasta

# pslx file
/media/projects/Lucy/programs/blatSuite/blat BRCARD1genetranscript.fasta all.fasta -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -out=pslx allgenet.pslx

# blast8 file
/media/projects/Lucy/programs/blatSuite/blat BRCARD1genetranscript.fasta all.fasta -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -out=blast8 allgenet.blast8

# Convert pslx file to sam file and samtools process - need to put in the header separately (samtools view -ht)
/media/projects/Lucy/programs/psl2sam.pl allgenet.pslx > allgenet.sam
samtools faidx BRCARD1genetranscript.fasta
samtools view -ht BRCARD1genetranscript.fasta.fai allgenet.sam
samtools view -T BRCARD1genetranscript.fasta -bS allgenet.sam | samtools sort - allgenet.sorted
samtools index allgenet.sorted.bam
```
D.4 Separate BRCA1 and BARD1 out into different files

Python script to separate out BRCA1 and BARD1 alignments from BRCARD1 references:
<file://Y:\Lucy\Data\scripts\Isoforms.py>

Note: green means fields to change depending on what I want. (Name of file, and BRCA1/BARD1)

```python
import csv

# opens the file produced by blast8 - make sure extension is .txt
with open("allgenet.blast8", 'r') as tsvin:
    allexonstab = csv.reader(tsvin, delimiter='\t')
    allexons = list(allexonstab)
    tsvin.close()

# extract all lines containing the gene it's referenced to. Change gene names appropriately
allexonsGENE = []
for line in allexons:
    for i in range(0, len(line)):
        if "BRCA1" in line[i]:
            allexonsGENE.append(line)

file = open("allgenetBrca1.csv", "w")
for item in allexonsGENE:
    file.write("%s\n %",".join(item))
file.close()
```

D.5 R script

Objective: R script to get all exon hits on one line per read:
<file://Y:\Lucy\Data\scripts\R_manipulation.R>

Note: green means fields to change depending on what I want.

```r
# import from the split file python made
# Add header to file
colnames(allgenetBrca1) = c("readname", "ref", "%identity", "alignment_length", "mismatches", "gap_opens", "query_start", "query_end", "subject_start", "subject_end", "evalue", "bit_score")

# Expand ref column into matches per exon
refCountsBRCA1 = dcast(allgenetBrca1, readname~ref)

# export to csv txt file - saves in Documents
write.csv(refCountsBRCA1, "BRCA1genetTable2.txt")
```
D.6 Excel methods

Excel methods for each set, BARD1 and BRCA1:

1. Import table and organise Exon columns into the right order, then Introns following that immediately. Copy sheet twice, labelled FullLengthTable and IsoformsTable.

2. **FullLengthTable**: To get the right number of full length reads, changed all 0's to 1's in both Exon 1 and Exon 11/24 columns (BARD1/BRCA1). Copy worksheet, name copy to CatFullLength (short for concatenate). Insert a column break between Exon block and intron block in both sheets.

3. **CatFullLength**: Above the reference names put exon number followed by comma (when concatenated for easy reading), e.g. "1,". Use lowercase letters for the introns (e.g. a, b, c, etc.)

   Use this "if" statement for each cell in the exon block: IF(FullLengthTable!B3=0, CatFullLength!B$1, ""). This states that if the corresponding cell in the FullLengthTable is 0, put the contents of what is above the reference into the cell that the statement is in. Expand so formula covers all cells in table. See Below

4. **CatFullLength**: Do something similar for the intron block. If statement: IF(IsoformsTable!AB3=0,"",CatIsoforms!AB$1). This states that if the corresponding cell in the FullLengthTable is 0, leave the cell blank, otherwise put the contents of what is above the reference into the cell that the statement is in. Expand so formula covers all cells in table. This will result in the cell being filled by the letter corresponding to the intron that is included in the isoform.

5. **CatFullLength**: At end of Exon references, concatenate all the cells in the row to get a list of missing exons, using the CONCATENATE function. Name this column "DelExons". Do the same for the intron block, naming it InsIntron. In the column after this, use the CONCATENATE function again on both the DelExons column and InsIntron column to get a list of both deleted exons and inserted introns. Label it DelExons+InsIntrons.

6. **SortedFullLength**: Select the two columns containing row names and concatenated reference names, copy them into a new sheet called SortedFullLength. Out of these two columns make a pivot table. In the PivotTable Fields toolbar drag "DelExons+InsIntrons" into the ROWS box, and "readname" into the VALUES box. Ensure the readnames is set to "Count of readnames" (not sum).

7. **SortedFullLength**: Sort "count of readname" column largest to smallest. Find the number that corresponds to the empty cell under Row Names, and this is the correct number of full length reads from this dataset.

8. **IsoformsTable**: Get rid of reads that have a 0 at either the first or last exon by sorting the table by the first exon column smallest to largest, deleting all rows with 0 in this column, and repeat with the last exon column. Copy worksheet, name copy to CatFullLength (short for concatenate). Insert a column break between Exon block and intron block in both sheets.

9. Repeat steps 3, 4, 5 and 6 but with name Isoforms instead of FullLength.

10. **SortedIsoforms**: Sort "count of readname" column largest to smallest. This is the list of isoforms present in the data.

**Result**: BRCA1 Isoforms file & BARD1 Isoforms file
D.8 Samcat.py Program

```python
#!/usr/bin/env python

import csv
import re
import collections

Alignment.py: Takes a reference file and an alignment file in SAM format and combines each set of reads into a single cigar string

# This program requires an additional comma separated file containing the reference names used to generate the
# SAM file, and the associated size of the reference in base pairs. The Reference file must be in order.
# eg. Reference,96
# This program works by looking first at all the alignments occurring with the same reference and the same read,
# concatenating these CIGAR strings, then looking at the alignments occurring for each reference within the same read
# and concatenating these CIGAR strings in the order presented in the additional reference file.
__author__ = "Cade Picard"

# Opens the file containing the reference name used and the size of the reference
# in base pairs, in csv format.
# Returns a list of lists of the references e.g. [['BRCA1_Exon1',213],['BRCA1_Intron1',1155]]
def get_references():
    with open("RefSizes.csv") as file:
        refReader = csv.reader(file)
        references = list(refReader)
        file.close()
    return references

# Returns a dictionary of the reference file given
# Key: exon name
# Value: reference length
def generate_reference_dict(references):
    ref = {}
    for i in range(0, len(references)):
        ref[references[i][0]] = [references[i][1], i]
    return ref

# Finds all rows pertaining to a single read, removes them from the list of all reads. Returns 2 lists of lists.
# Output = rows pertaining to the read
# Rows = rest of the remaining reads from the file
def find_all_rows(readNumber, rows):
    output = []
    indexes = []
    for i in range(0, len(rows)):
        if int(rows[i][0]) == int(readNumber):
            output.append(rows[i])
            # Add index to the indexes array for deletion
            indexes.append(i)
    indexes.reverse()
    # Delete the rows pertaining to the read in reverse order
```

101
for index in indexs:
del rows[index]
return output, rows

# Sort the rows by the order of the references in the reference csv file. Returns sorted list
# Output = sorted list of all the rows for the read
def sort_rows(rows, references):
    output = []
    # Sorts the reads by the name of the reference in the order specified in the reference file
    for i in range(0, len(references)):
        for j in range(0, len(rows)):
            if references[i][0] == rows[j][2]:
                output.append(rows[j])
    # Sorts references with the same name by the references start value in ascending order
    k = 0
    while k < len(output) - 1:
        if output[k][2] == output[k + 1][2]:
            if int(output[k][3]) > int(output[k + 1][3]):
                temp = output[k]
                output[k] = output[k + 1]
                output[k + 1] = temp
                k = 0
                k += 1
    return output

# Find all numbers associated with M, =, and X in a CIGAR string. Total them up and return the total
def count_matches(cigarString):
    total = 0
    matchM = re.findall(r'\d+M', cigarString)
    matchEquals = re.findall(r'\d+=', cigarString)
    matchX = re.findall(r'\d+X', cigarString)
    for num in matchM:
        total += int(num)
    for num in matchEquals:
        total += int(num)
    for num in matchX:
        total += int(num)
    return total

# Find all numbers associated with I in a CIGAR string. Total them up and return the total
def count_insertions(cigarString):
    total = 0
    matchI = re.findall(r'\d+I', cigarString)
    for num in matchI:
        total += int(num)
    return total

# Find all numbers associated with D in a CIGAR string. Total them up and return the total
def count_deletions(cigarString):
    total = 0
    matchD = re.findall(r'\d+D', cigarString)
    for num in matchD:
        total += int(num)
    return total
# Calculate the length of the gap from the previous aligned reference end to the
current aligned reference start

def calculate_length(reference, stopPoint, startPoint=0):
    total = 0
    for i in range(startPoint, stopPoint):
        total += int(reference[i][1])
    return total

# Combining the CIGAR string for the first row with the CIGAR string for the
second row within
# alignments of the same read.
# firstCigar = The cigar string of the first row to be combined
# referenceLength1 = The length of the reference that is associated with the
first row
# referenceStart1 = The start position of the alignment in the first row
# secondCigar = The cigar string of the second row to be combined
# gapLength = The length of the gap from the previous aligned reference end to
the current aligned reference start.
# This is how the program accounts for Intron gaps and Exon skipping.
# referenceStart2 = The start position of the alignment in the second row
# lastCigar = The last cigar string that was combined with the overall cigar
string, before it was combined

def calculate_new_cigar_string(firstCigar, referenceLength1, referenceStart1,
secondCigar, gapLength, referenceStart2, lastCigar=
""):  
    # Find all numbers associated with H in the 3 CIGAR strings.
    firstCigarHList = re.findall(r'(\d+)H', firstCigar)
    secondCigarHList = re.findall(r'(\d+)H', secondCigar)
    lastCigarHList = re.findall(r'(\d+)H', lastCigar)

    # Check if cigar2 is inside cigar1, if so, use cigar1
    if int(secondCigarHList[0]) > int(lastCigarHList[0])
       and int(firstCigarHList[1]) < int(secondCigarHList[1]):
        return firstCigar

    # a = The gap left until the next reference starts.
    # Calculated by taking the Reference (Exon/Intron) length (from file) MINUS
    # the start position of alignment1
    # MINUS 1. Gives the number of bases left for the read to align to. The
    # matches and deletions tell how far into the
    # reference the alignment goes.
    a = int(referenceLength1) - (int(referenceStart1) - 1) -
       count_matches(lastCigar) - count_deletions(lastCigar)

    # b = The gap between the beginning of the second reference and where the
    # alignment starts.
    # Calculated by taking the gapLength and ADDING the start position of
    # alignment2 MINUS 1
    b = gapLength + int(referenceStart2) - 1

    # totalGap = The total gap (deletions)
    totalGap = a + b

    # Decide whether it is an insertion or deletion between the two reference
    # alignments (will usually be a deletion).
    newCigarValue = 
    if totalGap > 0:
        newCigarValue = str(totalGap) + "D"
    elif totalGap < 0:
        newCigarValue = str(abs(totalGap)) + "I"

    # Construct the new cigar string out of the first string and the second
    # string, with the newCigarValue inbetween.
    # Also the last H from the first string and the first H from the second string
    # are removed.
    return firstCigar[:(len(firstCigarHList[1]) + 1)] + newCigarValue +

""
def construct_new_cigar_string(rows, refs, references):
    cigarString = ""
    for i in range(0, len(rows) - 1):
        # The length of the gap from the previous aligned reference end to the current aligned reference start.
        gapLength = calculate_length(references, refs[rows[i + 1][2]][1],
            refs[rows[i][2]][1] + 1)
        if i == 0:
            cigarString = calculate_new_cigar_string(rows[i][5],
                refs[rows[i][2]][0], rows[i][3], rows[i + 1][5],
                gapLength, rows[i + 1][3], rows[i][5])
        else:
            cigarString = calculate_new_cigar_string(cigarString,
                refs[rows[i][2]][0], rows[i][3], rows[i + 1][5],
                gapLength, rows[i + 1][3], rows[i][5])
        # If there is only 1 row in the read, add the only rows cigar string to the cigar string
        if cigarString == "":
            cigarString = rows[0][5]
        return cigarString

# Add in 0H for the beginnings and ends that don’t have a number so the program doesn’t break
def add_zero_Hs(line):
    indices = [m.start() for m in re.finditer('H', line[5])]
    if line[5].find("H") == -1 or indices[-1] != len(line[5]) - 1:
        line[5] += "0H"
    if line[5].find("H") == len(line[5]) - 1:
    return line

# Main function that starts the program
if __name__ == "__main__":

    # SAM file BLAT output file to be read and combined
    filename = "over6k.sam"
    with open(filename) as file:
        reader = csv.reader(file, delimiter = \'\t\')
        # Convert the tab delimited data into a list of lists
        data = list(reader)
        file.close()

        # Get the references from the reference file
    references = get_references()
    # Generate a dictionary of the references
    dictionary = generate_reference_dict(references)
    # Create a template row for the output
    template = data[0]
    # While Loop Variable
    endReached = False
    # Original length of the SAM file data
    length = len(data)
    # Open the output file
    thefile = open('Output- ' + filename, 'w')
    while not endReached:
        # Get the first read number
        readNum = data[0][0]
        # Find all the rows for the read number
        rows, data = find_all_rows(readNum, data)
        # Sort the rows into the correct order
        output = sort_rows(rows, references)
# Find out which is the dominant read direction of the read

zero = 0
sixteen = 0

for line in output:
    if line[1] == "0":
        zero += 1
    if line[1] == "16":
        sixteen += 1

readOrder = "0" if zero > sixteen else "16"

i = 0
while i < len(output) - 1:
    # Add 0H's to the rows cigar strings if necessary
    output[i] = add_zero_Hs(output[i])
    output[i + 1] = add_zero_Hs(output[i + 1])

    # If the read order is not the dominant direction then delete the row
    if output[i][1] != readOrder:
        del output[i]
        i -= 1

    # If the rows have the same reference name
    elif output[i][2] == output[i + 1][2]:
        # Return whichever row has the most matches
        if count_matches(output[i][5]) >= count_matches(output[i + 1][5]):
            newline = output[i]
        else:
            newline = output[i + 1]
            output[i] = newline

        # Delete the unused line from the output
        del output[i + 1]
        i -= 1
    i += 1

# Calculate the gapLength to the first reference in the output
startNum = calculate_length(references, dictionary[output[0][2]][1]) + int(output[0][3])
newCigarString = construct_new_cigar_string(output, dictionary, references)

# Change the template to have the correct information for the read
template[0] = output[0][0]
template[1] = output[0][1]
template[2] = "BRCARD1_geneseq"
template[3] = str(startNum)
template[5] = newCigarString

# Write the line to the output file
thefile.write("\t".join(template) + "\n")

# Once the list of SAM BLAT data is empty exit the while loop and close the output file
if len(data) == 0:
    endReached = True
    thefile.close()
Appendix E: BRCA1 Long Range PCR variations

Figure E Part 1: Varying long range BRCA1 PCR results.
Full description under Part 2.
Figure E Part 2: Varying long range BRCA1 PCR results.
Samples that underwent MinION sequencing in Sample 1 or Sample 2 are indicated. Size in bp of relevant bands in the marker lane (Ladder) is indicated on each figure. 2µl of each reaction was run on 1% agarose gels. A: Touchdown protocol based on the alternative PCR for full length BRCA1. Annealing temperature started at 66°C and finished at 59°C. B: Standard PCR for full length BRCA1 protocol and alternative PCR protocol comparison, with MPW controls for each. C: Standard PCR for full length BRCA1. D: Multiple replicates of the standard PCR for full length BRCA1 performed at the same time. E: Temperature gradient based on the standard protocol for full length BRCA1. Annealing temperature is displayed for each sample. F and G: Standard PCR for full length BRCA1 multiple replicates performed at the same time.