UNIVERSITY *of* OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND
SAPERE AUDE

Targeted Enrichment and Nanopore Sequencing of RNA Transcripts from Breast Cancer Risk Genes

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

Department of Pathology and Biomedical Science

University of Otago, Christchurch

Jessie Jie-Youen Chang

October 2018

Abstract

Aberrant alternative splicing of hereditary breast and ovarian cancer genes has been associated with an increased susceptibility for breast and/or ovarian cancer. However, studying the effects of gene variants on splicing patterns proves to be difficult for classifying Variants of Uncertain Significance (VUS) – variants which lack the evidence to be categorised as benign or pathogenic. This is particularly the case for the *BRCA1* gene, which is poorly expressed, with a long transcript length.

In this study, a novel RNA enrichment technique using the Oxford Nanopore MinION long-read sequencer was developed and tested to alleviate these limitations. The main hypotheses were that direct RNA sequencing can identify full-length splice isoforms of genes of interest, and that these results can be enhanced using a novel gene enrichment technique.

Total RNA was extracted from a control lymphoblastoid cell line and reverse-transcribed into cDNA. Second-strand cDNA synthesis was carried out by incorporating a T7-polymerase binding site attached to a gene-specific primer for the genes of interest. *In vitro* transcription was carried out with T7-polymerase to enrich the RNA isoforms, which were then sequenced with the Direct RNA Sequencing Kit (Oxford Nanopore Technologies, UK). Several bioinformatic pipelines were evaluated, as bespoke pipelines for analysing this novel dataset were unavailable. The final pipeline consisted of alignment to the Human Genome Reference Build 38 with Minimap2, conversion to BAM files, sorting and indexing with Samtools. The output file was visualised on the Integrative Genomics Viewer, and the reads were manually counted and processed in Excel.

The results exhibited evidence of transcript enrichment for three genes of interest (*BRCA1*, *RAD51C*, *CHEK2*) and a control gene (*GAPDH*). Of note, *GAPDH* presented isoform abundance ratios which were similar in three separate experiments, indicating potentially quantitative and reproducible linear enrichment. Full-length isoforms of *GAPDH*, *RAD51C* and *BRCA1* were able to be sequenced with the Direct RNA Sequencing Kit. Additionally, novel isoforms which have not been previously described for *RAD51C* were detected, which include Δ 3,5-6, Δ 3,7 and Δ 4,7

Novel methodologies and proof-of-principle experiments were executed to enrich and analyse specific transcripts using direct RNA nanopore sequencing. The enrichment method should be further validated with other approaches such as RT-qPCR and optimised for use with relevant genes. Provided that a bespoke pipeline for long direct RNA sequencing becomes available, this novel method may provide a powerful new approach for interpreting the clinical significance of variants of uncertain significance which may impact on splicing patterns.

Acknowledgements

I owe my utmost gratitude to my six phenomenal supervisors who have successfully delivered me through the first phase of my research journey. I am indebted to your generous time commitment and encouragement throughout my project.

Firstly, I would like to extend my greatest gratitude to the two pillars of the supervisory team – Professor Martin Kennedy and Dr Logan Walker. Your enthusiasm and zeal for science has never ceased to amaze me and will be forever ingrained in my heart as I continue this journey of research. It is because of you that I have found a sense of positivity and excitement for this project. I would like to express my deepest gratitude for Ms Allison Miller and Dr Vanessa Lattimore for teaching me the foundations of research, which I am sure will provide a solid cornerstone for my future research experience. Thank you both for maintaining my sanity throughout the year by encouraging me at all times. I would also like to thoroughly thank my 'back-end' supervisors, Dr Simone Cree and Associate Professor John Pearson, who have provided me the ability to carry out the second half of my project. Thank you for answering all my naïve questions about bioinformatic analyses – you have taught me how integral this process can be in research. Without you, I would have never taken my project to the level that it is at currently. I would also like to sincerely thank Dr Amanda Spurdle for donating the cell line that I used in this project.

I give my sincere congratulations to my fellow colleagues, who have achieved so much throughout this year, and would like to commend them for their tremendous work ethics and ambition for science. I cannot wait to see what the future holds for each and every one of you.

My honours experience was only so wonderful due to many minds behind the programme. I give my sincere thanks to Professor Madhav Bhatia, Professor Mark Hampton, Dr Margaret Currie, Ms Alice Milnes and Ms Bridget de la Haye for organising such a valuable programme.

Finally, I would like to acknowledge and thank my family for being my ultimate support structure throughout this year. Without your sacrificial mentality, I would not have been able to pursue research.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLAT	BLAST-like Alignment Tool
Bps	Base pairs
BRCT	BRCA1 C-Terminus (functional protein domain)
°C	Degrees Celsius
cDNA	Complementary deoxyribonucleic acid
DNA	Deoxyribonucleic acid
DRS	Direct RNA sequencing
dsDNA	Double-stranded DNA
GSP	Gene specific primer
IGV	Integrative Genomics Viewer
KB	Kilobases
LCL	Lymphoblastoid cell line
MPW	MilliPore water
mRNA	Messenger ribonucleic acid
NFW	Nuclease-free water
NMD	Nonsense-mediated decay
Oligo(dT)	Poly-Thymines that are complementary to the Poly-A tail on an mRNA
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
RINe	RNA integrity number equivalent score
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
snRNP	Small nuclear riboprotein
SR proteins	Serine/Arginine-rich family of nuclear phosphoproteins
Tm	Melting temperature
µL	Microlitre
µM	Micromolar
VUS	Variants of Uncertain Significance
V	Volts

Chapter 1: Introduction

1.1 Overview

Breast cancer is a prevalent disease among the global population, thus, it is of importance to study the mechanisms behind this disease. Studies within the past two decades have identified that the onset of this disease is linked to pathogenic variants of hereditary breast and ovarian cancer genes, which include *BRCA1*. While genetic screening tests can detect gene variants in such genes, determining the significance of these variants proves to be difficult (Duzkale et al., 2013). A group of variants named Variants of Uncertain Significance (VUS) have an uncertain clinical significance and can cause complications for clinical action. Alternative splicing of pre-mRNA has been linked to the onset of breast cancer (Menon & Omenn, 2010; Stickeler, Kittrell, Medina, & Berget, 1999), and provides important evidence for classifying gene variants. However, determining which mRNA splice changes are abnormal and potentially deleterious can be challenging.

One main challenge can be attributed to RNA sequencing limitations. For instance, studying large genes such as *BRCA1* reveals difficulties with short-read sequencers such as Sanger or Illumina. Fortunately, the advent of nanopore sequencing technology developed by Oxford Nanopore Technologies (ONT) has mitigated the limitations of standard sequencers and may prove to be a better alternative. Nanopore sequencing has prospects of delivering affordable and accessible long-read Direct RNA Sequencing (DRS); the direct sequencing of RNA molecules. The second challenge can be associated with the low expression levels of the genes of interest, leading to an increased difficulty in analysing different isoforms. The purpose of this project was to use the MinION nanopore sequencer to develop a novel method to explore the potential of long-read DRS for understanding effects of VUS on RNA splicing. Thus, the use of this method to identify the effect of rare VUS impacting on RNA splicing may be valuable and ground-breaking in the fields of oncology and genetics.

1.2 Breast cancer statistics

In the 2018 report by the International Agency for Research on Cancer, female breast cancer was determined as the second most prevalent cause of cancer death worldwide, comprising 11.6 % of all cancer mortalities across 185 countries. Furthermore, it is estimated that approximately 2.1 million new cases of female breast cancer will arise in 2018 (F. Bray et al., 2018). In contrast, male breast cancer has found to only comprise around 0.5-1 % of total breast cancer cases (Yalaza, Inan, & Bozer, 2016). Similar trends have been observed in the New Zealand Ministry of Health dataset in 2015, where breast cancer was a part of the ten most prevalent cancers in the new cancer registration (New Zealand Ministry of Health, 2017). Age-standardised registration rates show that breast cancer constitutes approximately 51.6 registrations per 100,000 of the total population. Males contributed to 0.7 registrations per 100,000 of the male population and females contributed to 98.4 registrations per 100,000 of the female population, which shows that female registrations dominate over male registrations in New Zealand. Registration rates of the Maori and non-Maori population were 70.5 registrations per 100,000, and 49.5 registrations per 100,000 respectively, which shows that a greater proportion of breast cancer patients are of Maori descent in New Zealand (New Zealand Ministry of Health, 2017). Although absolute numbers of New Zealand breast cancer cases are low, New Zealand female breast cancer rates are among the highest in the world, which needs to be addressed in the near future (F. Bray et al., 2018).

1.3. Hereditary breast and ovarian cancer genes

Hereditary breast cancer accounts for 5-10 % of all cases of breast cancer (F. Bray et al., 2018) and the associated genes include *ATM*, *BARD1*, *BRCA1*, *BRCA2*, *BRIP1*, *CHEK2*, *FANCM*, *NBN*, *PALB2*, *RAD51B*, *RAD51C* and *RAD51D*. The list of full names of these genes are outlined below (**Table 1.1**). These genes have a wide range of different functions, transcript lengths and expression levels. Furthermore, pathogenic variants of hereditary breast and ovarian cancer genes increase the risk of developing breast and/or ovarian cancers.

Table 1. 1 List of hereditary breast and ovarian cancer genes with common and full names

Common gene name	**Full gene name**
ATM	Ataxia Telangiectasia Mutated gene
BARD1	*BRCA1*-Associated Ring Domain Protein 1 gene
BRCA1	Breast Cancer Type 1 Susceptibility Protein gene
BRCA2	Breast Cancer Type 2 Susceptibility Protein gene
BRIP1	*BRCA1* Interacting Helicase 1 gene
CHEK2	Checkpoint Kinase 2 gene
FANCM	Fanconi Anaemia Complementation Group M gene
NBN	Nibrin gene
PALB2	Partner and Localiser of *BRCA2* gene
RAD51B	*RAD51* Paralog B gene
RAD51C	*RAD51* Paralog C gene
RAD51D	*RAD51* Paralog D gene

Certain variants of hereditary breast and ovarian cancer genes have been found to affect the function of the resulting protein by disrupting mechanisms such as transcription (Nadeau et al., 2000) and translation (Dacheux et al., 2013). One example of a post-translational modification affected by gene variants is phosphorylation (Tram, Savas, & Ozcelik, 2013). Phosphorylation is the attachment of a phosphoryl group, which regulates the function of many enzymes (Burnett & Kennedy, 1954). It is important for gene function, where missense mutations may cause disruptions by altering kinase recognition and phosphorylation patterns of the cell cycle.

BRCA1 was identified in Marie Claire-King's laboratory in 1990 and it is amongst the most widely studied of these genes (Hall et al., 1990). The *BRCA1* gene is located on chromosome arm 17q at region 2 band 1 (Hall et al., 1990; Narod et al., 1991) and contains a Zinc Finger CH3C4 Type (RING Finger) domain (Miki et al., 1994) and a *BRCA1* C Terminus (BRCT) domain (Bork et al., 1997; Koonin, Altschul, & Bork, 1996) (**Figure 1.1**). Functions of *BRCA1* are well studied, which include being involved in DNA repair (Scully et al., 1997), checkpoint control (Yu & Chen, 2004), ubiquitination (Morris & Solomon, 2004), chromatin remodelling (Bochar et al., 2000) and tumour suppression.

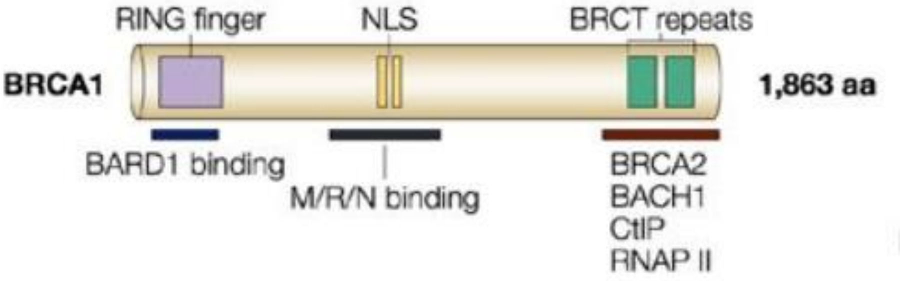


Figure 1. 1 The structure of the BRCA1 protein and its interactions.

The protein is composed of 1853 amino acids. BARD1 binds to the RING finger domain, BRCA2, BACH1, CtIP, RNAP II bind to the BRCT repeats and the MRN complex binds to the Nuclear Localisation Signals (NLS). Reproduced with permission (West, 2003).

In 1994, *BRCA2*, which functions in DNA repair of double-strand breaks by homologous recombination (Patel et al., 1998; F. Xia et al., 2001), was localised to 13q12-q13 (Wooster et al., 1994). The BRCA2 protein interacts closely with the RAD51 protein (Sharan et al., 1997). *RAD51* is a well-conserved gene, with a protein which functions in homologous recombination and double-stranded break repair (Shinohara, Ogawa, & Ogawa, 1992). The gene has paralogs such as *RAD51B* (Albala et al., 1997), *RAD51C* (Dosanjh et al., 1998) and *RAD51D* (Pittman, Weinberg, & Schimenti, 1998), which are hereditary breast and ovarian cancer genes. *ATM* encodes for a Serine/Threonine kinase, which is activated by DNA damage including double-stranded breaks (Shiloh, 2006). This gene was found to be important for regulating the phosphorylation of *BRCA1* (Cortez, Wang, Qin, & Elledge, 1999) and regulation of *CHEK2* (Matsuoka, Huang, & Elledge, 1998). *CHEK2* encodes for the protein Checkpoint Kinase 2 (CHK2), which is phosphorylated in response to DNA damage, in an *ATM*-dependent manner (Matsuoka et al., 1998). It is known to phosphorylate Serine 988 of BRCA1 after DNA damage, which allows the dispersal of the BRCA1 protein (Zhang et al., 2004). Furthermore, it is important in restoring survival (J. S. Lee, Collins, Brown, Lee, & Chung, 2000) and is involved in cell cycle regulation at the G2 phase. *FANCM* is one of the Fanconi Anemia (FA) genes, known to repair the inter-strand cross-linking regions by acting as a helicase or translocase and recruits the core FA complex (Meetei et al., 2005). BARD1 protein relocates BRCA1 to DNA damage sites through tandem BRCT modules, which may allow repair of DNA regions in cells which have lost phosphorylation-dependent pathways (M. Li & Yu, 2013).

BRIP1/BACH1/FANCJ encodes for a DNA-dependent ATPase and a 5’ to 3’ helicase (S. Cantor et al., 2004), necessary for the normal functioning of the double-strand break repair mechanism of BRCA1 by interacting with the BRCT domain (S. B. Cantor et al., 2001). PALB2 interacts with and promotes the localisation of BRCA2 by co-localisation and allows the repair and checkpoint regulation (B. Xia et al., 2006). The protein products of the *NBN* gene are named p53/Nibrin/NBS1, and form a part of the MRN complex along with MRE11 and RAD50, which mediates double-stranded break sensing functions (Petrini & Stracker, 2003).

1.4 Carrier testing

High-risk individuals with a family history of breast and/or ovarian cancer may be referred for a genetic screening test to determine their predisposition to hereditary cancer. Many variations of genetic screening services are available for high-risk individuals, whereby the type of laboratory test utilised is dependent on each clinic. Initial clinical genetic screening for mutations within *BRCA1* were monopolised by Myriad Genetics, which made these tests unaffordable and inaccessible to many high-risk family members. However, a court ruling in 2010 revoked the patent, which allowed other companies to competitively decrease the prices of these genetic screening tests (Kesselheim & Mello, 2010).

Carrier testing for hereditary breast and ovarian cancer genes by multiplex gene panels, include, but are not restricted to *BRCA1* and *BRCA2* (Beaudet, 2015; Crawford et al., 2017; Grindedal et al., 2017; Howarth et al., 2015). These panels range from small panels with high penetrance genes, to larger panels with a broader scope of genes (Shiovitz & Korde, 2015). Next-generation sequencing - a relatively rapid, massively parallel sequencing method (Behjati & Tarpey, 2013) - was the most favoured variant detection method in non-American laboratories, whereas next generation and Sanger sequencing were equally favoured in American laboratories. Furthermore, Multiplex Ligation-dependent Probe Amplification - a multiplex PCR assay (Schouten et al., 2002) - was the most favoured approach for deletion/duplication detection in non-American laboratories (Toland et al., 2018).

1.5 The Variants of Uncertain Significance (VUS) problem

Common obstacles to screening processes are VUS; genetic variants with an unclarified clinical significance (Millot et al., 2012). The approximate proportion of VUS of BRCA1/2 genetic testing outcomes are 5-10 % (Lattimore et al., 2015), and rare VUS which impact on alternative splicing are difficult to identify without patient tissue. Individuals carrying a VUS in a susceptibility gene are unlikely to receive prophylactic treatment based on the genetic test alone, as these risk-reducing measures may involve invasive surgical procedures (Eggington et al., 2014). However, family and medical history are implemented to allow prevention schemes to be applied to these individuals (Weitzel et al., 2003). The outcome of a VUS was reported to generate higher psychological stress (O'Neill et al., 2009), although this concept is conflicting with other studies (S. van Dijk et al., 2004). Overall, constructing more evidence for classifying these VUS is paramount. Unfortunately, while many international collaborations are taking place to further classify these VUS, more novel VUS are being continually identified (Duzkale et al., 2013). Evidence shows that in 2012, over 500 missense VUS were noted for *BRCA1* alone, and this number is hypothesised to increase over the years (Millot et al., 2012; National Human Genome Research Institute). Furthermore, approximately 100 New Zealanders in the past decade have been notified with the discovery of VUS during a screening test (Lattimore et al., 2015). Regardless of these findings, genetic screening should be offered on a wider scale, as the benefits of detecting pathogenic variants outweigh the expanding problem of VUS amassment (Duzkale et al., 2013; Grindedal et al., 2017). However, a contrasting position is taken by another study which showed that in fact, the VUS rate declined from 12.8 % to 2.1 % of all results (Eggington et al., 2014). These results may be attributed to the increased usage of tests and the advent of improved screening methods (Lattimore et al., 2015).

1.6 Classifying variants of uncertain significance

According to existing surveys, detecting rare VUS which impact on alternative splicing and providing evidence to classify these VUS, are laborious processes (Cheon, Mozersky, & Cook-Deegan, 2014) (**Figure 1.2**). A few approaches have been proposed for classifying VUS. These include using *in silico* approaches such as multifactorial probability-based models (Lindor et al., 2012), and biological approaches such as RNA transcript analysis (Thomassen et al., 2012) and embryonic stem cell-based functional assays (Chang, Biswas, Martin, Stauffer, & Sharan, 2009). Broader methods may include finding the same variants in family members (Cheon et al., 2014) and using animal models (Hakem et al., 1996). Furthermore, increasing the accuracy of current approaches has been beneficial for classification.

In particular, RNA transcript analysis has been indicated by the Evidence-based Network for the Interpretation of Germline-Mutant Alleles (ENIGMA) consortium to be a substantial method of providing evidence for classifying a subset of VUS. The consortium has provided evidence that capillary electrophoresis, Polymerase Chain Reaction (PCR) and sequencing are valid semi-quantitative methods for transcript splicing analysis (Colombo et al., 2014; Whiley et al., 2014). Historically, PCR methods have been used to analyse aberrant splicing by evaluating variants located near or within intron-exon boundaries (Chen et al., 2006; de Jong et al., 2017). Further studies have utilised mini-gene splicing assays for characterising the impact of splicing alterations (Ahlborn et al., 2015). Quantitative PCR methods have also been utilised as a quantitative method of analysing splice isoforms for VUS classification (Vandenbroucke, Vandesompele, Paepe, & Messiaen, 2001; Walker et al., 2010). However, they are unable to provide qualitative information about specific transcript sequences. Recently, CloneSeq, a novel high-throughput massively parallel RNA sequencing assay which is able to provide qualitative and quantitative data, was developed (Farber-Katz et al., 2018). Overall, all approaches of visualising alternative splicing effects utilise cDNA as the molecule of interest instead of RNA, which may be a considerable limitation for these approaches.

The abovementioned methods may aid in classification of VUS, however, the scope of classification levels is variable between different clinics. For instance, one study uses a 5-tier system composed of 'deleterious', 'suspected deleterious', 'VUS', 'genetic variant favour polymorphism' and 'polymorphism' levels (Eggington et al., 2014). In contrast, another implemented a 4-tier categorisation method composed of 'non-mutation', 'pathogenic variant', 'polymorphism/neutral variant' and 'VUS' levels (Millot et al., 2012). The wide variety of categorisation systems reveals that there is no standardisation for variant classification.

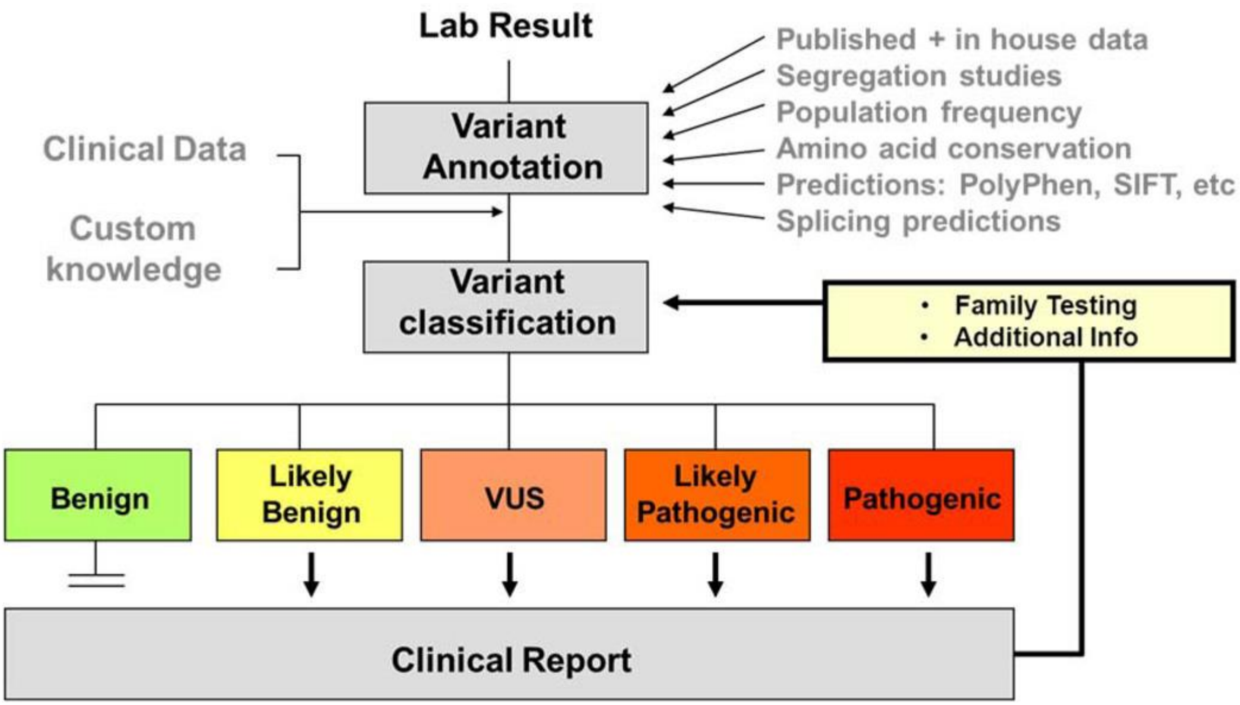


Figure 1. 2 Classification workflow of variant classification.

A laboratory result reveals the variant carried by the tested individual. The variant is annotated with external data, and classified into 5 categories; benign, likely benign, VUS, likely pathogenic, and pathogenic. Subsequently, a clinical report is produced. Reproduced with permission (Duzkale et al., 2013).

1.7 Splicing

A subset of VUS which impact on splicing may be classified by visualising the effect of gene variants on alternative splicing patterns with RNA transcript analyses. Splicing is the process whereby a section of a pre-mRNA transcript is removed, and the remaining sequences are joined together (**Figure 1.3**). This principle explains the non-uniformity between the number of human genes and proteins, which is a widely accepted concept. Hence, the concept of 'one gene–one RNA-one protein' is no longer recognised, as splicing has been found to affect 95% of human multi-exon genes (Pan, Shai, Lee, Frey, & Blencowe, 2008).

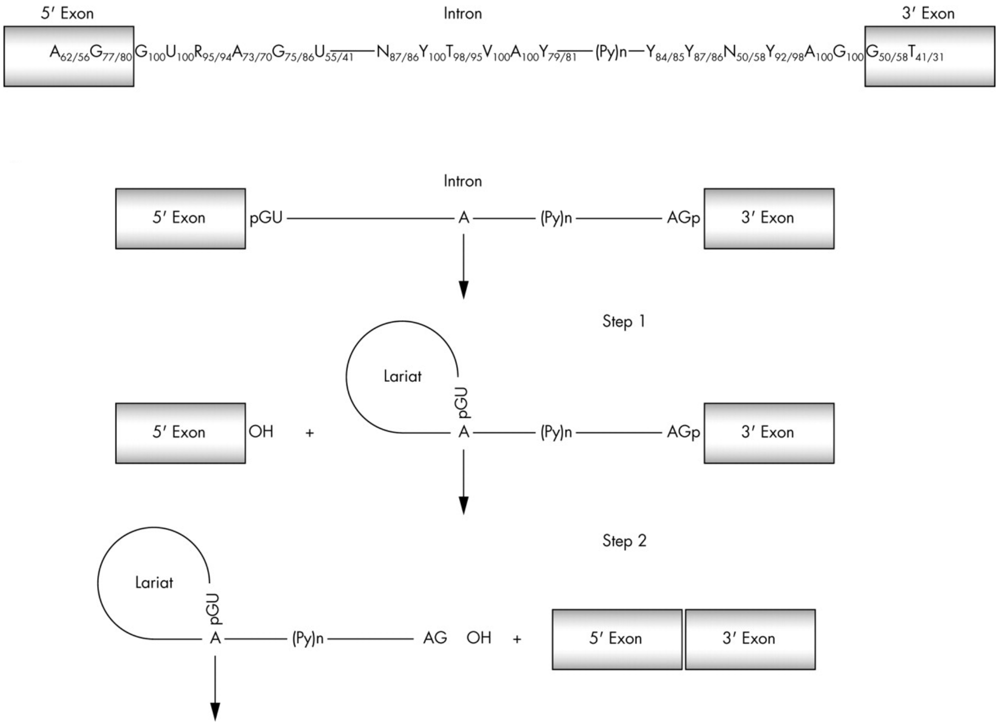

5′ Exon
Intron
3′ Exon
$A_{62/56}G_{77/80}G_{100}U_{100}R_{95/94}A_{73/70}G_{75/86}U_{55/41}$—$N_{87/86}Y_{100}T_{98/95}V_{100}A_{100}Y_{79/81}$—(Py)n—$Y_{84/85}Y_{87/86}N_{50/58}Y_{92/98}A_{100}G_{100}G_{50/58}T_{41/31}$
Intron
5′ Exon
pGU
A
(Py)n
AGp
3′ Exon
Step 1
Lariat
5′ Exon
OH
+
pGU
A
(Py)n
AGp
3′ Exon
Step 2
Lariat
pGU
A
(Py)n
AG
OH
+
5′ Exon
3′ Exon

Figure 1. 3 A brief schematic of a splicing event.

The 5’ and 3’ exons are spliced together by intron removal through the formation of a lariat structure.

Reproduced with permission (Baralle & Baralle, 2005).

Two forms of splicing exist; constitutive and alternative splicing. Constitutive splicing is the exon-joining and intron-removal in the order of most genes, whereas alternative splicing facilitates exon-skipping to produce different mature mRNA isoforms (Maki et al., 1981). The interplay between alternative transcription start sites, transcription end sites, alternative splice sites, exon-skipping and intron-retention all contribute to the final form of the mRNA transcript.

Splicing is regulated by trans-acting factors and cis-acting elements, and their mechanisms are well documented in the literature (Berget, Moore, & Sharp, 1977; Breitbart & Nadal-Ginard, 1987; Delsert, Morin, & Klessig, 1989; Fu & Manley, 1987). The cis-acting elements include Intronic Splicing Enhancers (ISE), Intronic Splicing Silencers (ISS), Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS). The enhancers are bound by positive trans-acting factors, such as the Serine/Arginine-rich family of nuclear phosphoproteins (SR proteins) (Ge & Manley, 1990) (**Figure 1.4**), and the silencers are bound by negative trans-acting factors such as heterogeneous nuclear ribonucleoproteins (hnRNPs). The splicing itself is carried out by a protein-RNA complex called the spliceosome by targeting the 5' and 3' regions. The spliceosome is formed by more than 170 proteins and 5 small nuclear RNAs (snRNAs) (Behzadnia et al., 2007). The recognition of the 5' GU and the 3' AG of introns leads to the out-splicing of certain introns to form the mature mRNA. This mechanism commences with the U1 small nuclear riboprotein (snRNP) binding to the 5' splice site (Siliciano & Guthrie, 1988), and the Splicing Factor 1 (SF1) binding to the branch point (Berglund, Chua, Abovich, Reed, & Rosbash, 1997). This leads to the U2 snRNP auxiliary factor (U2AF) binding to the 3' splice site to facilitate the binding of U2 to the branch point (Ruskin, Zamore, & Green, 1988). The commitment to splicing occurs through the substitution of SF1 by U2 snRNP, which forms the pre-spliceosomal A complex (Kramer & Utans, 1991). The A complex converts to the B (Wolf et al., 2009) and C

(Zhan, Yan, Zhang, Lei, & Shi, 2018) complexes which facilitates the removal of the intron and ligation of the exons through a transesterification reaction.

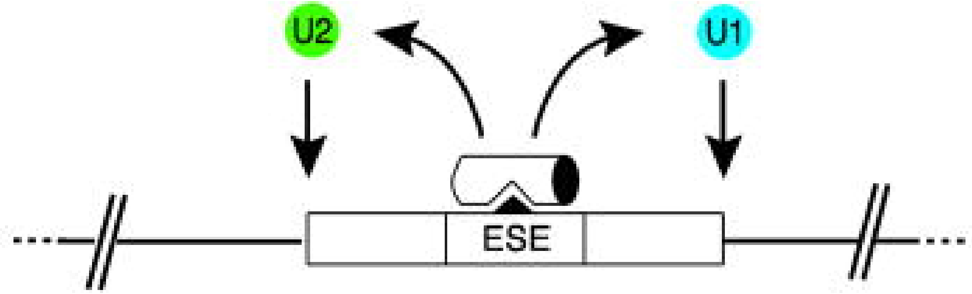


Figure 1. 4 Brief schematic of spliceosomal action during splicing.

Factors bind to regulatory regions in order to ensure that the exon which is to be spliced by the spliceosome can be recognised. The figure represents two SR-proteins binding to an Exonic Splicing Enhancer, to facilitate the stable binding of U1 and U2 snRNPs. Reproduced with permission (Kelemen et al., 2013).

1.8 Alternative splicing

Alternative splicing appears to have a profound effect in contributing to proteome diversity (Gueroussov et al., 2015) and isoform function by affecting most cellular processes. Firstly, isoforms are involved in changing global cellular mechanisms. An example is the process of apoptosis, whereby alternative splicing acts as a switch for pro-apoptotic or anti-apoptotic enzymes (Vegran, Boidot, Solary, & Lizard-Nacol, 2011). This mechanism may be cell-type dependent (Caldas, Fangusaro, Boue, Holloway, & Altura, 2007). Alternative splicing also affects transcription factors by affecting the protein-complex forming system. For instance, the skipping of an exon with DNA binding ability will produce transcription factors which have lost the ability to bind to promoters. Moreover, the regulation of such mechanisms is also affected by alternative splicing (Goodson, Jonas, & Privalsky, 2005). Additionally, the localisation of proteins is affected by splicing. Certain variants of the enzyme Drosha were localised to the cytoplasm while two others were localised to the nucleus (Link, Grund, & Diederichs, 2016). Additionally, the Divalent Metal Transporter 1 (DMT1) exhibits clear intracellular localisation patterns which are isoform-dependent (Tabuchi, Tanaka, Nishida-Kitayama, Ohno, & Kishi, 2002), which further supports this concept. Furthermore, the splicing mechanism can inactivate and modulate enzymes in a variety of ways. For example, the cytosolic phospholipase A2 beta changes the substrate specificity due to alternative splicing (Ghosh, Tucker, Burchett, & Leslie, 2006). Other functions of alternatively spliced isoforms include involvement in forming protein domains (Kelemen et al., 2013), channel proteins (Vallejo-Illarramendi, Domercq, & Matute, 2005) and binding properties between proteins (Belfiore, Frasca, Pandini, Sciacca, & Vigneri, 2009). Moreover, an interesting study suggests that certain isoforms may only be functional as a result of other related isoforms being expressed by a change in splicing (Kelemen et al., 2013).

Aberrant alternative splicing is known to be strongly associated with diseases, which is expected given the involvement of alternative splicing in protein function and regulation. The relevance to cancer diagnosis, management and therapeutics is important to acknowledge, as numerous studies have shown that aberrant alternative splicing contributes to cancer formation (S. Li, Hu, Zhao, Huang, & He, 2018; Z. Li et al., 2015; Miyamoto et al., 2018). An example of a well-noted alternative splicing isoform of *BRCA1* is the Δ11/11q isoform, which involves skipping of the whole or part of the largest exon in the *BRCA1* gene, respectively (Tammaro, Raponi, Wilson, & Baralle, 2014). Exon 11 contains regulatory sequences which are important in alternative splicing regulation (Raponi, Douglas, Tammaro, Wilson, & Baralle, 2012) and the skipping of this exon is believed to promote growth and survival of tumour cells (Tammaro, Raponi, Wilson, & Baralle, 2012). Thus, new therapeutic approaches may be found by recognising the relevance of these aberrant alternative splice isoforms (S. C. Lee & Abdel-Wahab, 2016). This may be achieved using multiple bioinformatic prediction programmes (Spurdle et al., 2008).

Nonsense-Mediated Decay (NMD) is a method of splicing regulation by degradation of mRNA isoforms with premature stop codons 50 bps upstream of the exon junction complex (Nagy & Maquat, 1998). However, other functions of NMD have been found, which include regulation of physiological transcript expression (Mendell, Sharifi, Meyers, Martinez-Murillo, & Dietz, 2004). NMD inhibitors such as cycloheximide allow retention and visualisation of transcripts that would otherwise be degraded by the NMD process.

Additionally, while exon-skipping is the most common type of alternative splicing in humans, other types of alternative splicing such as NAGNAG splicing exist. This involves tandem acceptors separated by three nucleotides and results in two splice variants; the E isoform and the I isoform (Sinha et al., 2009). The most common alternative splicing events have been presented in **Fig. 1.5**, and the nomenclature for alternative splicing events are described below (**Table 1.2**).

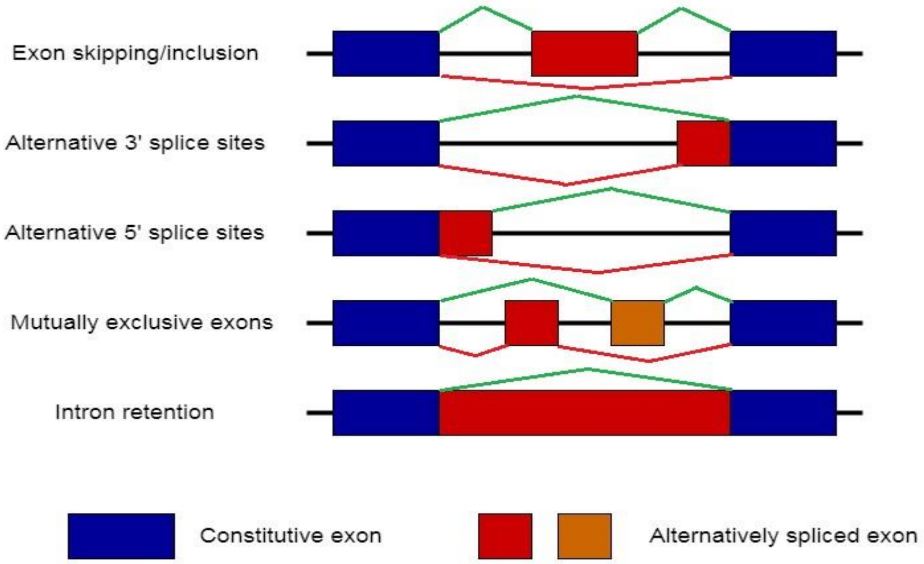

Exon skipping/inclusion
Alternative 3' splice sites
Alternative 5' splice sites
Mutually exclusive exons
Intron retention
Constitutive exon
Alternatively spliced exon

Figure 1. 5 A diagram of possible splicing events.

Five main mechanisms of alternative splicing exist; exon inclusion or skipping, alternative 3’ splice-site selection, alternative 5’ splice-site selection, mutually exclusive exons and intron retention. Redrawn and reproduced with permission (Cartegni, Chew, & Krainer, 2002).

Table 1. 2 Description of nomenclature used in splicing events

Nomenclature	**Description**
Δ **x**	Exon-skipping of exon x
▼**x**	Intronic retention in intron x

* Exon-skipping is represented by the symbol Δ, and intron retention is represented by the symbol ▼. The numeric value which follows the symbol refers to the exon or intron that is affected.

1.9 Molecular methods for analysing mRNA transcripts

1.9.1 Polymerase Chain Reaction (PCR)

PCR is a non-linear amplification method which enables amplification of DNA. Three major steps are involved; denaturation, annealing and elongation. The reaction incorporates a DNA template, DNA polymerase, forward and reverse 3' primers, deoxynucleoside triphosphates (dNTPs), buffer solution, bivalent cations and monovalent cations (Mullis & Faloona, 1987). The method is useful for visualising target samples with small quantities (Erlich, Gelfand, & Sninsky, 1991). Reverse-transcription PCR is a variation of the PCR concept and works by reverse-transcribing RNA samples into cDNA.

1.9.2 Quantitative PCR (qPCR)

This method provides a quantitative assessment of the nucleic acids in the sample of interest. The approach provides real-time quantification, where the relative abundance of PCR products can be visualised. The approach works by utilising a fluorescent molecule to detect and quantify the nucleic acids of interest. There are two main methods of carrying out qPCR; use of intercalating fluorescent dyes which bind non-specifically to double-stranded DNA (Ponchel et al., 2003), and the use of a sequence-specific probe with fluorescent molecules which are activated by the hydrolytic cleavage by an enzyme (Holland, Abramson, Watson, & Gelfand, 1991).

1.9.3 *In vitro* transcription

In vitro transcription is a method of linear amplification of target RNA, which is commonly used in situations with a low quantity of starting sample (J. Schneider et al., 2004). The most common use of this approach involves a T7 polymerase bacteriophage promoter attached to an oligo-dT primer upstream of the sequence to be targeted by the T7 RNA polymerase (Tabor & Richardson, 1985). This primes the cDNA strand generated by reverse-transcribing mRNA. This amplification is highly reproducible, and the amplified RNA correlates to the total RNA (J. Schneider et al., 2004). Furthermore, the average correlation coefficient between *in vitro* amplified and non-amplified samples was shown to be 0.82 (Zhao, Hastie, Whitfield, Borresen-Dale, & Jeffrey, 2002). Therefore, this approach may be beneficial for visualising poorly expressed genes such as *BRCA1*.

1.10 Sequencing and genome-wide techniques for studying alternative splicing

1.10.1 Microarrays

Traditionally, hybridisation microarrays were used to evaluate gene fusion (Lovf et al., 2013), mutations/single nucleotide polymorphisms (Park et al., 2004), gene expression (Dhanasekaran et al., 2001) and alternatively spliced transcripts (Pan et al., 2004). However, constraints such as poor quantification of poorly and highly expressed genes and cross-hybridisation artefacts accelerated the development of alternative approaches (Casneuf, Van de Peer, & Huber, 2007; Eklund et al., 2006; Okoniewski & Miller, 2006).

1.10.2 First & second-generation sequencing

Sanger or first-generation sequencing with Expressed Sequence Tags (EST) was widely used before the transition into next-generation sequencing or second-generation sequencing (Dias Neto et al., 2000). Although Sanger sequencing was regarded as the gold standard, it has a relatively low throughput. With the invention of next-generation sequencing, Illumina HiSeq has been dominating this field as the new gold standard sequencer with its low cost, high accuracy and high throughput properties (Bentley et al., 2008). Furthermore, the Flow Cell Surface Reverse Transcription sequencing (FRT-seq) method was developed as an alternative method for transcriptome sequencing with the Illumina Genome Analyser. This method uses poly(A)+ RNA template as opposed to cDNA template, which maintains strand specificity and reduces PCR bias (Mamanova et al., 2010). These methods have been used to define single and multiple exon-skipping events which occur within a few hundred bases (Colombo et al., 2014; Davy et al., 2017; Lattimore et al., 2018). However, these methods are suboptimal for examining exon connectivity in long genes such as *BRCA1*.

1.10.3 Third Generation Sequencing

The arrival of third generation sequencers with single-molecule detection initiated a new era of sequencing, which includes PacBio and nanopore sequencing. Coupled with the ability of real-time sequencing of single molecules (E. L. van Dijk, Auger, Jaszczyszyn, & Thermes, 2014) and long-read sequencing, these two methods are powerful (Hoenen et al., 2016; Rhoads & Au, 2015). Additionally, the ability of nanopore sequencing to carry out DRS - sequencing of RNA as opposed to cDNA offers new potential approaches to understanding RNA splicing (Byrne et al., 2017).

1.10.3.1 Pacific Biosciences (PacBio) sequencing

PacBio delivers a fast, long-read, single molecule, third generation sequencing which is able to provide data in real-time (Au et al., 2013; Eid et al., 2009; Sharon, Tilgner, Grubert, & Snyder, 2013). The long reads are suitable for de novo transcriptome assembly (Brown et al., 2014), and assessing alternative splicing This is optimal for analysing and detecting alternative splice isoforms due to a higher percentage of complete sequence reads. Hence, the PacBio Iso-Seq method has been used to study exon splicing mechanisms and bears a pipeline designed specifically for this application (Tabor & Richardson, 1985). However, the inability of carrying out DRS and its high error rates (5%) contributes to insertions and deletions and result in the loss and misalignment of reads (Carneiro et al., 2012).

1.10.3.2 Nanopore sequencing

Comparatively, nanopore sequencing is a novel technology, derived from the ideas of two researchers David Deamer and George Church. The ONT MinION (**Figure 1.6 A**) is a single-molecule sequencer similar to the PacBio with the ability to determine exon connectivity in a single transcript (Bolisetty, Rajadinakaran, & Graveley, 2015). This is possible even with *Dscam1*, the most complicated alternatively spliced gene known (Bolisetty et al., 2015). The flow-cell of a MinION (**Figure 1.6 A**) consists of a membrane with continuous electrical current, in which approximately 2048 nanopores are embedded (Jain, Olsen, Paten, & Akeson, 2016). Double-stranded nucleic acids pass one of its strands into the nanopores, disrupting the current within the membrane (Kasianowicz, Brandin, Branton, & Deamer, 1996) (**Figure 1.6 B**). The disruption patterns can be attributed to specific bases, and the output data can be analysed further via bioinformatic approaches. Unfortunately, this sequencing method also has a high error rate (Jain et al., 2016), however, the rolling circle amplification appears to allow improvements in the quality (Eid et al., 2009). The approach works by using a circular template to generate a long single-stranded DNA or RNA, which can be manipulated to suit the purpose of the experiment (Ali et al., 2014).

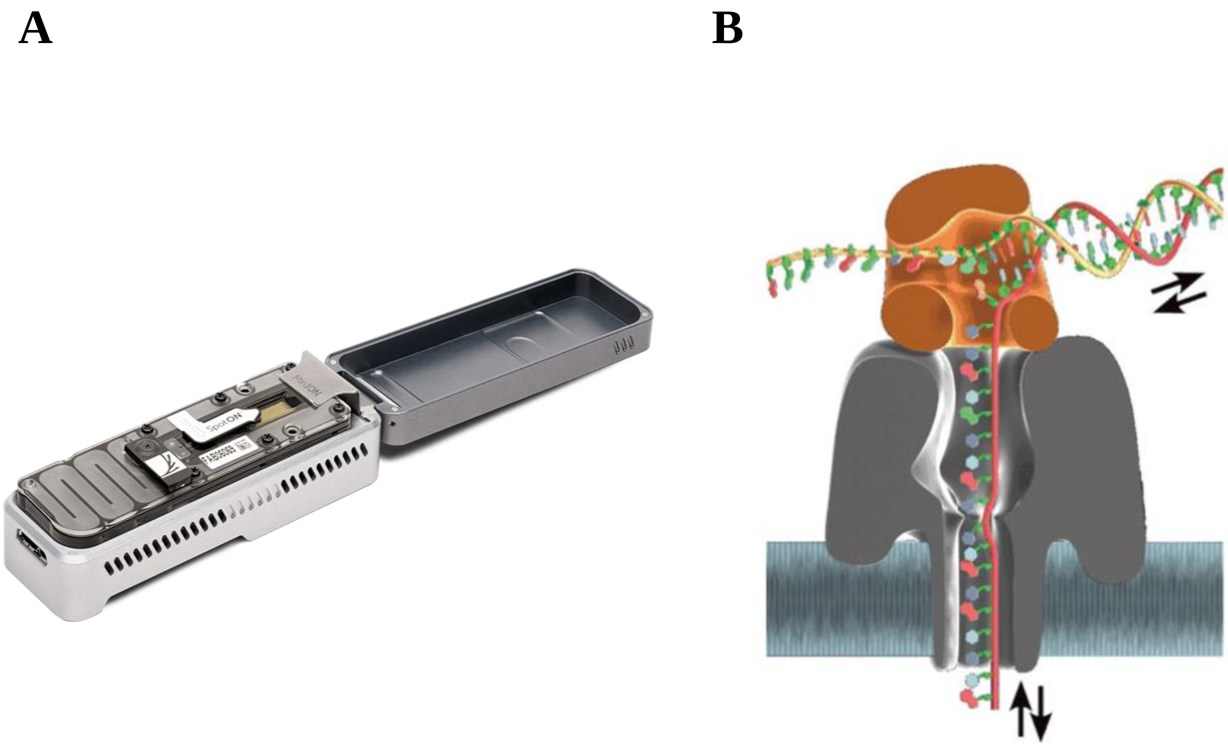


Figure 1. 6 Parts involved in nanopore sequencing.

A) A diagram of a MinION nanopore sequencer with a flow-cell attached. Each flow-cell carries approximately 2048 nanopores. Reproduced from Oxford Nanopore Technologies (https://nanoporetech.com/products/minion). **B)** Figure of a nanopore passing a molecule to be sequenced. One strand of the double-stranded molecule bears an anchor to stabilise the structure, in order to allow the other strand to penetrate through the nanopore. This allows the nucleic acid to be sequenced. Reproduced with permission (G. F. Schneider & Dekker, 2012).

1.10.3.3 Advantages of nanopore sequencing

It is now clear that extra benefits are associated with nanopore sequencing compared to PacBio other than their mutual long-read sequencing ability. These added advantages include low capital cost (Bolisetty et al., 2015), portability, detection of cytosine methylation (Rand et al., 2016; Simpson et al., 2017) and DRS ability (Garalde et al., 2018). This may provide more complete sequence and modification annotations. Furthermore, the quality of nanopore reads does not decline with length, which is advantageous (McCabe, Cormican, Johnston, & Earley, 2018). Additionally, the MinION was able to sequence a DNA fragment over 2 million bases, which confirms the strength of the sequencer (Payne A, 2018).

Whether in the context of educational institutions or diagnostic clinics, the low capital cost of the MinION is advantageous for the end-user. For example, the high capital cost of PacBio sequencing deters institutions and clinics from implementing its long-read ability (Bolisetty et al., 2015). In contrast, the MinION has the lowest costs among all single molecule sequencers (Rhoads & Au, 2015).

Portability allows the MinION to be utilised in in-field diagnostics, unlike other sequencers (Jain et al., 2016). The MinION was used during the Ebola and Zika outbreaks and thus its usefulness has been well validated (Quick et al., 2017; Quick et al., 2016). Similarly, it was implemented in rapid bovine pathogen diagnostics, increasing the feasibility of this technology to be applied in other research fields (McCabe et al., 2018).

Overall, the high consensus among studies suggests that the biggest drawback of using nanopore sequencing is the high error rate. Nanopore sequencing has difficulties with high G + C content sequences (Laver et al., 2015), homopolymer tracts (Loman, Quick, & Simpson, 2015), and low base-call accuracy (Faucon et al., 2017; McCabe et al., 2018). To illustrate the severity of this issue, 1D (template strand) error rates of > 20 % and 2D (template and complement strands) error rates of 10-15 % were noted in 2016 (Ip et al., 2015). Nevertheless, improvements in accuracy and throughput have occurred over the past few years and are likely to continue (Bolisetty et al., 2015). The use of Unique Molecular Identifiers may improve the variation in transcript quantification (Islam et al., 2014). Contrastingly, another disregards the feasibility of this approach as the length of the Unique Molecular Identifiers is required to be > 30 bp with the current state of errors (Byrne et al., 2017). It is suggested that base-calling algorithms need to be improved to further address this problem and software need to be optimised for nanopore sequencing (Garalde et al., 2018). Furthermore, as the motor-DNA complex determines the read length and accuracy, searching for more sophisticated motor proteins with longer processivity in the future may also be beneficial (Y. Wang et al., 2015).

1.11 RNA-seq vs Direct RNA Sequencing (DRS)

It is evident that there had been an ongoing effort to deliver the best version of RNA-Seq technology. The perpetual drive for creating superior RNA-Seq technology derives from the fact that this technology is useful and may be applied to various research studies (Dillies et al., 2013). The term 'RNA-Seq' has been well incorporated into scientific literature, however, the term has become a misnomer with the advancement of sequencing technology. While countless older studies have used 'RNA-Seq' as a term for direct sequencing of cDNA, the term was incorrectly accepted as an accurate term because DRS was not possible. However, the advent of DRS has made the use of this term problematic, where recent studies have had to implement the term 'direct RNA-Seq' to distinguish from ‘RNA-Seq’. However, certain studies use these two terms interchangeably, which defeats the purpose of distinguishing these two terms (Q. Xu et al., 2017). Thus, it is important that these terms are used appropriately.

RNA-Seq reads cDNA molecules which has been reverse-transcribed from an RNA sample, and the generated reads are aligned to a reference genome for further analysis (Soneson & Delorenzi, 2013). This method has been utilised for de novo assembly of transcripts (Cao et al., 2015), studying the transcriptome (Nagalakshmi et al., 2008), gene expression studies (Zhuo, Emerson, Chang, & Di, 2016) and detecting alternative splicing isoforms (Davy et al., 2017). Bioinformatic analysis is efficient with RNA-seq as more tools have been developed for RNA-seq compared to DRS. However, limitations include error propagation (Roberts et al., 1989), artificial splicing (Cocquet, Chong, Zhang, & Veitia, 2006), duplication errors (Salzberg et al., 2017) and loss of strandedness information (Haddad, Qin, Giger, Guo, & Baldwin, 2007). Long transcripts are also over-represented compared to shorter transcripts, which poses a statistical bias (Oshlack & Wakefield, 2009). Problems with analysing complex genomic regions, isoform detection and identifying methylations also exist specifically with short-read RNA-seq (Byrne et al., 2017; Rhoads & Au, 2015). The uneven coverage along the genome is problematic as the differences in the sequencing depths are troublesome (Soneson & Delorenzi, 2013). However, methods to mitigate the loss of strandedness are available (Chu & Corey, 2012), and alternative splicing analyses, computational approaches and repetitive sequencing can compensate the bias effect (Oshlack & Wakefield, 2009), but at the cost of increasing the cost per base for a given accuracy rate.

In contrast, the main advantage of DRS is its ability to sequence RNA without the need for converting to cDNA prior to sequencing. Microarrays and RNA-Seq techniques required this cDNA conversion step as RNA was not able to be sequenced directly (Aird et al., 2011). The elimination of the reverse transcription, ligation and amplification steps is advantageous as it minimises the distortion RNA template representation (Ozsolak & Milos, 2011a). DRS only requires 3′ polyadenylated templates to allow short and long RNAs to be sequenced together in a single experiment (Ozsolak & Milos, 2011b). These advantages simplify the overall workflow and reduce reagent costs. Long-read DRS is necessary to gain enough coverage at 5' and 3' ends to identify full-length transcripts, which cannot be reciprocated in conventional RNA-Seq (Levin et al., 2010). This method will also allow detection of nucleotide analogues, reduction of PCR bias, and strand specificity (Salzberg et al., 2017). Finally, fusion events and RNA modifications can also be visualised with this method (Garalde et al., 2018). One limitation is lower base-calling accuracy with DRS compared to RNA-Seq (Garalde et al., 2018). Nonetheless, as with improvements in RNA-Seq, rapid optimisation is expected to occur with growing interest in DRS methods.

1.12 Analysis

The basic flow of downstream sequencing analyses is similar between all next-generation sequencing and third-generation sequencers. Sequencing with next-generation sequencing and single molecule platforms create FASTQ-format files for the generated reads. These reads are rid of sequencing artefacts and errors, aligned to reference genomes, annotated and quantified (Chu & Corey, 2012). Short-read sequencers with limited ability to recognise exon connectivity in mRNAs have been coupled with software programmes to account for this constraint (Bolisetty et al., 2015). However, examining exon co-occurrence in *BRCA1* transcripts has been shown to overcome this problem with the use of the MinION (de Jong et al., 2017). Bioinformatic tools are available from ONT, and many are also developed by the community which are freely accessible via online repositories (Chu & Corey, 2012). As nanopore technology is a relatively novel concept, the topic of optimum bioinformatic software for downstream analyses is well debated. Overall, as it is difficult to determine the best progression of bioinformatic software associated with the MinION, tools should be carefully selected depending on the application.

All RNA-Seq and DRS approaches display advantages and disadvantages, and this suggests that it is best to select sequence platforms based on the downstream application (Soneson & Delorenzi, 2013). If higher accuracy is a priority over the ability to examine alternative splicing, traditional methods should be used over single-molecule based platforms (Chu & Corey, 2012).

However, nanopore sequencing appears to be the best overall approach to study VUS effects on alternative splicing patterns of RNA transcripts due to its long-read, high throughput, and DRS capabilities.

1.13 Aims and Hypotheses

The primary hypothesis of this project was that the MinION nanopore sequencer can directly read and detect full-length RNA isoforms of *BRCA1*. A secondary hypothesis was that a novel enrichment method is able to linearly amplify RNA isoforms in a quantitative manner by *in vitro* transcription. The aims of this project are to develop and test a novel approach for linear enrichment of RNA transcripts from a control lymphoblastoid cell line (LCL) and to quantify the different isoforms, which will be valuable in understanding any splicing alterations caused by VUS. Nanopore sequencing is a relatively new technique and testing its parameters and using this approach to understand patterns of VUS is the key goal of this research project.

Chapter 2: Materials and Methods

Table 2. 1 List of products and reagents used in experiments

Name of Product/Reagent	**Manufacturer/Source**	**Location**
Tris Base	Thermo Fisher Scientific	Massachusetts, United States
EDTA	Ajax Finechem by Thermo Fisher Scientific	Auckland, New Zealand
Glacial Acetic Acid	Thermo Fisher Scientific	Massachusetts, United States
Lymphoblastoid Cell Line (07.111.0085)	kConFab	Melbourne, Australia
Cycloheximide Solution	Sigma-Aldrich	Missouri, United States
Fetal Bovine Serum	Hyclone™ (GE Life Sciences)	Auckland, New Zealand
Penicillin/Streptavidin	Gibco® by Life Technologies™	Auckland, New Zealand
RPMI Medium 1640 1X (+ L-Glutamine)	Gibco® by Life Technologies™	Auckland, New Zealand
RNeasy® Mini Kit	Qiagen	Hilden, Germany
polyA-Spin™ mRNA Isolation Kit	New England BioLabs® Inc.	Massachusetts, United States
NanoDrop™ 8000 Spectrophotometer	Thermo Fisher Scientific	Auckland, New Zealand
Qubit™ RNA HS Assay Kit	Invitrogen by Thermo Fisher Scientific	Oregon, United States
Qubit™ 1X dsDNA HS Assay Kit	Invitrogen by Thermo Fisher Scientific	Oregon, United States

Tapestation RNA ScreenTape	Agilent Technologies	California, United States
Primers	Integrated DNA Technologies	Queenstown, Singapore
RNAseZap™	Invitrogen by Thermo Fisher Scientific	Vilnius, Lithuania
Superscript® III Reverse Transcriptase	Invitrogen by Thermo Fisher Scientific	California, United States
Ambion™ RNAseH, *E.coli*	Invitrogen by Thermo Fisher Scientific	Vilnius, Lithuania
RNAseA/T1 Mix	Thermo Scientific	Massachusetts, United States
Betaine	Thermo Fisher Scientific	Massachusetts, United States
KAPA 5X Long Range Buffer	KAPA Biosystems	Massachusetts, United States
25 mM $MgCl_2$	KAPA Biosystems	Massachusetts, United States
10 mM dNTP mix	KAPA Biosystems	Massachusetts, United States
KAPA Long Range Hotstart (2.5 U/µL)	KAPA Biosystems	Massachusetts, United States
Fisher TAQ-Ti DNA Polymerase (5 U/ µL)	Fisher Biotec	Wembley, Australia
Fisher TAQ-Ti DNA Polymerase 10X Reaction Buffer	Fisher Biotec	Wembley Australia
Nuclease-Free Water - UltraPure™ Distilled Water	Gibco® Invitrogen	Auckland, New Zealand
HyAgarose™	HydraGene	New Jersey, United States
SYBR® Safe DNA Gel Stain	Invitrogen by Thermo Fisher Scientific	California, United States

Electrophoresis System – PowerPac Basic™ (300 V/400 mA/75 W)	BIO-RAD	California, United States
KAPA Universal DNA Ladder	KAPA Biosystems	Massachusetts, United States
6X KAPA Loading Dye	KAPA Biosystems	Massachusetts, United States
Alliance 4.7 Imaging System	UVItec	Cambridge, United Kingdom
LongAmp® *Taq* PCR Kit	New England BioLabs® Inc.	Massachusetts, United States
HighPrep™ PCR Magnetic Beads	MAGBIO Genomics Inc.	Maryland, United States
MEGAscript™ T7 Transcription Kit	Invitrogen by Thermo Fisher Scientific	Vilnius, Lithuania
RNAClean™ XP	Beckman Coulter®	California, USA
MinION Flow-Cell – FLO-MIN106 R9.4.1 Version	Oxford Nanopore Technologies	Oxford, England
Direct RNA-seq Kit -SQK-RNA001	Oxford Nanopore Technologies	Oxford, England
1.7 mL Clear Microtubes	Axygen®	California, United States
RNAseOUT™	Invitrogen by Life Technologies	California, United States
1.5 mL DNA LoBind Tubes	Eppendorf™	Hamburg, Germany

2.1 General stocks

2.1.1 Tris/Acetic acid/EDTA (TAE) buffer

A solution of 10 X TAE buffer was made with 48.4 g of Tris Base, 3.7 g of EDTA, and 11.4 mL of glacial acetic acid. The solution was made up to 1 L with MilliPore Water (MPW). The buffer was diluted to 1x for further use.

2.1.2 Tris/EDTA (TE) buffer

TE buffer was formulated by adding 1 mL of 1 M Tris Base, and 200 µL of 0.5 M EDTA, to a final volume of 100 mL with MPW. The final concentrations of the solution comprised of 10 mM of Tris Base, and 1 mM EDTA.

2.2 RNA source/cell culturing

A control LCL (07.111.0085) was grown in 37 °C and 5 % CO_2 incubation, suspended in 20 µL of media. Media used to culture the cells consisted of 55 mL of Fetal Bovine Serum, 5.5 mL of Penicillin/Streptavidin and 500 mL of RPMI 1640. Cells were fed and split into new passages every 1-3 days.

Shortly before RNA extraction, the cells were treated with cycloheximide solution (NMD inhibitor) for 4 hours. The cultured cells were used to extract total RNA via the RNeasy® Mini Kit according to the protocol described in **Appendix A**. Twenty-three RNeasy® Mini Kit extractions were performed in total. PolyA-Spin™ mRNA Isolation Kit was used to isolate the mRNA according to the protocol described in **Appendix B**. One PolyA-Spin™ mRNA Isolation Kit extraction was performed in total. NanoDrop™ 8000 Spectrophotometer, Qubit™ RNA HS Assay Kit and the Agilent RNA ScreenTape/Tapestation were used to determine the quality and concentration of the extracted RNA.

2.3 Primer design

The software Primer 3 (version 0.4.0) and Geneious (version 11.1.4) were used for designing primers (**Table 2.2**).

Table 2. 2 List of primers designed for PCR reactions, with characteristics

Name	Primer sequence	Tm (°C)	Description
T7-*BRCA2*_F1	GGTAATACGACTCACTATAGGGAGAAGCGTGAGGGGACAGATTTG	66.2	T7 polymerase binding site attached to a *BRCA2* exon 1 forward primer
T7-*PALB2*_F1	GGTAATACGACTCACTATAGGGAGACCATCTTTTTGTCCAGCCAGC	66.4	T7 polymerase binding site attached to a *PALB2* exon 1 forward primer
T7-*RAD51B*_F1	GGTAATACGACTCACTATAGGGAGAAAGTTGGATGCTGCAGACCC	66.4	T7 polymerase binding site attached to a *RAD51B* exon 1 forward primer
T7-*RAD51C*_F1	GGTAATACGACTCACTATAGGGAGAAAATGCAGCGGGATTTGGTG	66.0	T7 polymerase binding site attached to a *RAD51C* exon 1 forward primer
T7-*RAD51D*_F1	GGTAATACGACTCACTATAGGGAGATCCATTCCCTGGTGTCTTCG	65.9	T7 polymerase binding site attached to a *RAD51D* exon 1 forward primer
T7-*BRIP1*_F1	GGTAATACGACTCACTATAGGGAGAAAGGGTGTGGTGATTGCCAA	66.5	T7 polymerase binding site attached to a *BRIP1* exon 1 forward primer
T7-*BARD1*_F1	GGTAATACGACTCACTATAGGGAGAATTCACCCAGCAAGGTCAGG	66.3	T7 polymerase binding site attached to a *BARD1* exon 1 forward primer
T7-*ATM*_F1	GGTAATACGACTCACTATAGGGAGACGCGCTTACCCAATACAAGC	66.4	T7 polymerase binding site attached to an *ATM* exon 1 forward primer
T7-*CHEK2*_F1	GGTAATACGACTCACTATAGGGAGAACATCAGTGACTGTGAAAAAGCA	65.0	T7 polymerase binding site attached to a *CHEK2* exon 1 forward primer

T7-*FANCM*_F1	GGTAATACGACTCACTATAGGGAGAGGTGGTTGTCGGCCTAATGA	66.7	T7 polymerase binding site attached to a *FANCM* exon 1 forward primer
T7-*NBN*_F1	GGTAATACGACTCACTATAGGGAGAAGGGATTGTGGGCATGACAG	66.3	T7 polymerase binding site attached to a *NBN* exon 1 forward primer
T7-*GAPDH*_F1	GGTAATACGACTCACTATAGGGAGATCTCTGCTCCTCCTGTTCGA	66.2	T7 polymerase binding site attached to a *GAPDH* exon 1 forward primer
T7-*BRCA1*_F1	GGTAATACGACTCACTATAGGGAGAGCGCGGGAATTACAGATAAA	53.0	T7 polymerase binding site attached to a *BRCA1* exon 1 forward primer
T7_F	GGTAATACGACTCACTATAGGGAGA	54.6	T7 polymerase binding site forward primer
Anchored Oligo-dT$_{(20)}$	TTTTTTTTTTTTTTTTTTTTVN	41.0	Oligo-dT20 primer with a TN anchor
***GAPDH*_R9**	GGGGAGATTCAGTGTGGTGG	57.7	*GAPDH* exon 9 reverse primer
***BRCA1*_1F**	GCGCGGGAATTACAGATAAA	68.1	*BRCA1* exon 1 forward primer
***BRCA1*_24pR**	AAGCTCATTCTTGGGGTCCT	56.2	*BRCA1* exon 24 reverse primer on the p side
***BRCA1*_7R**	CCAATTCAATGTAGACAGACG	51.2	*BRCA1* exon 7 reverse primer

2.4 First cDNA strand synthesis

Preparation of cDNA was executed in a pre-PCR room free of post-PCR contaminants. RNAseZAP™ was used to clean workbenches and pipettes to minimise RNAse contamination. RNA extracted from the control LCL was reverse-transcribed using the enzyme Superscript® III Reverse Transcriptase according to the protocol described in **Appendix C**. An aliquot of this cDNA sample was diluted 1:5 with NFW for use in quality control tests.

An anchored oligo-dT$_{(20)}$ was utilised instead of a standard oligo-dT$_{(20)}$ primer as it carries a TN anchor, which allows it to bind exactly to the end of the 5' of the poly-A tail of mRNA (Nam et al., 2002). This minimizes the binding to internal intronic poly-A sites, which was likely to reduce spurious transcription events.

2.4.1 Quality control for first cDNA strand synthesis

2.4.1.1 Standard long-range PCR

All standard long-range PCR reactions were made to a total reaction volume of 20 µL. Each reaction contained 1.3 M betaine, 1x KAPA long range buffer, 1.75 mM of $MgCl_2$, 0.5 µM of each forward and reverse primer (*BRCA1*_1F and *BRCA1*_24pR respectively), 300 µM of KAPA 10mM dNTP mix, 0.5 units of KAPA Long Range HotStart and 1.2 µL of 1:5 diluted cDNA. Thermal cycling conditions consisted of an initial denaturation step of 94 °C for 4 minutes, then 35 cycles of 94 °C denaturation step for 30 seconds, a 59 °C annealing step for 30 seconds, and an extension step at 68 °C for 12 minutes, with a final extension of 72 °C for 12 minutes.

2.4.1.2 Alternative long-range PCR

The alternative long-range PCR was carried out to carry out an improved method of long-range PCR. This PCR consisted of a higher $MgCl_2$ and cDNA concentration, lower dNTP concentration and lower denaturation and extension times compared to the standard long-range PCR. All alternative long-range PCR reactions were made to a total reaction volume of 10 µL. Final reactions contained 1 M betaine, 1x KAPA long range buffer, 2 mM of $MgCl_2$, 0.7 µM of each forward and reverse primer (BRCA1_1F and BRCA1_24pR respectively), 200 µM of KAPA 10 mM dNTP mix, and 0.5 units of KAPA Long Range HotStart, and 3 µL of 1:5 diluted cDNA. Thermal cycling conditions consisted of an initial denaturation step of 94 °C for 2 minutes, then 35 cycles of 94 °C denaturation step for 30 seconds, a 59 °C annealing step 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 Second cDNA strand synthesis

The LongAmp® *Taq* DNA Polymerase kit was used to generate the second cDNA strand. A Gene Specific Primer (GSP) attached to a T7-polymerase binding site was used to generate the second strand of cDNA with this kit, according to the modified protocol (**Appendix D**). The modifications include the number of PCR cycles and using an optimum annealing temperature. The manufacturer recommends an annealing temperature ranging from 45 to 65 °C (https://international.neb.com/protocols/2012/10/15/m0323-longamp-taq-dna-polymerase-protocol). Therefore, a temperature gradient PCR was carried out to determine the optimum annealing temperature for the T7-GSP primer for *BRCA1* and *GAPDH*.

2.5.1. Temperature gradient PCR

Temperature gradient PCR was carried out to find the optimal annealing temperature for second strand cDNA synthesis. All temperature gradient reactions consisted of 12 reactions of 10 µL. T7-*BRCA1* exon 1 forward and exon 7 reverse primers were used for *BRCA1*, and T7-*GAPDH* exon 1 forward and exon 9 reverse primers were used for *GAPDH*. Each reaction volume comprised of 1x TAQ buffer, 0.1 mM dNTP mix, 1.5 mM $MgCl_2$, 0.5 units Fisher TAQ-Ti DNA polymerase, 0.5 µM forward primer, 0.5 µM reverse primer, 1 mM betaine, and 1 µL of 1:5 dilution cDNA, with the rest of the reaction volume made up to 10 µL with NFW. The cycling conditions were 1 cycle of 94 °C for 2 minutes, 45 cycles of 94 °C for 15 seconds, 50-65 °C gradient across 12 lanes for 15 seconds, 72 °C for 2 minutes, 1 cycle of 94 °C for 5 minutes and a final step of 1 cycle of 25 °C for 2 minutes.

2.5.2 Quality control for second cDNA strand synthesis

2.5.2.1 T7-only primer PCR

Products generated from cDNA second strand synthesis in experiment 4 were tested using the T7-only primer and a respective reverse primer to visualise the incorporation of the T7-GSP. The T7 only primer and *BRCA1* exon 7 reverse primer were used for *BRCA1*, and T7 only primer and *GAPDH* exon 9 reverse primer were used for testing *GAPDH*. All reactions were made up to a total volume of 20 µL. Each reaction volume consisted of 1x Fisher buffer, 0.2 mM dNTP mix, 1.5 mM $MgCl_2$, 0.5 units Fisher TAQ-Ti DNA polymerase, 2 µL of dscDNA, 0.5 µM T7-only forward primer, 0.5 µM reverse primer, 1 M betaine, and the rest of the reaction volume made up to 20 µL with NFW. The cycling conditions were 1 cycle of 94 °C for 2 minutes, 45 cycles of 94

°C for 15 seconds, 50 °C for 15 seconds, 72 °C for 2 minutes, 1 cycle of 94 °C for 5 minutes and 1 cycle of 25 °C for 2 minutes.

2.5.2.2 T7-only and oligo-dT$_{(20)}$ primer (full-length) PCR

Products of the cDNA second strand synthesis were tested using the T7-only and an oligo-dT$_{(20)}$ primer in experiment 4. All reactions were made up to a total volume of 20 µL. Each reaction volume consisted of 1x Fisher buffer, 0.2 mM dNTP mix, 1.5 mM $MgCl_2$, 0.5 units Fisher TAQ-Ti DNA polymerase, 2 µL of dscDNA, 0.5 µM T7-only forward primer, 0.5 µM oligo-dT$_{(20)}$ reverse primer, 1 M betaine, and the rest of the reaction volume made up to 20 µL with NFW. The cycling conditions were 1 cycle of 94 °C for 2 minutes, 45 cycles of 94 °C for 15 seconds, 50 °C for 15 seconds, 72 °C for 2 minutes, 1 cycle of 94 °C for 5 minutes and 1 cycle of 25 °C for 2 minutes.

2.6 Agarose gel electrophoresis

Agarose gels (1 % in TAE) were used for visualising all PCR products. Agarose powder was dissolved in 1x TAE buffer and 1 µL of SYBR® Safe DNA Gel stain was added for each 25 mL of 1x TAE buffer used. This gel was set in a PowerPac Basic™ electrophoresis system, which was submerged in 1x TAE buffer. 4 µL of KAPA Universal DNA ladder was loaded into the first lane. 1x volume of 6x KAPA loading dye was added to each PCR product and loaded on the gel. Electrophoresis was carried out with the electrophoresis system at 90 V for 30-60 minutes. The gel was examined via the image system Alliance 4.7.

2.7 HighPrep™ magnetic bead clean-up

HighPrep™ magnetic beads were used to isolate dscDNA from the products of the cDNA second-strand synthesis according to the protocol described in **Appendix E**.

2.8 *In vitro* transcription

The MEGAscript™ T7 Transcription Kit was used to generate RNA via *in vitro* transcription according to the protocol described in **Appendix F**. The extension was run for 4 hours in experiment 2, 16 hours in experiment 3 and 20 hours in experiment 4. The quality of the *in vitro* transcribed RNA was checked via the Tapestation assay.

2.9 RNAClean™ XP magnetic bead clean-up

The RNAClean™ XP beads were used to isolate the RNA products from the *in vitro* transcription step in experiment 4, according to the protocol described in **Appendix G**. In experiment 4, 20 µL of *RAD51C, BRCA1, CHEK2* in vitro products and 1 µL of *GAPDH* in vitro product were pooled and cleaned with the RNAClean™ XP beads, and eluted with 15 µL of NFW before the MinION library preparation step.

2.10 MinION library preparation and sequencing

The protocol for library preparation was as described by Oxford Nanopore Technologies (protocol version: DRS_9026_v1_revN_15Dec2016), which is in **Appendix H**. The kit control was omitted in the reaction and was replaced with the same volume with *in vitro* transcribed RNA, which was isolated by the RNAClean™ XP beads and reverse-transcribed as a part of the protocol, and then adapters were ligated to the mRNA (**Figure 2.1**). The same Direct RNA Sequencing Kit was used in all experiments (SQK-RNA001), and all flow-cells were checked for quality when first received and shortly before the sequencing had been carried out.

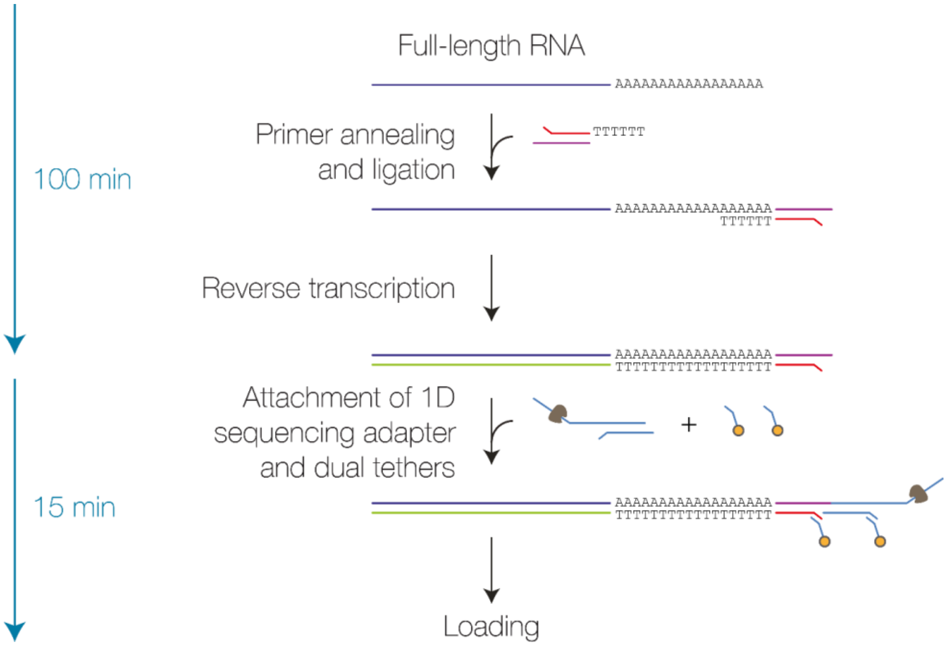


Figure 2. 1 A brief schematic of library preparation with the ONT Direct RNA Sequencing Kit.

The steps involved are; primer annealing and ligation, reverse transcription, attachment of 1D sequencing adapters and dual tethers, and loading this resulting sample on a flow-cell. These steps ensure that the mRNA molecule is able to be sequenced by the nanopores. Reproduced from https://store.nanoporetech.com/direct-rna-sequencing-kit.html.

2.11 MinION sequencing analysis

2.12.1 Basic data acquisition

MinKNOW is the default software developed by ONT for the purposes of carrying out data acquisition, data streaming, real-time analysis and feedback, and sample identification and tracking (https://nanoporetech.com/analyse). MinKNOW version 1.13.1 and 1.14.1 (**Table 2.3**) were used to derive FAST5 files from the ONT MinION sequencing reads from experiment 1 and experiments 2-4 respectively. The concatenated FAST5 files were basecalled via Albacore version 2.3.0, a basecalling software which identifies DNA/RNA sequence information from the raw data from the MinKNOW programme (https://nanoporetech.com/about-us/news/new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy). The basecalling was carried out with the command 'nice read_fast5_basecaller.py --input [pathway to fast5 files] --worker_threads 4 --save_path [pathway and name of output files] --flowcell FLO-MIN106 --kit SQK-RNA001 --output fastq --recursive --files_per_batch_folder 0'. This created FASTQ output files. The reads were then concatenated with the script 'cat *fastq > filename.fastq'. This created concatenated FASTQ files. Total counts of reads were calculated via the command 'awk '{s++} END {print s/4}' path to fastq file'.

Table 2. 3 List of programmes used for basic data acquisition

Programme	**Source**
MinKNOW v1.13.1 & v1.14.1	https://community.nanoporetech.com/protocols/experiment-companion-configuration/v/cte_1000_v1_revak_17nov2015/installing-minknow-for-con-50942?search_term=MinKNOW
Albacore v2.3.0	https://community.nanoporetech.com/protocols/albacore-offline-basecalli/v/abec_2003_v1_revan_29nov2016/linux

2.12.2 Alignment

Table 2. 4 List of reference genomes, transcriptomes and datasets used in bioinformatic analyses

Reference Genomes/ Transcriptomes/Datasets	**Source**	**Description**
Human Genome hg38	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/	GRCh38.12 Human Genome Reference Build hg38
Compiled Human Genome hg38	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz	Genome compiled by Dr. Heng Li (Broad Institute) - more suitable for the purposes of using Minimap2 than the original Human Genome Reference Build hg38
Human Transcriptome	https://www.gencodegenes.org/releases/current.html	Gencode v28 (GRCh38.12)
Reference_panel_genes.fasta	https://www.ncbi.nlm.nih.gov/	File of FASTA sequences from all genes studied in this thesis
Gencode v28 annotation GFF3 file	https://www.gencodegenes.org/releases/current.html	Gencode v28 (GRCh38.12) annotation GFF3 file
BRCARD1_geneseq.fasta	Reference taken from the storage file of Gene Structure and Function Laboratory (GSFL), University of Otago, Christchurch	File of joined FASTA sequences of *BRCA1* and *BARD1*. Prepared by Lucy de Jong (GSFL).
Nanopore RNA Consortium data	https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md	DRS data generated by an international Nanopore RNA Consortium with RNA isolated from a lymphoblastoid cell line (NA12878)
Nanopore RNA Consortium subset data	https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md	A subset of DRS data generated by an international Nanopore RNA Consortium with RNA isolated from a lymphoblastoid cell line (NA12878). This consists of pass reads from one MinION run (Notts_Run1).
Makenzie Cancer Research Group (MCRG) RNA-seq data	**Appendix I**	File of a subset of RNA-seq data generated in the Mackenzie Cancer Research Group (MCRG) with a control LCL (08.111.0064)
Davy et al. (2017) data	Supplementary table 1 from Davy et al. (2017)	Table of *BRCA1*, *RAD51C*, *CHEK2* alternative splicing events observed in RNA-seq data from Davy et al. (2017)

2.12.2.1 Alignment with Minimap2

Minimap2 (**Table 2.5**) is a pairwise aligner for genomic and spliced nucleotide sequences, which has specific parameters for nanopore DRS reads (H. Li, 2018b). The aligner Minimap2 was used to index various reference genomes (**Table 2.4**) with commands 'minimap2 -d ref.mmi [reference file].fasta'. These indices were used to align our concatenated FASTQ files mainly with the commands 'minimap2 -ax splice -uf -k14 [pathway to index file] [input files].fasta > [output file].sam'. SAM files were created through this alignment, and different parameters were also tested during this step. From these SAM files, BAM files were created with the commands 'samtools view –S –b [pathway to file].sam > [pathway and name of output file].bam', sorted with the commands 'samtools sort [pathway to file].bam –o [pathway and name of input file].sorted.bam', and indexed with the commands 'samtools index [pathway to sorted file].bam'. The sorted BAM files were visualized on the Integrated Genomics Viewer (IGV) version 2.4.

2.12.2.2 Alignment with BLAST-Like Alignment Tool (BLAT)

BLAT (**Table 2.5**) is a rapid alignment tool which is commonly used for mRNA/DNA and cross-species protein alignments (Kent, 2002). The concatenated FASTQ file was converted to FASTA files via the command 'seqtk seq -a in.fastq.gz > [output file].fasta'. The aligner BLAT was used to align our FASTA files to various reference genomes with a range of parameters. PSLX and BLAST8 files were created as a result. After these files were checked via Microsoft Office Excel 2013, the PSLX files were converted into SAM files with the psl2sam.pl script, which were then in turn converted to BAM files. The BAM files were sorted and indexed with the programme Samtools, and visualized on IGV version 2.4.

2.12.2.3 Alignment with Genomic Mapping and Alignment Programme (GMAP)

GMAP (**Table 2.5**) is generally used for mRNA and EST sequences (T. D. Wu & Watanabe, 2005). The concatenated RNA FASTQ files were converted to FASTA files via the command ‘seqtk seq –a in.fastq.gz > [output file].fasta’. The FASTA files were converted to cDNA via converting Uracils to Thymines by the GALAXY FASTA Manipulation RNA/DNA Converter. The cDNA FASTA file was then used for GMAP alignment.

2.12.2.4 Alignment with LAST

LAST (**Table 2.5**) was designed for aligning long RNA or DNA reads (M. C. Frith, Wan, & Horton, 2010). The LAST databases were created by the command ‘lastdb -P8 -uNEAR -R01 mydb [path to reference genome/transcriptome file].fasta’. This created FASTA files. The query files were aligned with the commands ‘parallel-fasta ‘lastal -p myseq.par -d90 -m50 -D10 mydb | last-split -m1 -d2 -g mydb" < [path to query file].fasta> myseq.maf’.

Table 2. 5 List of programmes used for alignment

Programme	**Source**
Minimap2 - v2.11-r819-dirty	https://github.com/lh3/minimap2#install
BLAT - v36 x 2	http://www.blat.net/examples/install&profile.html
GMAP - v2016-09-23	http://research-pub.gene.com/gmap/
LAST - v941	https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md
Samtools - v1.7	http://www.htslib.org/download/
GALAXY FASTA manipulation RNA/DNA converter	https://usegalaxy.org/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fdevteam%2Ffasta_nucleotide_changer%2Fcshl_fasta_nucleotides_changer%2F1.0.2&version=1.0.2&__identifer=ixz50yagcqh/
Psl2sam.pl	https://github.com/samtools/samtools/blob/master/misc/psl2sam.pl

2.12.3 Quantification

2.12.3.1 Isoform Counts with Kallisto

Kallisto (**Table 2.6**) is a programme for quantifying reads generated from high throughput RNA-seq experiments (N. L. Bray, Pimentel, Melsted, & Pachter, 2016). Kallisto index files were created with a reference transcriptome (**Table 2.4**) via the commands ‘kallisto index –i [pathway to output file] [reference transcriptome].fasta’. The index was then used to quantify the different transcripts produced by the MinION run, using the concatenated FASTQ files. The command used was 'kallisto quant –i [pathway to index file] -o [pathway to output directory] --single –l [mean read length] -s [standard deviation of reads] [pathway to query file].fastq --bias –t 4’. This produced a TSV file including the name, length, effective length, estimated counts and Transcripts Per Million (TPM), which were visualised via Microsoft Office Excel 2013. The mean read length and standard deviation of reads were generated by the StatsFasta.py script (**Table 2.6**).

2.12.3.2 Isoform Counts with Salmon

Salmon (**Table 2.6**) is a programme for rapid quantification of transcripts from RNA-seq experiments (R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, & C. Kingsford, 2017). Various parameters of Salmon were used to test its ability to quantify ONT DRS reads. Quasi-non-aligned-based mode was tested by creating an index with the commands 'salmon index –t [path to reference transcriptome].fa -i transcripts_index --type quasi –k 31'. The index was used to quantify with the commands 'salmon quant -i [pathway to index file] -l A --fldMean [mean read length] --fldSD [standard deviation of reads] -r [pathway to query file].fastq -o [name of output file]'. The mean read length and standard deviation of reads were generated by the StatsFasta.py script (**Table 2.6**). Non-quasi alignment-based mode was tested with the commands 'salmon quant –t [pathway to reference transcriptome].fa -l A –a [pathway to minimap2-aligned bam/sorted.bam file].bam -o [name of output file]'. The two methods produced an SF file as a result including names, length, effective length, Transcripts Per Million (TPM) and number of reads, which was visualised via Microsoft Office Excel 2013.

2.12.3.3 Isoform Counts with Mandalorion – Episode II

Mandalorion – Episode II (**Table 2.6**) is a pipeline which determines how many reads overlap with the exons of a gene to produce a reads/gene per 10,000 reads (R. Volden et al., 2018; R. Volden & Vollmers, 2018). The commands ‘python3 defineAndQuantifyWrapper.py -c [path to content_file] -p [path to query file] -u 5 -d 30 -s 200 -r 0.05 -R 3 -i 0 -t 0 -I 100 -T 60 -g [path to gencode v28 annotation GTF file] -m NUC.4.4.mat -f [patht to example_config file]’ were used to align the output of the MinION DRS parameters, which were the exemplar experiments given by the Mandalorion creators.

2.12.3.4 Isoform Counts with Structural and Quality Annotation of Novel Transcript Isoforms (SQANTI)

SQANTI (**Table 2.6**) is a pipeline which was created specifically to characterise PacBio cDNA long reads (de la Fuente & Tardaguila, 2018). Various parameters were tested with SQANTI, with alignment to either the human genome or the human transcriptome. SQANTI utilised GMAP databases, which were already created from the alignment steps in chapter 2.12.2.3. The general command used was ‘sqanti_qc.py -x [path to GMAP database] -o [name of output file] -d [path to output directory] [query file].fasta [annotation file].gtf [reference file].fasta’.

2.12.3.5 Isoform counts with manual quantification

BRCA1, *CHEK2*, *RAD51C* and *GAPDH* transcripts including full-length and non-full-length isoforms were counted manually from the sorted BAM file being visualised on IGV. The counts were organised into a table of counts. These counts were compared to counts generated from external datasets which include, DRS data from the Nanopore RNA Consortium, RNA-seq data from Mackenzie Cancer Research Group and RNA-seq data from a study by Davy et al. (2017) (**Table 2.4**). Initially, a subset of nanopore pass reads (Consortium Subset Data) from the Nanopore RNA Consortium were aligned and manually quantified due to the interest of time and lack of satisfactory analysis software. Subsequently, the full dataset was aligned with Minimap2 and quantified manually (including *RAD51C*, *BRCA1* and *CHEK2*). The *GAPDH* reads from the full dataset were not counted due to an interest of time, hence the subset data was used for downstream analysis for *GAPDH* transcripts.

Table 2. 6 List of programmes used for quantification

Programme	**Source**
Kallisto-v0.44.0	http://pachterlab.github.io/kallisto/download
Salmon-v0.11.2	https://github.com/COMBINE-lab/salmon/releases
Mandalorion – Episode II	https://github.com/rvolden/Mandalorion-Episode-II
SQANTI-v1.2	https://bitbucket.org/ConesaLab/sqanti/src
StatsFasta.py	https://cs.gmu.edu/~mlbio/LSH-DIV/

Chapter 3: Direct RNA sequencing of total Poly-A RNA (Experiment 1)

3.1 Introduction

For this project, four independent successive experiments (named Experiment 1-4) were carried out, each one involving development and analysis of an RNA library on a MinION flow-cell (**Figure 3.1**). Experiment 1 was carried out as the control experiment, by sequencing poly-A RNA isolated from a control LCL. In experiment 2, *GAPDH* was utilised as a control gene and was also enriched and sequenced along with *BRCA1*. In experiment 3, *BRCA1* and 11 other hereditary breast and ovarian cancer genes with a range of different transcript lengths and expression levels were utilised to test the enrichment with this novel enrichment technique. In experiment 4, *GAPDH*, *BRCA1*, *CHEK2* and *RAD51C* were used to test the enrichment technique. The purpose of the control experiment was to establish the DRS library preparation procedures in this laboratory, and to test the ability of the MinION to successfully carry out DRS. The data generated from this experiment was used as a control reference dataset to observe findings in subsequent experiments.

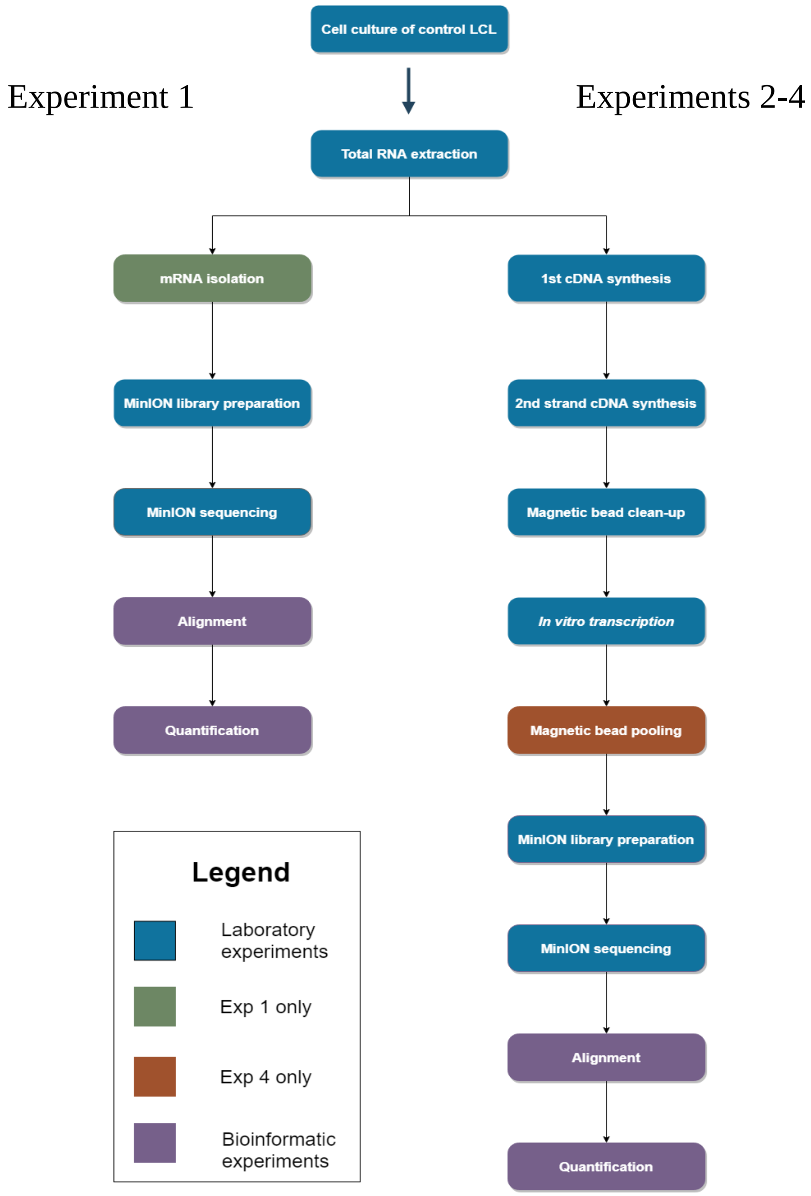


Figure 3. 1 A schematic diagram of the sequence of methods used in all experiments.

Experiment 1 utilised an mRNA isolation step, which was not employed in the other three experiments. Each experiment involved an RNAClean™ XP magnetic bead clean-up step as a part of the MinION library preparation. However, an extra magnetic pooling step with the RNAClean™ XP beads was carried out in experiment 4. All experiments consisted of cell culture, RNA extraction, MinION library preparation, MinION sequencing, alignment and quantification. Experiments 2-4 also included 1st and 2nd cDNA strand synthesis, clean-up with HighPrep™ magnetic beads and *in vitro* transcription.

3.2 RNA source/cell culturing

A control LCL was grown within RPMI 1640 media in suspension culture. After cycloheximide treatment, total RNA was extracted from these cells with the RNeasy® Mini Kit. The RNeasy® Mini Kit preparations were carried out 8 times for this experiment. The extracted total RNA presented with satisfactory quality from NanoDrop™, Qubit™ RNA HS and Tapestation results. Tapestation results were consistently above an RNA Integrity Number equivalent (RINe) score of 9.3. NanoDrop™ and Qubit™ results varied across the 8 RNA extraction events. Then, the RNA from 8 extractions were pooled to increase the concentration per given volume. The average NanoDrop™ concentration was 934.68 ng/µL and the average Qubit™ RNA HS result was 402.86 ng/µL. Poly-A RNA was isolated from the total RNA, using the polyA-Spin™ mRNA Isolation Kit. This step was carried out once, and Qubit™ RNA HS result from this mRNA isolation step revealed a concentration of 7.56 ng/µL and a total amount of 151.2 ng in a 20 µL volume.

3.3 MinION sequencing and bioinformatic analyses

The ONT protocol for the DRS kit was used according to the protocol described in **Appendix H**. The library preparation steps involved using the Direct RNA Sequencing Kit to prepare the mRNA for sequencing. This consisted of annealing and ligating the primers, reverse-transcription, attachment of 1D sequencing adapters and dual tethers (**Figure 2.1**). The prepared library was loaded on a MinION flow-cell, and sequencing was carried out for 48 hours. The total number of 889,850 sequenced reads was acquired by MinKNOW v1.13.1. Albacore v2.3.0 was used to basecall these reads and generate FASTQ files. This produced ‘pass’ and ‘fail’ folders, which contained 539,776 and 104,902 FASTQ reads respectively. All FASTQ files in the ‘pass’ folder were concatenated.

3.3.1 Alignment of MinION-generated reads

The concatenated FASTQ files from the data acquisition step were attempted to be aligned to various alignment programmes, by testing different parameters (**Tables 3.1-3.9**).

Using Minimap2, the parameters ‘k’ and ‘G’ were examined with the compiled Human Genome hg38 (**Table 2.4**). The ‘G’ parameter sets the maximum intron length, and the ‘k’ parameter sets the k-mer size (H. Li, 2018a). Each index was used to align FASTQ files to the compiled Human Genome hg38 with Minimap2, and the resulting SAM file was converted to BAM, sorted and indexed with Samtools, to be visualised on IGV. The results revealed no major differences between the successful indices, hence default parameters were used for subsequent experiments (-k14, -G200000) (**Table 3.1**).

Table 3. 1 Comparison of parameters ‘k’ and ‘G’ for creating Minimap2 indices by mapping to the compiled Human Genome hg38

Method number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Compiled Human Genome hg38	-k 1 (index) -G 50000 -d index file --secondary=no	Testing low k and low G	Unsuccessful - did not produce output
2	Compiled Human Genome hg38	-k 1 (index) -G 200000 -d index file --secondary=no	Testing low k and moderate G	Unsuccessful - did not produce output
3	Compiled Human Genome hg38	-k 1 (index) -G 500000 -d index file --secondary=no	Testing low k and high G	Unsuccessful - did not produce output
4	Compiled Human Genome hg38	-k 14 -G 50000 -d index file --secondary=no	Testing moderate k and low G	Successful
5	Compiled Human Genome hg38	-k 14 -G 200000 -d index file --secondary=no	Testing moderate k and moderate G	Successful - produces similar output to **4**
6	Compiled Human Genome hg38	-k 14 -G 500000 -d index file --secondary=no	Testing moderate k and high G	Successful - produces similar output to **4**
7	Compiled Human Genome hg38	-k 28 -G 50000 -d index file --secondary=no	Testing high k and low G	Successful - produces similar output to **4**
8	Compiled Human Genome hg38	-k 28 -G 200000 -d index file --secondary=no	Testing high k and moderate G	Successful - produces similar output to **4**
9	Compiled Human Genome hg38	-k 28 -G 500000 -d index file --secondary=no	Testing high k and high G	Successful - produces similar output to **4**

All Minimap2 indices were able to be successfully created using all the different types of references genomes. The index ‘k’ overrides the ‘k’ value set in the subsequent command for alignment, whereas this is not the case for the ‘G’ parameter. Therefore, only the ‘k’ parameter was set in indexing as opposed to using both ‘k’ and ‘G’ parameters (**Table 3.2**).

Table 3. 2 List of tested methods for creating Minimap2 indices

Index number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Compiled Human Genome hg38	-k14	Making an index with the compiled Human Genome Reference Build hg38 setting k-mer as 14	Successful
2	Human Genome hg38	-k14	Making an index with the Human Genome Reference Build hg38, setting k-mer as 14	Successful
3	Human Transcriptome	-k14	Making an index with the Human Transcriptome, setting k-mer as 14	Successful
4	Gencode v28 annotation GFF3 file	-k14	Making an index with the Gencode v28 aannotation file	Successful
5	Reference_genome_panel.fasta	-k14	Making an index with fasta sequences of genes of interest	Successful

Minimap2 was used to align FASTQ files to various indices created in **Table 3.2**. The parameter ‘ax splice’ allowed splicing information to be considered, and the parameter ‘uf’ sets the transcript strand to find the GT-AG sites. ‘--secondary=no’ allowed only primary alignments to be included in the output data (H. Li, 2018a). Method number 2 produced the best alignment according to the output results on IGV (**Table 3.3**).

Table 3. 3 List of tested parameters for aligning with Minimap2

Method number	**Index**	**Parameters**	**Description of action**	**Description of results**
1	**1**	-ax splice -uf	Aligning FASTQ files to the index made via the compiled Human Genome Reference Build hg38	Successful - can see mostly well-aligned reads on IGV
2	**2**	-ax splice -uf	Aligning FASTQ files to the index made via the compiled Human Genome Reference Build hg38	Successful - similar but superior mapping compared to **1**.
3	**3**	-ax splice -uf	Aligning FASTQ files to the index made via the Human Transcriptome	Successful - however, difficult to visualise on IGV with gene names alone. Need transcript ID information
4	**4**	-ax splice -uf	Aligning FASTQ files to the index made via the Gencode v28 annotation GFF3 file	Unsuccessful - Error seen on IGV
5	**5**	-ax splice -uf --secondary=no	Aligning FASTQ files to the index made via FASTA files of genes of interest	Unsuccessful - Error: segmentation fault (core dumped) when the concatenated pass file was tried to be aligned

FASTQ files were converted to FASTA files to be aligned with BLAT. The parameter ‘t’ is the database type, ‘q’ is the query type, ‘stepSize’ is the spacing between tiles, ‘repMatch’ is the number of repetitions allowed before it is marked as overused. ‘minScore’ sets the minimum score, ‘minIdentity’ sets the minimum sequence identity. The parameter ‘out’ sets the type of output file and ‘fine’ allows the search for small initial and terminal exons (Kent, 2002). All three trials resulted in unsuccessful outcomes, with non-interpretable outputs (**Table 3.4**).

Table 3. 4 List of tested parameters for aligning with BLAT

Method number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Human Genome hg38	-t=dna -q=rna -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -out=pslx -fine	Aligning FASTA files to the Human Genome hg38. Output file = PSLX	Unsuccessful - Introns and exons could not be differentiated on IGV
2	BRCARD1_ geneseq.fasta	-t=dna -q=rna -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -out=pslx -fine	Aligning FASTA files of to the BRCARD1_geneseq.fasta file. Output file = PSLX	Unsuccessful – Output on IGV could not be interpreted
3	BRCARD1_ geneseq.fasta	-t=dna -q=rna -stepSize=5 -repMatch=2253 -minScore=0 -minIdentity=0 -out=blast8 -fine	Aligning FASTA files of to the BRCARD1_geneseq.fasta file. Output file = BLAST8	Unsuccessful - Output on IGV could not be interpreted

Gmap_build was used to build the databases which were used as references for the alignment step (**Table 3.5**). The databases were used to align FASTA files converted from FASTQ files. The Uracil in these files were converted to Thymines using a GALAXY cDNA converter (**Table 2.5**) after the DRS FASTA files were unable to produce successful outcomes. The parameter ‘t’ sets the number of threads used, ‘f’ sets the output file type, ‘n’ sets the number of paths (T. Wu, 2018). All four trials of GMAP were unsuccessful, in which both the Human Genome Reference Build hg38 dataset and Human Transcriptome were used to align the reads (**Tables 3.5-3.6**).

Table 3. 5 Comparison of parameters creating GMAP databases

Database number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Human Transcriptome	gmap_build -d name of database -D path to reference file	Building a GMAP database with the Human Transcriptome	Successful
2	Human Genome hg38	gmap_build -d name of database -D path to reference file	Building a GMAP database with the Human Genome hg38	Successful

Table 3. 6 List of tested parameters for aligning with GMAP

Method number	Database	Parameters	Description of action	Description of results
1	**2**	-g path to reference file -t 4 -f samse	Aligning FASTA files to a database made with Human Genome hg38. Output = SAM file	Unsuccessful -Error: genomic sequence is unusually long. Error: No paths found Processed 17 queries
2	**1**	-g path to reference file -t 4 -f samse	Aligning FASTA files to a database made with Human Transcriptome. Output = SAM file	Unsuccessful -Error: No paths found Processed 17 queries
3	**1**	-g path to reference file -f 2 -n 0 -t 4	Aligning FASTA files to a database made with Human Transcriptome. Output = GFF3 file	Unsuccessful –Error: - gmap.avx2 does not exist -No paths found -Processed 173 reads from folder of over 200,000 reads
4	**2**	-D output directory -d gmap database -f 2 -n 0 -t 4	Aligning DRS FASTA files converted to cDNA to the gmap database created with the Human Transcriptome	Unsuccessful -Produced a GFF3 file with only two genes

LAST was used to create a single database using the Human Genome Reference Build hg38 as the reference. The parameters implemented in **Table 3.8** were used as they were pre-defined parameters set by the creators of LAST for long RNA sequencing. The parameter ‘P’ sets the number of processors, ‘uNEAR’ finds alignments with low rates of substitution, ‘d’ sets the minimum score for gapless alignments, ‘m’ sets the maximum initial matches per query position and ‘D’ sets query letters per random alignment (M. Frith, 2018). However, the default parameters for long RNA sequencing alignment produced unsuccessful results with the data from experiment 1 (**Tables 3.7-3.8**)

Table 3. 7 Comparison of parameters creating LAST databases

Database number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Human Genome hg38	lastdb -P8 -uNEAR Name of database Path to reference FASTA file	Making a database with default parameters for long nanopore reads	Successful

Table 3. 8 List of tested parameters for aligning with LAST

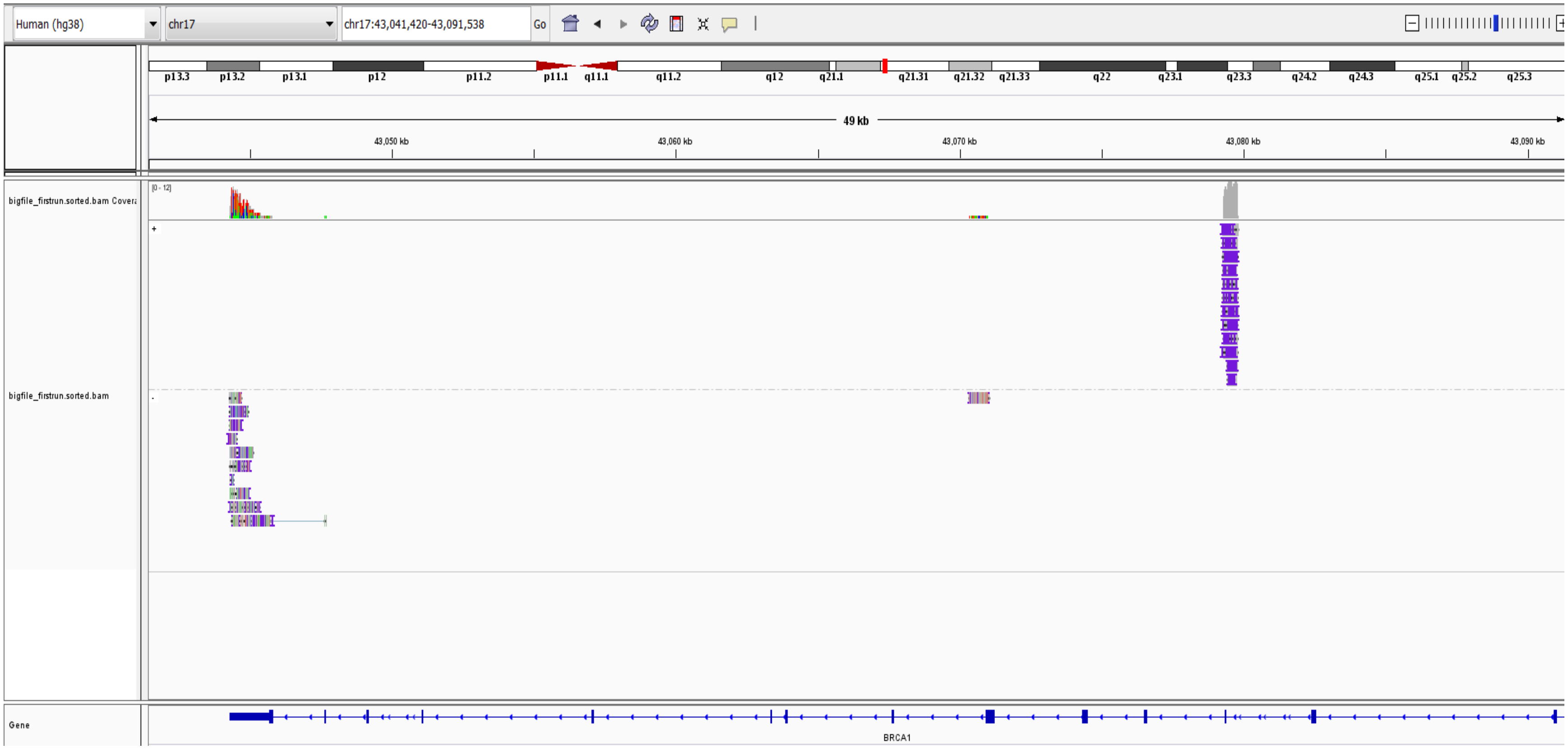
Method number	**Database**	**Parameters**	**Description of action**	**Description of results**
1	**1**	Parallel-fasta Lastal -p path to par file -d90 -m50 -D10 database | Last-split -m1 -d2 -g database	Aligning FASTA files to the database made with Human Genome hg38.	Unsuccessful - produced count text file with no content and a par file with lack of information

3.3.2 Summary of alignment

Minimap2 provided the best alignment and parameter combination among the four aligners tested, and therefore the rest of the analyses were carried out using Minimap2 with the suggested ONT DRS parameters. Minimap2 was used to align the reads to the Human Genome or the Human Transcriptome (**Table 2.4**). The results obtained from alignment were compared to an external dataset from the Nanopore RNA Consortium (**Table 2.4**).

3.3.3. Visual output of sequenced reads

The resulting BAM files from the alignment step were visualised on IGV as shown in **Fig. 3.2-3.9**. IGV screenshots of mRNA alignments to the Human Genome are shown in **Fig. 3.2, 3.4, 3.6 & 3.8.** IGV screenshots of mRNA alignments to the Human Transcriptome are shown in **Fig. 3.3, 3.5, 3.7 & 3.9.** Alignments of DRS reads for several genes of interest were illustrated, including the hereditary breast and ovarian cancer genes (*BRCA1*, *CHEK2* and *RAD51C*) and a housekeeping gene (*GAPDH*). The output of the sequenced reads revealed a 3' bias for most genes. High levels of control genes, such as *GAPDH* (**Figures 3.6-3.7**), and very low expression of genes of interest, such as *BRCA1* were observed (**Figures 3.2-3.3**). Some reads were ambiguous in alignment, with very short fragments aligned in the middle of the gene as opposed to the 5' or 3' ends. For example, *BRCA1* showed 12 reads spanning only exon 14, compared to only 11 reads being present at the 3' end (**Figure 3.2**). These ambiguous alignments were more common in genes with lower expression level and longer transcript length, such as *BRCA1*. Furthermore, full-length isoforms were more common in genes with high expression levels and shorter transcript lengths, such as *GAPDH* and *RAD51C* (**Table 3.6**).



Human (hg38)

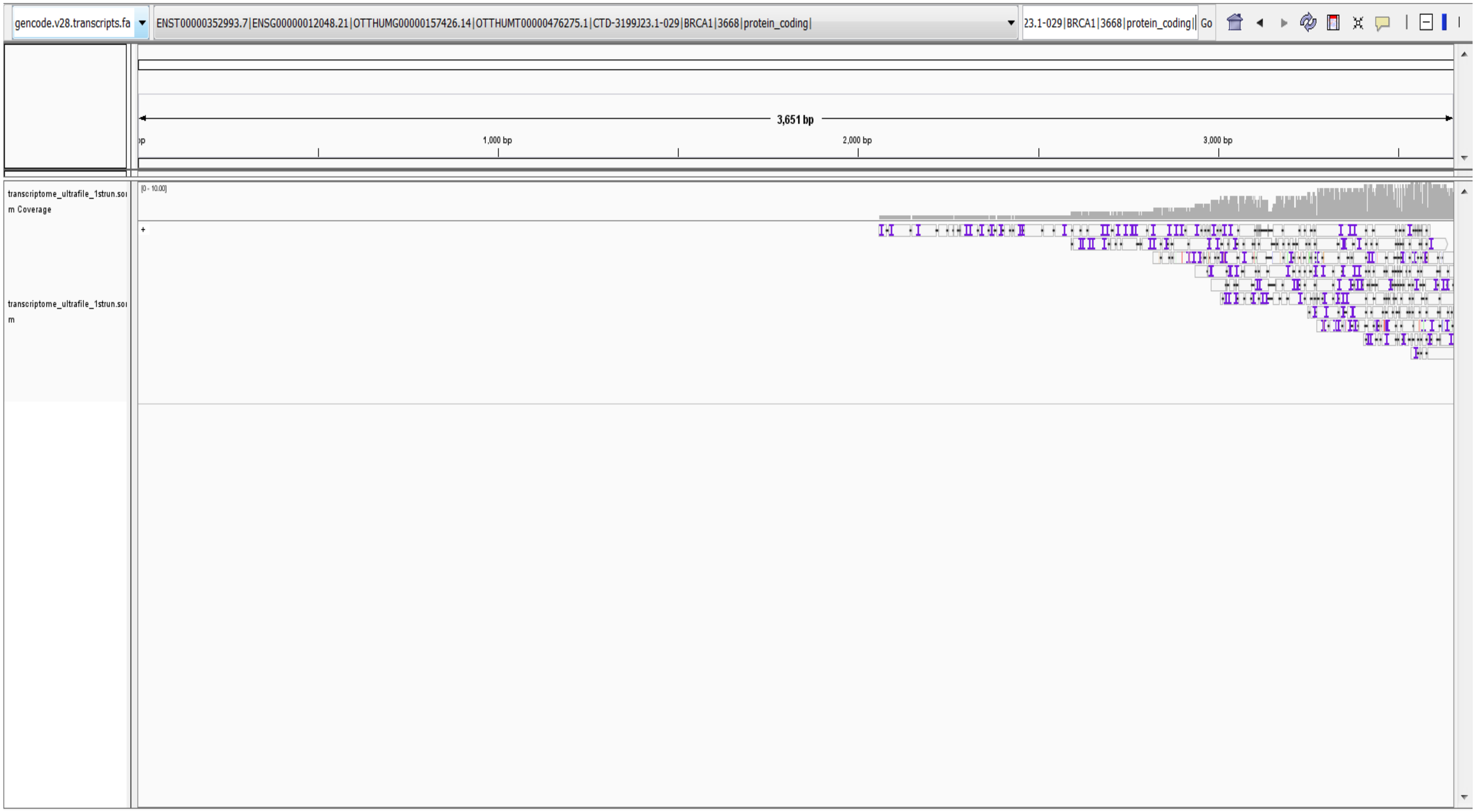
chr17

chr17:43,041,420-43,091,538

Go

Figure 3. 2 IGV screenshot showing *BRCA1* RNA transcripts from experiment 1, aligned to the Human Genome via Minimap2.

The output shows the scarcity of *BRCA1* transcripts within the isolated mRNA sample and a 3' bias. First 23 reads of *BRCA1* transcripts are illustrated with high rates of indels, which are characteristic of nanopore sequencing data. Gene structure is shown at the bottom of the track, with exons shown as blue blocks and introns represented as blue horizontal lines. Transcription direction is from right to left. The reference genome, chromosome location and coordinates are shown at the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.



gencode.v28.transcripts.fa

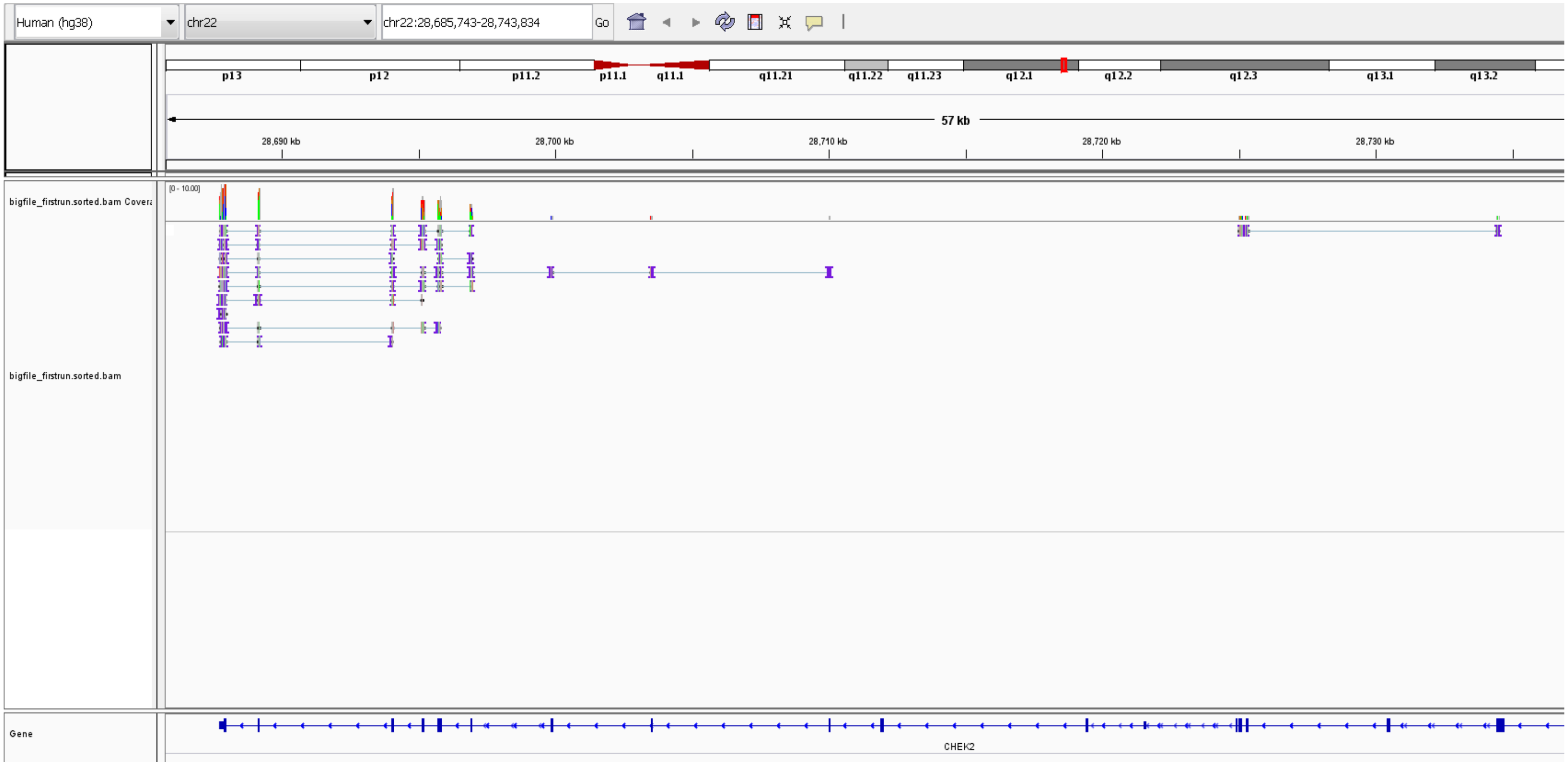
ENST00000352993.7|ENSG00000012048.21|OTTHUMG00000157426.14|OTTHUMT00000476275.1|CTD-3199J23.1-029|BRCA1|3668|protein_coding|

23.1-029|BRCA1|3668|protein_coding|

Go
3,651 bp
bp
1,000 bp
2,000 bp
3,000 bp
transcriptome_ultrafile_1strun.sorm Coverage
[0 - 10.00]
transcriptome_ultrafile_1strun.sorm
+

Figure 3. 3 IGV showing *BRCA1* RNA transcripts from experiment 1, aligned to the Human Transcriptome via Minimap2.

The output shows the scarcity of *BRCA1* transcripts within the isolated mRNA sample and a 3’ bias. First 10 reads of *BRCA1* transcripts are illustrated, with high rates of indels, which are characteristic of nanopore sequencing data. The reference transcriptome and transcript type are shown in the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.



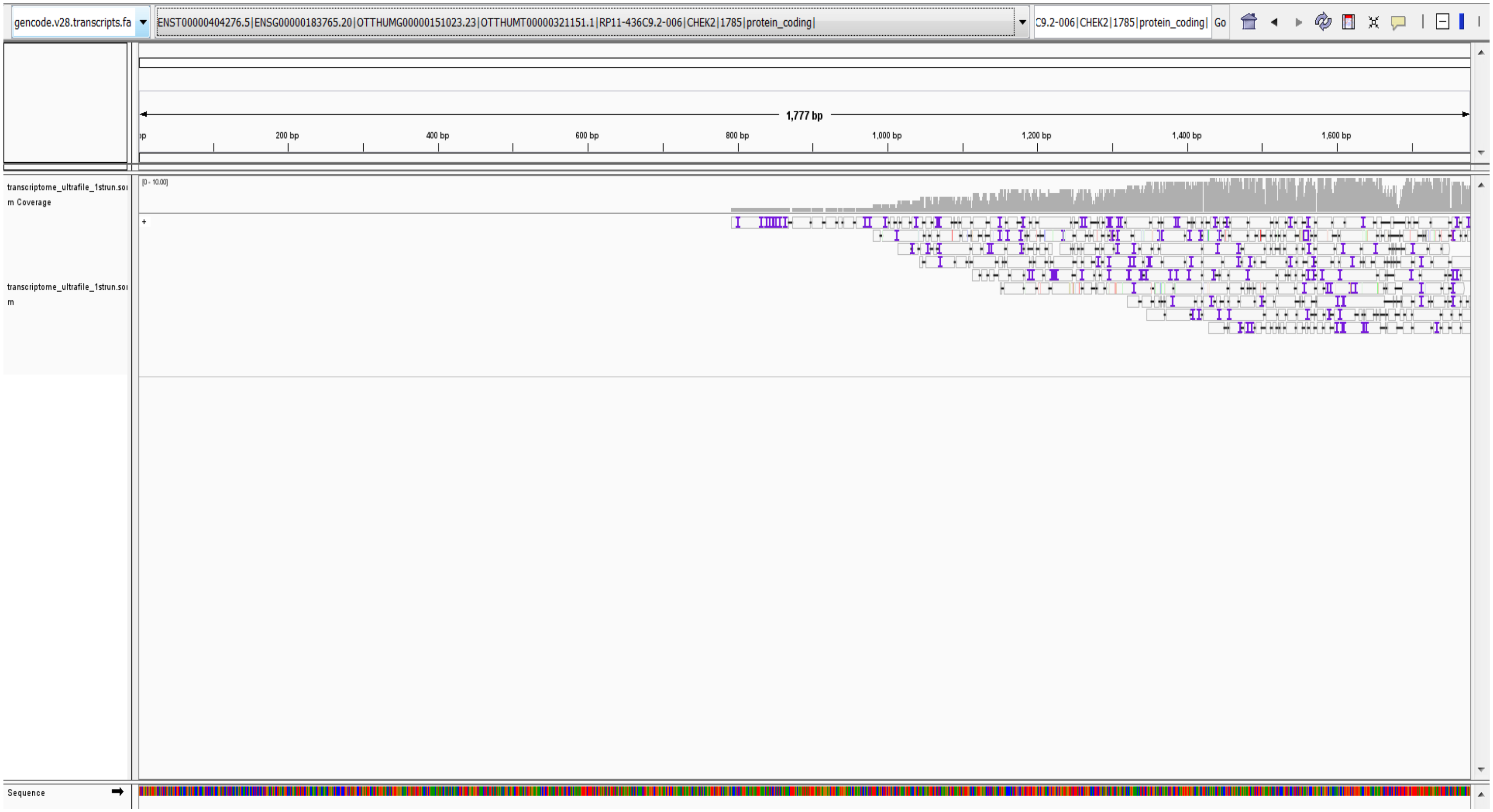
Human (hg38)

chr22

chr22:28,685,743-28,743,834

Figure 3. 4 IGV showing *CHEK2* RNA transcripts from experiment 1, aligned to the Human Genome via Minimap2.

The output shows the scarcity of *CHEK2* transcripts within the isolated mRNA sample, and a 3' bias. First 11 reads of *CHEK2* transcripts are illustrated with high rates of indels, which are characteristic of nanopore sequencing data. Gene structure is shown at the bottom of the track, with exons shown as blue blocks and introns represented as blue horizontal lines. Transcription direction is from right to left. The reference genome, chromosome location and coordinates are shown at the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.



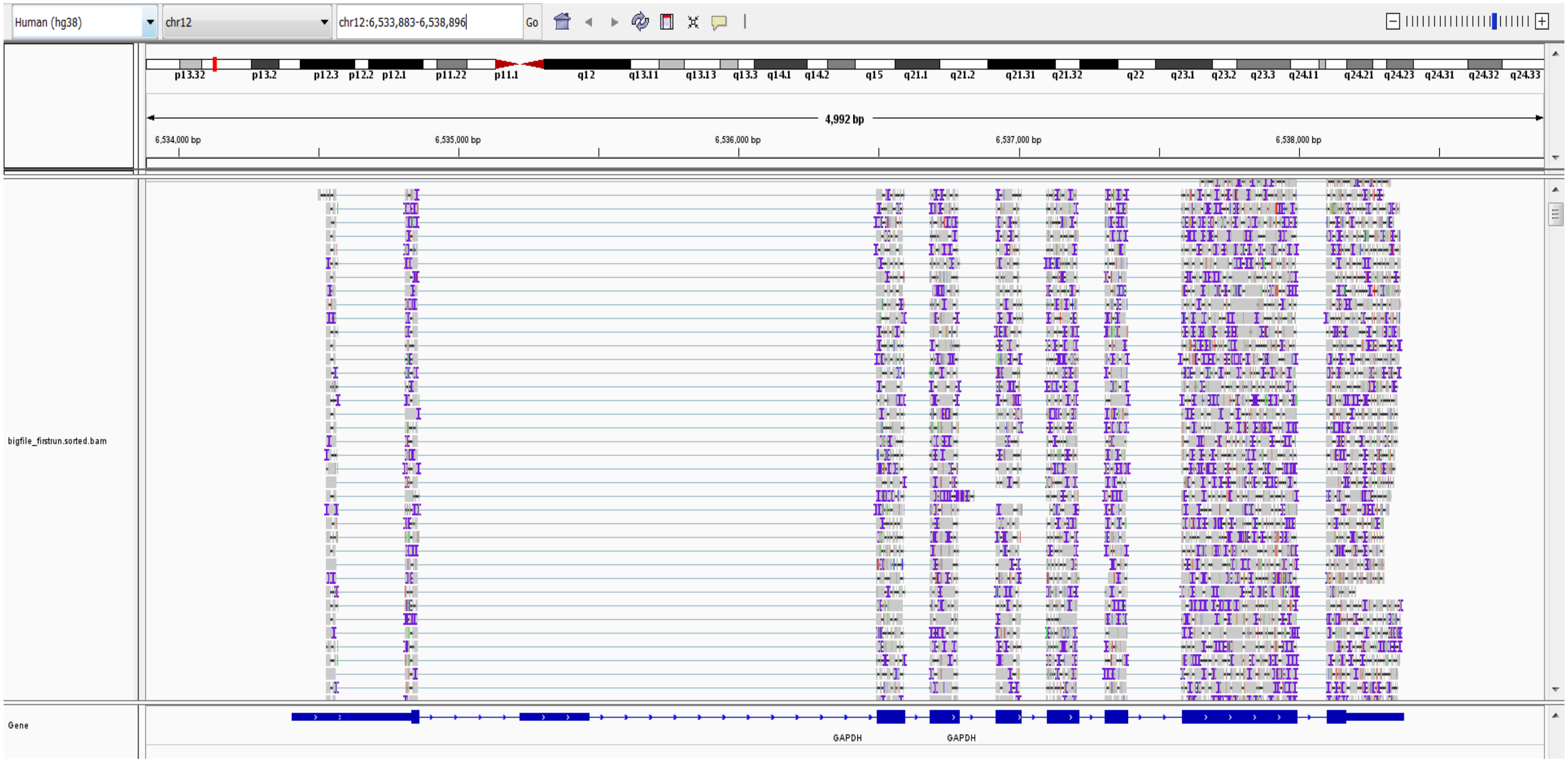
gencode.v28.transcripts.fa

ENST00000404276.5|ENSG00000183765.20|OTTHUMG00000151023.23|OTTHUMT00000321151.1|RP11-436C9.2-006|CHEK2|1785|protein_coding|

C9.2-006|CHEK2|1785|protein_coding|

Figure 3. 5 IGV showing *CHEK2* RNA transcripts from experiment 1, aligned to the Human Transcriptome via Minimap2.

The output shows the scarcity of *CHEK2* transcripts within the isolated mRNA sample and a 3’ bias. First 9 reads of *CHEK2* transcripts are illustrated, with high rates of indels, which are characteristic of nanopore sequencing data. The reference transcriptome and transcript type are shown in the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.



Human (hg38)
chr12
chr12:6,533,883-6,538,896

Figure 3. 6 IGV screenshot showing *GAPDH* RNA transcripts from experiment 1, aligned to the Human Genome via Minimap2.

The output shows the abundance of *GAPDH* transcripts within the isolated mRNA sample. First 38 reads of *GAPDH* transcripts are illustrated with high rates of indels, which are characteristic of nanopore sequencing data. Gene structure is shown at the bottom of the track, with exons shown as blue blocks and introns represented as blue horizontal lines. Transcription direction is from left to right. The reference genome, chromosome location and coordinates are shown at the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.

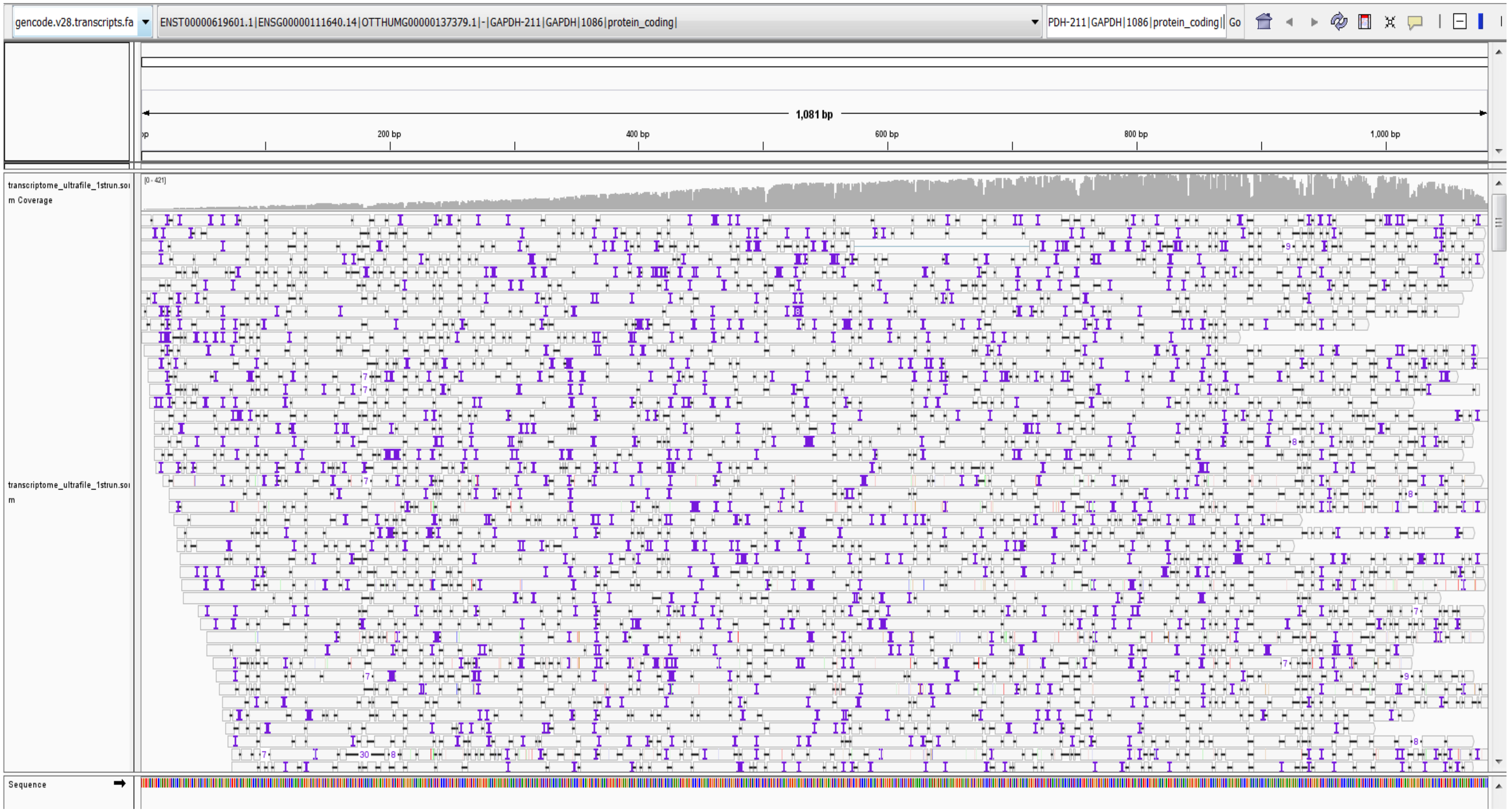

gencode.v28.transcripts.fa
ENST00000619601.1|ENSG00000111640.14|OTTHUMG00000137379.1|-|GAPDH-211|GAPDH|1086|protein_coding|
PDH-211|GAPDH|1086|protein_coding|
Go
1,081 bp
bp
200 bp
400 bp
600 bp
800 bp
1,000 bp
transcriptome_ultrafile_1strun.sorm Coverage
[0 - 421]
transcriptome_ultrafile_1strun.sorm
Sequence

Figure 3. 7 IGV showing *GAPDH* RNA transcripts from experiment 1, aligned to the Human Transcriptome via Minimap2.

The output shows the abundance of *GAPDH* transcripts within the isolated mRNA sample. First 43 reads of *GAPDH* transcripts are illustrated, with high rates of indels, which are characteristic of nanopore sequencing data. The reference transcriptome and transcript type are shown in the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.

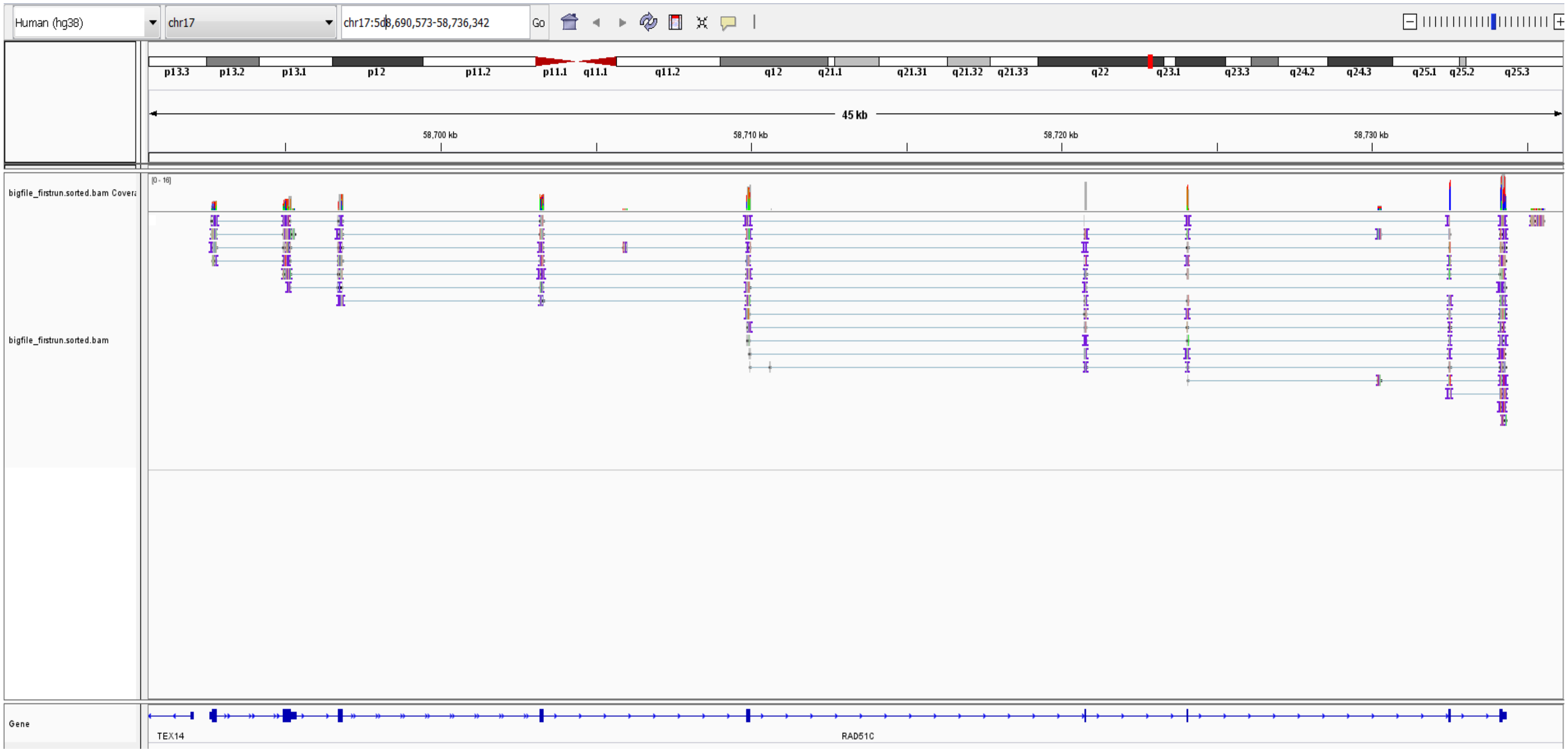


Figure 3. 8 IGV showing *RAD51C* RNA transcripts from experiment 1, aligned to the Human Genome via Minimap2.

The output shows the abundance of *RAD51C* transcripts within the isolated mRNA sample. First 16 reads of *RAD51C* transcripts are illustrated with high rates of indels, which are characteristic of nanopore sequencing data. Gene structure is shown at the bottom of the track, with exons shown as blue blocks and introns represented as blue horizontal lines. Transcription direction is from left to right. The reference genome, chromosome location and coordinates are shown at the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.

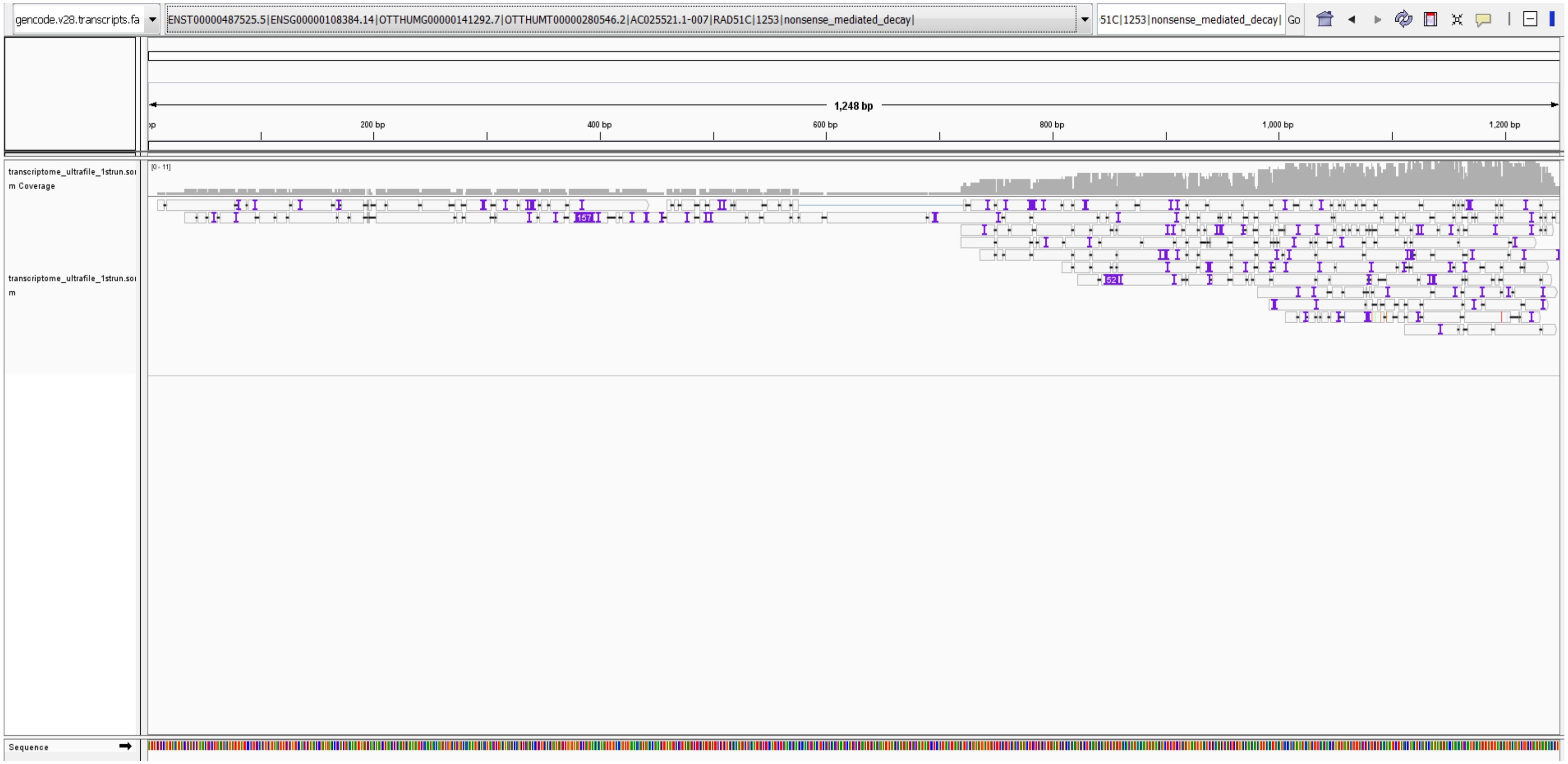

gencode.v28.transcripts.fa
ENST00000487525.5|ENSG00000108384.14|OTTHUMG00000141292.7|OTTHUMT00000280546.2|AC025521.1-007|RAD51C|1253|nonsense_mediated_decay|
51C|1253|nonsense_mediated_decay|
Go
1,248 bp
bp
200 bp
400 bp
600 bp
800 bp
1,000 bp
1,200 bp
transcriptome_ultrafile_1strun.sor
m Coverage
[0 - 11]
157
52
transcriptome_ultrafile_1strun.sor
m
Sequence

Figure 3. 9 IGV showing *RAD51C* RNA transcripts from experiment 1, aligned to the human transcriptome via Minimap2.

The output shows the abundance of *RAD51C* transcripts within the isolated mRNA sample. First 11 reads of *RAD51C* transcripts are illustrated, with high rates of indels, which are characteristic of nanopore sequencing data. The reference transcriptome and transcript type are shown in the top left corner. Coverage is shown by the density and height of vertical lines above the transcripts.

3.3.4 Quantification of MinION-generated reads

Different quantification software was trialled by testing different parameters, and the results are shown below (**Tables 3.9-3.14**). All quantification programmes were unable to quantify the aligned reads with the parameters outlined in **Tables 3.9-3.14**.

The Kallisto programme was used to create indices based on the Human Genome Reference Build hg38 and the Human Transcriptome (**Table 2.4**). The parameter ‘single’ allows quantification of single-end reads, ‘l’ sets the estimated average fragment length, ‘s’ sets the estimated standard deviation of the fragment length, and ‘t’ sets the number of threads (Pachter, 2016). The parameters used in Kallisto experiments were unsuccessful, with failure to produce integer count data (**Tables 3.9-3.10**).

Table 3. 9 Comparison of parameters creating Kallisto indices

Method number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Human Genome hg38	index -i index file name	Building a Kallisto index with the Human Genome hg38	Successful
2	Human Transcriptome	index -i index file name	Building a Kallisto index with the Human Transcriptome	Successful

Table 3. 10 List of tested parameters for DRS quantification using Kallisto

Method number	**Index**	**Parameters**	**Description of action**	**Description of results**
1	**1**	--single -l 738 -s 384 -t 4	Quantifying the transcripts from MinION run aligned to the Human Genome hg38	Unsuccessful – produced no count data
2	**2**	--single -l 738 -s 384 -t 12	Quantifying the transcripts from MinION run aligned to the Human Transcriptome	Unsuccessful – produced non-integer count data

Salmon indices were created with only the Human Transcriptome, as this was the default reference for Salmon. Both alignment and non-alignment modes were trialled. The parameter ‘l’ sets the library type, ‘p’ sets the number of threads, ‘fldMean’ sets the expected mean fragment length, ‘fldSD’ sets the expected standard deviation, ‘validateMappings’ runs an extension alignment dynamic on the quasi-mapppings and ‘rangeFactorizationBins’ is a parameter that may allow improved quantification estimates. The parameter ‘gcbias’ allows correction for GC bias in the data, ‘writeUnmappedNames’ makes Salmon write out the names of unmapped reads, and ‘minAssignedFrags’ sets the minimum number of assigned fragments (R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, & C. Kingsford, 2017) (**Tables 3.11-3.12**). Alignment-based mode with sorted BAM files of Minimap2 aligned reads were able to generate count data, which were non-integer values. Therefore, Salmon was unable to successfully quantify the reads.

Table 3. 11 Parameters for creating Salmon indices

Method number	**Reference**	**Parameters**	**Description of action**	**Description of results**
1	Human Transcriptome	index -t reference genome --type quasi -k 31	Building a Salmon index with the Human Transcriptome in quasi-index mode with default parameters	Successful

Table 3. 12 List of tested parameters for DRS quantification using Salmon

Method number	**Index**	**Parameters**	**Description of action**	**Description of results**
1	**1**	quant -l A -p 4 --fldMean 738 --fldSD 384 --validateMappings --rangeFactorizationBins 4 --gcBias --writeUnmappedNames u	Quantifying the transcripts from MinION run aligned to the Human Genome hg38 with an index created from the Human Transcriptome in quasi-index mode	Unsuccessful
2	**1**	quant -l A -p 4 --fldMean 73 --fldSD 384 --validateMappings --rangeFactorizationBins 4 --writeUnmappedNames --minAssignedFrags 1	Quantifying the transcripts from MinION run aligned to the Human Genome hg38 with an index created from the Human Transcriptome in quasi-index mode	Unsuccessful
3	**1**	quant -l A	Quantifying the transcripts from MinION run aligned to the Human Genome hg38 with an index created from the Human Transcriptome in quasi-index mode	Unsuccessful
4	**1**	quant -l A	Quantifying the transcripts from MinION run aligned to the Human Transcriptome with an index created from the Human Transcriptome in quasi-index mode	Unsuccessful
5	**None**	quant -l A (sorted BAM file)	Quantifying the transcripts from MinION run aligned to the Human Transcriptome in alignment-based mode with a sorted BAM file	Unsuccessful – produces count data with integer values which do not correspond to the output on IGV
6	**None**	quant -l A (unsorted bam file)	Quantifying the transcripts from MinION run aligned to the Human Transcriptome in alignment-based mode with an unsorted BAM file	Unsuccessful – produces non-integer count data

Mandalorion – Episode II was used in order to quantify the isoforms aligned by Minimap2. The parameter ‘u’ defines how lenient TSS and polyA sites are defined upstream, ‘d’ defines how lenient TSS and polyA sites are defined downstream, ‘s’ defines how many randomly sampled subreads are used to create an isoform consensus, ‘r’ determines the minimum ratio of reads aligned to a locus, ‘R’ sets the minimum number of reads that have to be assigned to an isoform for the isoform to be reported, ‘I’ sets the maximum 5’ overhang, ‘i’ sets the minimum 5’ overhang, ‘T’ sets the maximum 3’ overhang, and ‘t’ sets the minimum 3’ overhang (R. Volden & Vollmers, 2018) (**Table 3.13**).

Table 3. 13 List of tested parameters for DRS quantification using Mandalorion – Episode II

Method number	Reference	Parameters	Description of action	Description of results
1	Gencode v28 annotation GTF file	Python3 defineAndQuantifyIsoforms.py -c content file -p output directory -u 5 -d 30 -s 200 -r 0.05 -R 1 -I 0 -t 0 -I 100 -T 60 -g path to Mandalorion NUC.4.4.mat file Path to annotation GTF file -m config_file	Quantifying the transcripts from MinION run aligned to the Human Transcriptome and referenced to the Gencode v28 annotation GTF file with the defineAndQuantifyIsoforms.py python script.	Unsuccessful- Error: ValueError: invalid literal for int() with base 10
2	Gencode v28 annotation GTF file	Python3 defineAndQuantifyWrapper.py -c content file -p output directory -u 5 -d 30 -s 200 -r 0.05 -R 1 -I 0 -t 0 -I 100 -T 60 -g path to Mandalorion NUC.4.4.mat file Path to annotation GTF file -m config_file	Quantifying the transcripts from MinION run aligned to the Human Transcriptome and referenced to the Gencode v28 annotation GTF file with the defineAndQuantifyWrapper.py python script.	Unsuccessful -Error: UnicodeDecodeError: ‘utf-8’ codec can’t decode bytes in position 15-16: invalid continuation byte

SQANTI utilised GMAP databases (**Table 3.5**) and was used to quantify reads in a FASTA format. The parameter ‘n’ allows the gene name tag from the annotated GTF file to be used to define the gene and the parameter ‘t’ sets the number of GMAP threads (de la Fuente & Tardaguila, 2018) (**Table 3.14**). Various errors were produced through these steps, therefore, quantification using these parameters were unsuccessful (**Table 3.14**).

Table 3. 14 List of tested parameters for DRS quantification using SQANTI

Method number	**GMAP database**	**Parameters**	**Description of action**	**Description of results**
1	**1**	Python Sqanti_qc.py -x path to gmap database -t 4 -o name of output file -d path to output directory Query FASTA file Annotation GTF file Reference FASTA file	Quantifying the transcripts from MinION run aligned to the Human Transcriptome	Unsuccessful: -Many no paths found errors -OSError: [Errno 13] Permission denied
2	**2**	Python Sqanti_qc.py -x path to gmap database -t 4 -o name of output file -d path to output directory Query FASTA file Annotation GTF file Reference FASTA file	Quantifying the transcripts from MinION run aligned to the Human Genome (Used the cDNA FASTA)	Unsuccessful -Many no paths found errors -OSError: [Errno 13] Permission denied
3	**1**	Python Path to Sqanti_qc.py Query FASTA file Annotation GTF file Reference FASTA file -n -t 4 -x path to gmap database	Quantifying the transcripts from MinION run aligned to the Human Transcriptome (Used the cDNA FASTA)	Unsuccessful – IndexError: list index out of range

3.4 Discussion

3.4.1 Overview

The results of this chapter revealed that a successful first DRS run had been carried out in this laboratory by visualising the reads on IGV (**Figures 3.2-3.9**). Satisfactory quality mRNA was isolated and sequenced with the MinION via the Direct RNA Sequencing Kit. The results showed that the MinION had successfully sequenced full-length isoforms of many genes including *GAPDH* and *RAD51C*. However, full-length transcripts of many poorly expressed genes, such as *BRCA1*, were not well represented using a whole transcriptome DRS approach (**Figure 3.2**). Only four genes of interest have been described in this chapter, however, many other genes were also sequenced as a part of this whole transcriptome sequencing experiment. The data generated from this experiment was used as a control experiment for subsequent enrichment experiments outlined in chapter 4.

3.4.2 Basic data acquisition

Data acquisition with MinKNOW and Albacore was straightforward, as the software was provided by ONT. Basecalling with Albacore produced ‘pass’ and ‘fail’ reads, in which the number of ‘pass’ reads were greater than ‘fail’ reads in this experiment. This suggests that most of the reads were of satisfactory quality to be sequenced with the MinION.

3.4.3 Alignment

The results showed positive alignment with the use of Minimap2. However, BLAT, GMAP and LAST did not produce successful alignments. These four aligners were tested as they have been previously used with nanopore data (Croville et al., 2018; de Jong et al., 2017; Ip et al., 2015). Minimap2 indices were successfully created with moderate to high k-mers, whereas low k-mers were unable to produce an index. K-mers are all the possible subsequences that may be generated from a string of length k (Marcais & Kingsford, 2011). Therefore, using a k-mer length of 1 may have caused this unsuccessful result (**Table 3.1**). Genome alignment was preferred over transcriptome alignment as it provided exon-intron information, which was beneficial for detecting exon-skipping events on IGV. The compiled Human Genome hg38 (**Table 2.4**) by Dr Heng Li was used initially as it was believed that the standard Human Genome Reference Build hg38 was inadequate for alignment, with properties such as the exclusion of unplaced and unlocalised contigs (H. Li, 2017). However, upon closer examination, the alignment of *BRCA1* was similar or slightly improved compared to the standard Human Genome Reference Build hg38 (**Table 2.4**). Therefore, the rest of the analyses were carried out with the standard genome.

Minimap2 has defined parameters specific for ONT DRS data, whereas BLAT lacked this information. BLAT produced PSLX and BLAST8 files (**Table 3.4**), and the PSLX files were converted to SAM files via the script psl2sam.py. The resulting SAM file was converted to BAM, sorted, and indexed with Samtools. On IGV, the sorted BAM file was incomprehensible, showing unidentifiable reads with lack of exon-intron information even with aligning to the Human Genome instead of the Human Transcriptome. BLAT uses a client/server mode, and its cDNA genomic alignments are known to be ambiguous (T. D. Wu & Watanabe, 2005).

Alignment to the BRCARD1_geneseq.fasta file (**Table 2.4**) also followed these results. One reason which may have caused these results may be due to DRS reads being aligned instead of converted cDNA sequences.

In contrast, GMAP is beneficial for obtaining isoform information, however, it proved to be more laborious than Minimap2 for processing ONT DRS reads. As GMAP only processes cDNA reads, the DRS were converted to cDNA sequences and aligned to a transcriptome. All alignment trials presented with a ‘no paths found’ error (**Tables 3.5-3.6**), which may be due to the low quality of the reads. GMAP involves a single query/interactive mode instead of a client/server mode. This allows higher processing speed for individual queries, as GMAP searches for oligomers directly from the file instead of accessing the data from a server (T. D. Wu & Watanabe, 2005).

A LAST database was created successfully (**Table 3.7**), however, the given parameters for long DRS reads were unable to align the data generated from experiment 1 **(Table 3.8**). Therefore, this may suggest that the reads generated from the DRS experiment or the trialled parameters were suboptimal. The parameters for minimum score for gapless alignment, maximum initial matches per query position and query letters per random alignment should be tested in future experiments to optimise this alignment.

The bioinformatic errors observed in these experiments (**Tables 3.4-3.6 & 3.8**) were unable to be overcome due to time constraints. For these collective reasons, Minimap2 was used as the aligner for the final pipeline and was utilised for aligning all subsequent DRS data and the external Nanopore RNA Consortium dataset (**Table 2.4**).

3.4.4 Quantification of reads

Quantification programmes also proved to be problematic (**Tables 3.9-3.14**), with all tested programmes failing to quantify reads aligned with various reference genomes (**Table 2.4**). Kallisto was unable to produce count data, whereas Salmon was able to produce count data including TPM and number of reads by using the Transcriptome as the reference genome. However, the output of Salmon did not correlate with the reads visualised on IGV. Therefore, it appeared that Salmon was unable to accurately quantify reads from long DRS reads. This result may have occurred because these programmes have been designed for short or long RNA-seq reads, as opposed to DRS reads. Even with the reads being converted to cDNA sequences, the quantification programmes failed to count isoforms.

Mandalorion – Episode II (**Table 3.13**) and SQANTI (**Table 3.14**) have been designed for long nanopore and PacBio reads, respectively. Mandalorian – Episode II is specific for Rolling Circle Amplification to Concatemeric Consensus (R2C2) (R. Volden et al., 2018), and is not specific for DRS. Similarly, SQANTI has been designed for long-read RNA-seq and was suggested as a putative quantifier for long DRS reads. However, none of the attempted experiments were able to produce a successful outcome (**Table 3.14**).

Therefore, reads were manually counted by visualising them on IGV, and recording the data in an Excel spreadsheet. Each gene of interest was designated for a particular type of reference transcript variant to standardise the exon numbering across the datasets (*BRCA1* – NM_007294, *CHEK2* – NM_001257387, *GAPDH* – NM_001289745, *RAD51C* – NM_058216). During this process, human error may have been introduced, which means that the outputs of these analyses may be inaccurate. However, some inferences can be cautiously made with these data, as in the upcoming chapters.

3.4.5 Comparison with the Nanopore RNA Consortium dataset

As replicated experiments were not carried out, the data from experiment 1 was supplemented with an external dataset from the Nanopore RNA Consortium (**Table 2.4**). The consortium data involved 30 flow-cells which generated approximately 13 million DRS reads, whereas the results from experiment 1 involved 1 flow-cell, which sequenced 889,850 reads. Therefore, the throughput of this experiment 1 was satisfactory. The Nanopore RNA Consortium data implemented poly-A RNA instead of total RNA as in experiment 1. However, the consortium data used a control LCL (NA 12878) without NMD inhibition, whereas experiment 1 utilised a different control LCL (**Table 2.1**) with NMD inhibition. The specific comparisons between different experiments and external datasets will be outlined in subsequent chapters. In chapter 4, a novel method for enriching targeted transcripts will be explored.

Chapter 4: Developing a gene enrichment method for targeted direct RNA sequencing

4.1 Introduction

The purpose of these experiments was to test a novel linear enrichment method to amplify RNA isoforms of interest, in order to provide an improved method of examining the effects of VUS on splicing patterns. A brief overview of the sequence of steps involved in these experiments are outlined in **Figure 3.1** and **Figure 4.1**. In experiment 2, *GAPDH* was utilised as a control gene and was also enriched and sequenced along with *BRCA1*. In experiment 3, *BRCA1* and 11 other hereditary breast and ovarian cancer genes with a range of different transcript lengths and expression levels were utilised to test the enrichment with this novel enrichment technique (**Table 1.1**). Because this multi-gene approach did not prove to be very successful, experiment 4 was a more modest approach to enrichment of four genes, *GAPDH* (housekeeping gene), *BRCA1*, *CHEK2* and *RAD51C* (hereditary breast and ovarian cancer genes). *GAPDH*, *CHEK2* and *RAD51C* have similar transcript lengths (1.2 KB-1.9 KB) (**Table 4.1**), but different expression levels. Experiment 1 data, external RNA-seq data and external DRS data (**Table 2.4**) were implemented to compare the results of these experiments.

Table 4. 1 List of genes of interest implemented in this study and their characteristics

Gene	Gene Length	Transcript Length	Expression Level
BRCA1	81 KB	7.2 KB	Low
BRCA2	84 KB	11 KB	Moderate
PALB2	38.1 KB	4.7 KB	High
RAD51B	863 KB	3.3 KB	Very high
RAD51C	42.2 KB	1.2 KB	High
RAD51D	20 KB	2.5 KB	High
BRIP1	185 KB	6.4 KB	Moderate
BARD1	84 KB	4 KB	Moderate
ATM	147 KB	10 KB	Low
CHEK2	54.1 KB	1.9 KB	Moderate
FANCM	64.9 KB	6.4 KB	Low
NBN	51.4 KB	4.6 KB	High
GAPDH	3.8 KB	1.4 KB	Very high

*The transcript lengths were inferred from Geneious (version 11.1.4) and the expression level was inferred from an external RNA-seq dataset (**Table 2.4**).

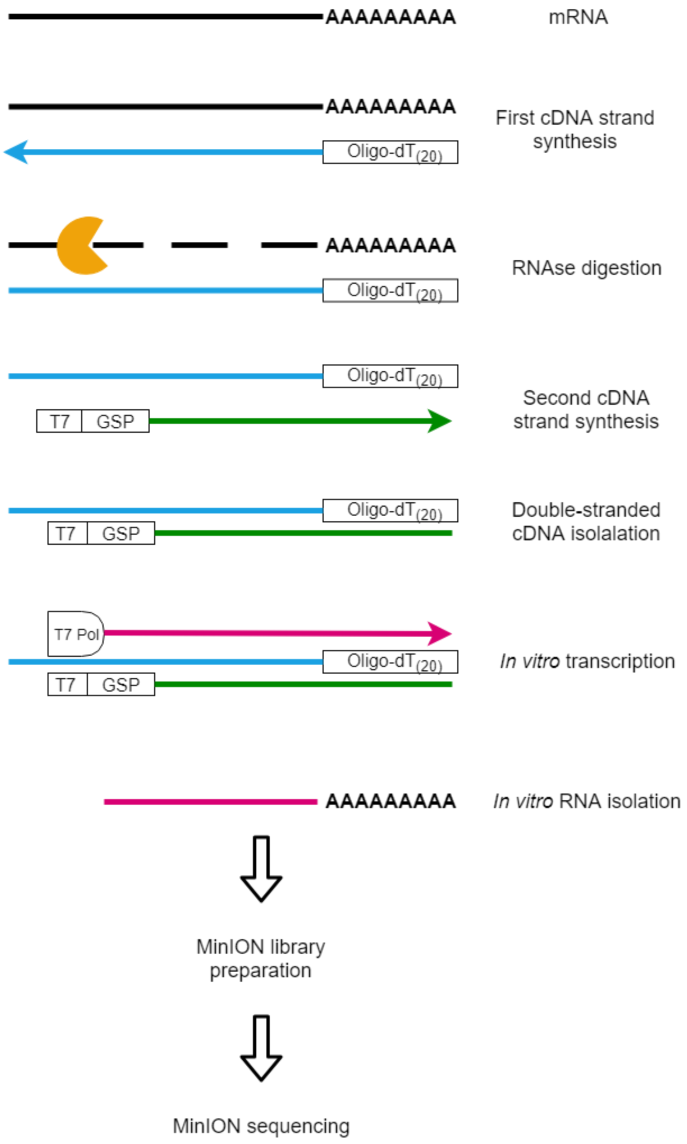


Figure 4. 1 A brief schematic of the enrichment approach.

The mRNA from the extracted total RNA was reverse-transcribed into an RNA-cDNA hybrid using an anchored oligo-dT$_{(20)}$ primer. An RNAse was implemented to digest the RNA strand, whereas the cDNA remained intact. A second cDNA strand was synthesised in order to incorporate a T7 polymerase binding site attached to a gene specific primer. The double-stranded cDNA was isolated by HighPrep™ magnetic beads. The T7 RNA polymerase was utilised to generate *in vitro* transcribed RNA. This RNA was isolated by RNAClean™ XP magnetic beads and used to generate a MinION library. The library was loaded on a flow-cell and sequenced with the MinION.

4.2 RNA source/cell culturing

More control LCL (**Table 2.1**) batches were grown to carry out experiments 2-4 with the same aforementioned conditions. From these cultures, another 15 RNA extractions were carried out via the RNeasy® Mini Kit. The extracted total RNA had satisfactory quality, determined from NanoDrop™, Qubit™ RNA HS and Tapestation assays. Tapestation results were consistently above a RINe score of 9.3. However, NanoDrop™ and Qubit™ results varied between all 15 RNA samples. The average NanoDrop™ concentration was 878.21 ng/µL and the average Qubit™ RNA HS result was 585.85 ng/µL.

4.3 First cDNA strand synthesis

The first strand of cDNA was reverse-transcribed from total RNA extracted from a control LCL using the enzyme Superscript® III Reverse Transcriptase according to **Appendix C**. Here, an anchored oligo-dT$_{(20)}$ primer was utilised. The RNA strand was degraded by the use of an RNAse enzyme (**Table 2.1**), in which the cDNA strand was left single-stranded (**Figure 4.1**). Quality of the cDNA from this step was examined by a series of different quality control experiments, which are outlined below.

4.3.1 Standard long-range PCR

Thirteen standard long-range PCR reactions were carried out to determine the efficacy of the first cDNA strand synthesis step. *BRCA1* exon 1 forward and exon 24p reverse primers were utilised to visalise any full-length *BRCA1* cDNA present in the single-stranded cDNA sample. The results shown on one gel revealed prominent bands of approximately 2.4 KB (**Figure 4.2 A**), and the results from the second gel revealed a bright, prominent band of approximately 2.4 KB and thinner bands around 6 KB and 2.2 KB (**Figure 4.2 B**). The 6 KB band represented full-length isoforms of *BRCA1*. However, the rest of the PCRs had failed, with smears observed as opposed to distinct bands. Therefore, the alternative long-range PCR was explored.

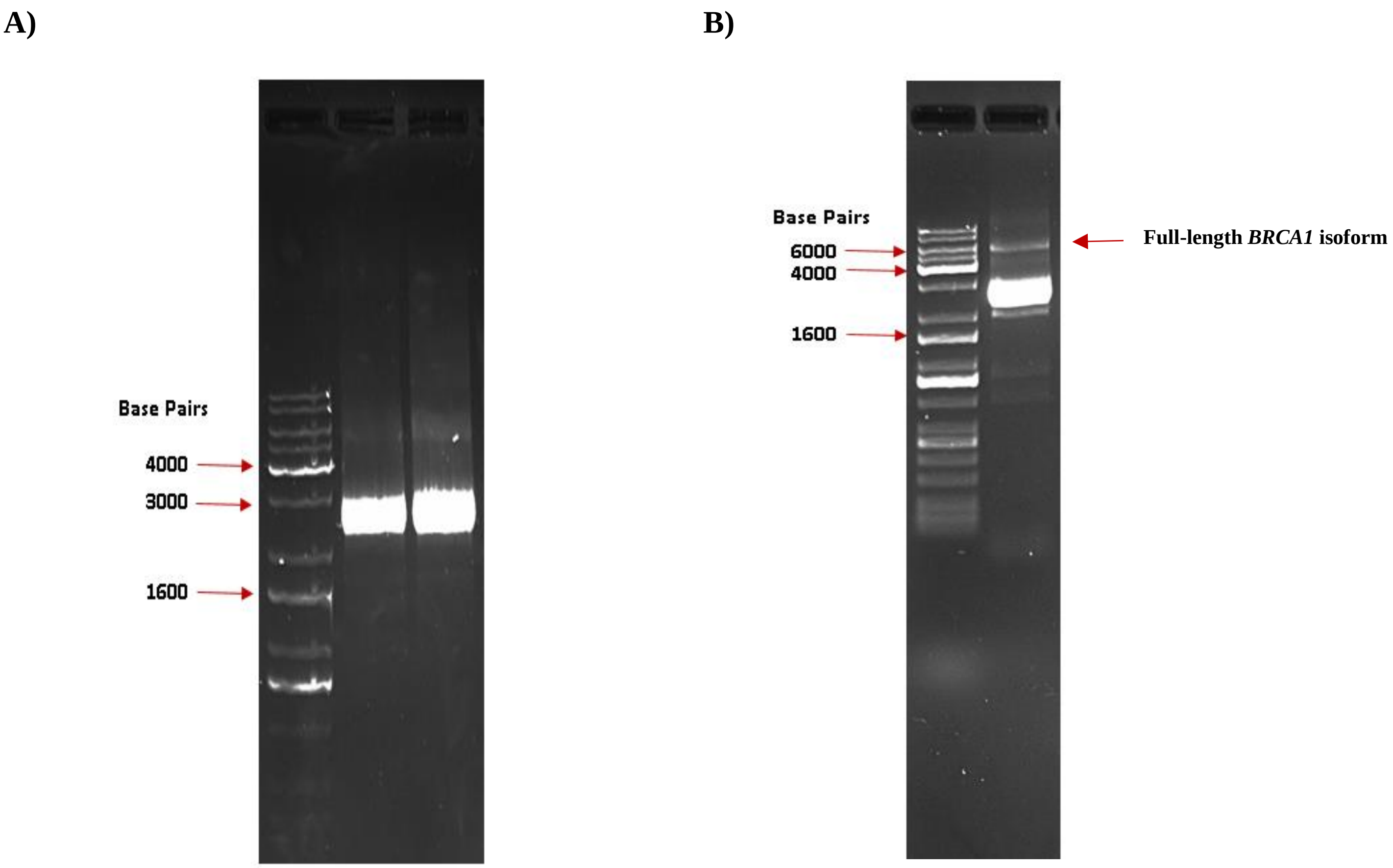


Figure 4. 2 Agarose electrophoresis gels of standard long-range PCR products.

BRCA1 exon 1 forward and exon 24p reverse primers were used in the PCR. **A)** Thick bands around the size of 2.4 KB were shown. **B)** A thick band around 2.4 KB, along with two thinner bands around 2.2 KB and 6 KB were shown. The gels were set at 1 % and run at 90 V for 30 minutes – 1 hour.

4.3.2 Alternative long-range PCR

An alternative long-range PCR (**Figure 4.3**) was carried out to determine the efficacy of the first strand cDNA synthesis step, as a substitute for the standard long-range PCR. This PCR consisted of a higher $MgCl_2$ and cDNA concentration, lower dNTP concentration and lower initial denaturation and extension times compared to the standard long-range PCR (**Figure 4.2**). *BRCA1* exon 1 forward and exon 24p reverse primers were utilised to visalise whether any full-length BRCA1 cDNA was present in the single-stranded cDNA sample. The gel produced two bright and prominent bands of approximately 5.8 and 2.8 KB sizes, which presented improved bands compared to **Fig. 4.2 A & B**.

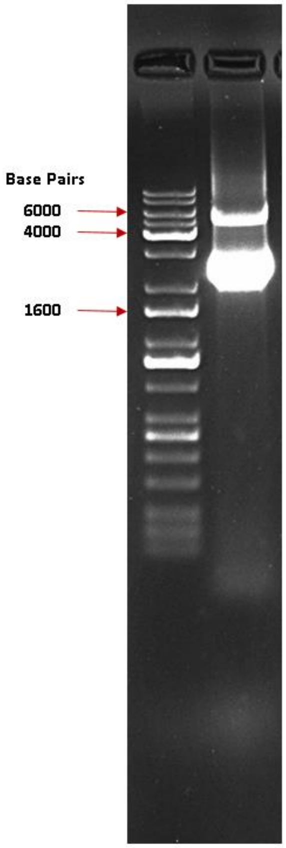


Figure 4. 3 Agarose electrophoresis gel of alternative long-range PCR products.

The gels were set at 1% at 90 V for 30 minutes – 1 hour. *BRCA1* exon 1 forward and exon 24p reverse primers were used in the PCR. The gel showed presence of bright and prominent bands of approximately 5.8 KB and 2.8 KB sizes.

4.4 Second cDNA strand synthesis

In experiment 2, *GAPDH* and *BRCA1* T7-GSP forward primers were added separately into two different reaction volumes for the targeted enrichment of *GAPDH* and *BRCA1*. In experiment 3, 12 T7-GSP forward primers were pooled and mixed, and an aliquot of 5 µM were used in one reaction volume (**Table 2.2**). In experiment 4, four reaction volumes were carried through the second strand cDNA synthesis step, one for each gene of interest; *GAPDH*, *BRCA1*, *RAD51C*, *CHEK2*. Different cycling conditions were used in experiment 4 to the other experiments. Experiments 2 and 3 used 1 cycle of 94 °C for 1 minute, 50 °C for 1 minute and 65 °C for 30 minutes. Experiment 4 used 3 cycles of 94 °C for 1 minute, 50 °C for 1 minute and 65 °C for 30 seconds. The product was then placed into a refrigerator for 10 minutes, and an extra step of 65 °C for 30 minutes were exerted on the dscDNA product.

4.4.1 Temperature gradient PCR

The temperature gradient PCR was used to determine the optimum annealing temperature for the second cDNA strand synthesis step. The manufacturer recommends an annealing temperature ranging from 45 to 65 °C for the second cDNA synthesis step (https://international.neb.com/protocols/2012/10/15/m0323-longamp-taq-dna-polymerase-protocol). However, as each melting temperature (Tm) of a primer was different to other primers (**Table 2.2**), the optimum annealing temperature was speculated to be also different for each primer. Hence, a temperature gradient PCR was carried out to determine the optimum annealing temperature for the T7-GSP primer for *BRCA1* and *GAPDH*. *BRCA1*_F1 and *BRCA1*_7R primers were used for *BRCA1*, and *GAPDH*_F1 and *GAPDH*_R9 primers were used for *GAPDH*. The results of the temperature gradient PCR showed that temperatures from 49.8 °C to 62.0 °C proved to be optimum annealing temperatures for the second cDNA synthesis for *BRCA1* (**Figure 4.4**). In comparison, temperatures from 49.8 °C to 65.1 °C were found to be optimum annealing temperatures for *GAPDH* (**Figure 4.5**). Therefore, 50 °C was used as the annealing temperature for the second cDNA strand synthesis for all enrichment experiments. Furthermore, expected band sizes of 900 bps and 1.2 KB were found for *BRCA1* (**Figure 4.4**) and *GAPDH* (**Figure 4.5**), respectively.

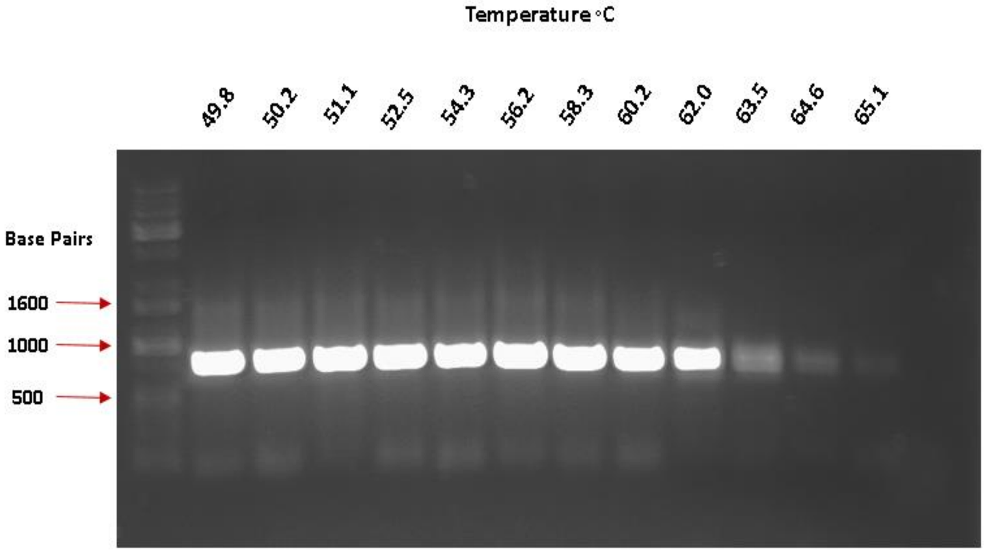


Figure 4. 4 Agarose electrophoresis gel of *BRCA1* temperature gradient PCR products.

BRCA1 exon 1 forward and exon 7 reverse primers were used in the PCR. Products with annealing temperatures from 49.8 °C to 62.0 °C produced prominent fragments of around 900 bps. With increasing temperatures above 62 ºC, the quantity of product tailed off markedly. The gels were set at 1 % at 90 V for 30 minutes – 1 hour.

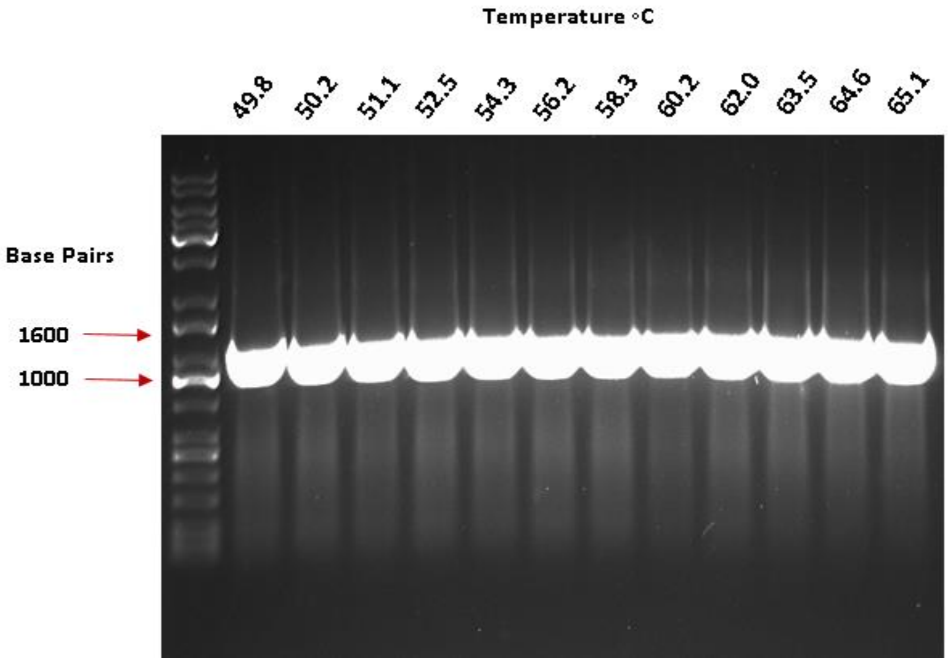


Figure 4. 5 Agarose electrophoresis gel of *GAPDH* temperature gradient PCR products.

GAPDH exon 1 forward and exon 9 reverse primers were used in the PCR. Products with annealing temperatures from 49.8 °C to 65.1 °C produced prominent fragments of around 1.2 KB. The gels were set at 1 % at 90 V for 30 minutes – 1 hour.

4.4.2 T7-Only Primer PCR

A T7-only primer PCR was utilised for testing the efficacy of the second cDNA strand synthesis. The T7-GSP primer was used as a means of providing a binding site for the T7 RNA polymerase, with a gene specific primer attached to target the gene of interest. This was necessary for the *in vitro* transcription step which followed the second cDNA strand synthesis step (**Figure 4.1**). A forward primer composed only of a T7-polymerase binding site, and a reverse primer close to the 5' end of the second cDNA strand were implemented to test for incorporation of the T7-polymerase binding site into the second cDNA strand (**Figure 4.6**). The resulting gel revealed a band around 810 bps (**Figure 4.7**), which is the approximate length of exon 1 to exon 7 on *BRCA1*. This provides evidence that the T7 polymerase binding site had been successfully incorporated in the double-stranded cDNA.

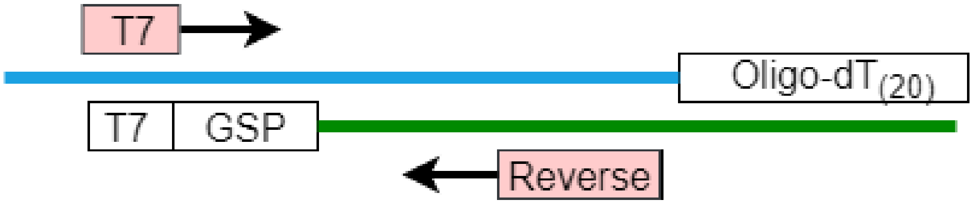


Figure 4. 6 Diagram of a T7-only PCR.

The forward primer was the T7-only primer and the reverse primer was an exon 7 reverse for *BRCA1*. This PCR determined whether the T7-GSP had been successfully incorporated into the second cDNA strand.

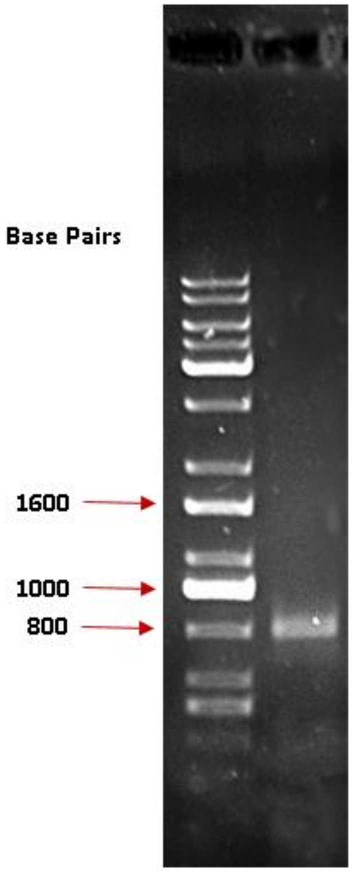


Figure 4. 7 Agarose electrophoresis gel of *BRCA1* T7-only forward and exon 7 reverse primer PCR.

A single band of approximately 810 bps was observed. The gel was set at 1 % at 90 V for 30 minutes – 1 hour.

4.4.3 T7-only & oligo-dT$_{(20)}$ (full-length) PCR

T7-only and oligo-dT$_{(20)}$ primers were used as forward and reverse primers for this PCR, respectively (**Figure 4.8**). This was carried out in order to determine whether the second cDNA synthesis step had been carried out in a satisfactory manner in experiment 4 to produce dscDNA which spanned the whole length of the gene of interest. The products of four genes of interest (*BRCA1*, *CHEK2*, *GAPDH*, *RAD51C*) were tested. Ideally, the second cDNA strand should overlap the oligo-dT$_{(20)}$ to allow the *in vitro* transcribed RNA to contain a poly-A tail, which is crucial for the sequencing step (**Figure 4.1**). The gels showed varied results between different genes of interest (**Figure 4.9**). Genes with relatively high expression levels (*GAPDH* and *RAD51C*) presented bright, thick bands of approximately 1.5 KB and 1.2 KB, respectively. The genes with relatively low expression (*BRCA1* and *CHEK2*) presented multiple thin, faint bands across a smear.

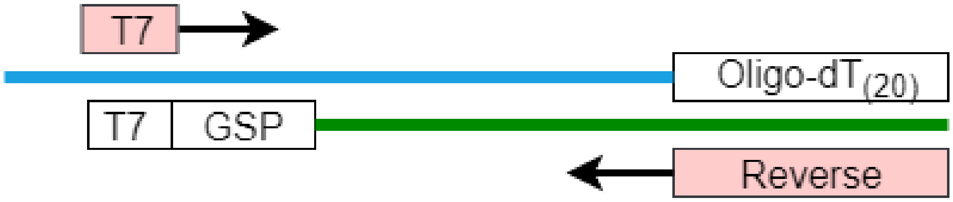


Figure 4. 8 Diagram of a T7-only & oligo-dT(20) PCR.

The T7 RNA polymerase binding site primer was utilised as the forward primer, and the anchored oligo-dT$_{(20)}$ primer was used as a reverse primer in these experiments. This PCR determines whether the second cDNA strand synthesis had been able to be completed to the 5’ end of the first cDNA strand.

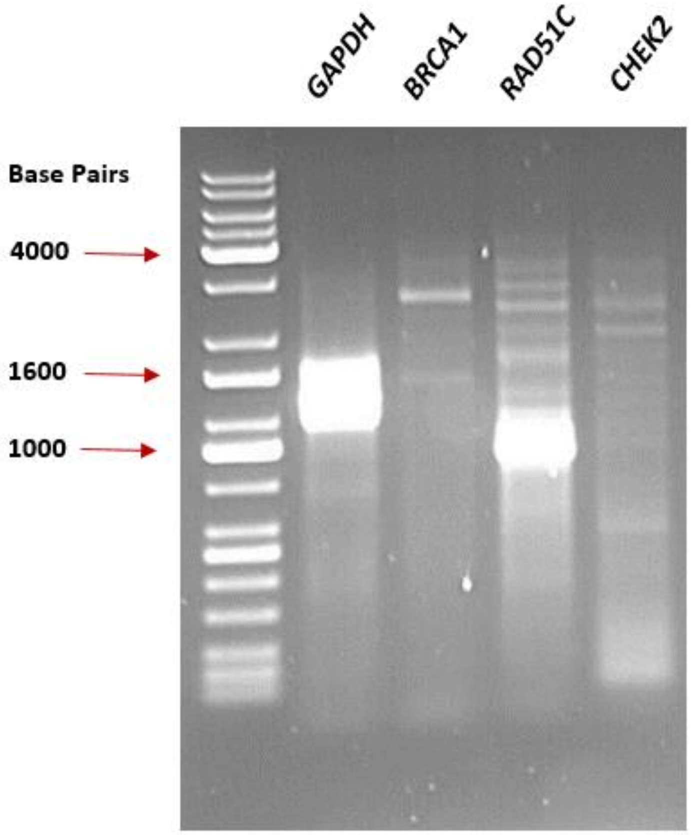


Figure 4. 9 Agarose electrophoresis gel of T7-only forward and anchored oligo-dT$_{(20)}$ reverse primer PCR. The products of *GAPDH*, *BRCA1*, *RAD51C, CHEK2*. *GAPDH* and *RAD51C* presented bright, thick bands with sizes of 1.5 KB and 1.1 KB respectively. All four genes of interest showed smears across the length of the ladder, and thin, distinct bands were shown for *BRCA1*, *RAD51C* and *CHEK2* samples above 2 KB. The gels were set at 1% at 90 V for 30 minutes – 1 hour.

4.5 HighPrep™ magnetic bead clean-up

The double-stranded cDNA produced with the T7-GSP primer was isolated by a HighPrep™ magnetic bead clean-up step. The double-stranded cDNA was eluted in 20 µL to concentrate the sample for use in the *in vitro* transcription step. The elution was analysed through the Qubit™ dsDNA HS Assay. In experiment 2, the result showed a concentration of 1.97 ng/µL for *BRCA1* products, and 1.96 ng/µL for *GAPDH* products. In experiment 3, the concentration was 16.6 ng/µL. In contrast, experiment 4 showed concentrations of 12.6 ng/µL for *GAPDH* products, 11.9 ng/µL for *BRCA1* products, 10.6 ng/µL for *CHEK2* products and 9.16 ng/µL for *RAD51C* products.

4.6 *In vitro* transcription

In experiment 2, the double stranded cDNA products were *in vitro* transcribed with the MEGAscript™ kit with 4 hours of incubation at 37 °C. Incubation times were 16 hours for experiment 3, and 20 hours for experiment 4. The Tapestation assay results were unable to be reported as the lower marker could not be detected by the reagents. Therefore, the quality of the *in vitro* RNA was not able to be determined.

4.7 MinION library preparation

The *in vitro* transcribed RNA from the previous step was used for the MinION library preparation step **(Figures 2.1, 3.1, 4.1**). The results of the RNAClean™ XP bead system in the DRS kit protocol of each experiment was measured by the Qubit™ RNA and dsDNA HS

Assays. The RNA concentrations of experiments 2-4 were 54.4 ng/µL, 38 ng/µL, and 110 ng/µL respectively. The dsDNA concentrations of experiments 2-4 were 8.9 ng/µL, 2.56 ng/µL and 9.26 ng/µL respectively.

4.8 MinION sequencing results

The ONT DRS kit was used according to the protocol described in **Appendix H**. The isolated *in vitro* mRNA was reverse transcribed as a part of the protocol, and adapters were ligated to the mRNA (**Figures 2.1, 3.1, 4.1**). The sequencing was carried out for 48 hours and the sequenced reads were acquired by MinKNOW v1.14.1. Experiment 2, 3 and 4 generated total FAST5 files which were vastly different (**Table 4.2**). Albacore v2.3.0 was used to basecall these reads and generate FASTQ files. This produced 'pass' and 'fail' folders, for each experiment (**Table 4.2**).

Table 4. 2 Results from basic data acquisition using MinKNOW and Albacore programmes

Counts	Experiment 2	Experiment 3	Experiment 4
Total FAST5	460,846	45,584	1,147,161
Pass reads from basecalling	288,100	11,494	651,277
Fail reads from basecalling	170,435	32,868	383,808

4.8.1 Alignment

Alignment was successfully carried out with Minimap2 by aligning the reads to the Human Genome hg38. The visualisation on IGV showed that in most datasets, 3' bias was present (**Figures 4.11-4.13**). Furthermore, the use of the Minimap2 generated superior alignment compared to other aligners.

4.8.2 Quantification

Each transcript isoform was counted and categorised manually by visualising the sorted BAM files on IGV. The counts of each isoform were based on one type of transcript – NM_007294 for *BRCA1*, NM_001257387 for *CHEK2*, NM_058216 for *RAD51C* and NM_001289743 for *GAPDH*. These genes formed a part of the 4th experiment, and as this experiment produced the best enrichment, only these genes were quantified to compare across experiments. Therefore, TPM values of these genes were calculated by multiplying the number of reads for each isoform with a factor consistent with each experiment (1,000,000/number of pass reads). Counts from experiment 3 were omitted for further analysis as insufficient reads were observed.

External data from an international Nanopore RNA Consortium was used to compare the results from these experiments (**Table 2.4**). The total and subset (Notts_Run1) datasets included 10,302,647 and 363,756 reads respectively, and were aligned, sorted, and indexed via parameters used in experiments 1-4. The sorted BAM file was visualised on IGV and transcript isoforms were counted manually for genes of interest (*BRCA1*, *CHEK2*, *GAPDH* and *RAD51C*).

4.8.2.1 *BRCA1* transcripts

Enrichment for *BRCA1* was evident in experiment 4 compared to the experiment 1 (control experiment) (**Table 4.3**). However, the coverage across all types of isoforms was lacking in depth, where most of *BRCA1* transcripts from experiment 4 appeared to have a 5’ bias, starting from exon 1. In comparison, experiment 1 had a clear 3’ bias (**Table 4.3**). Furthermore, experiment 4 generated longer isoforms compared to experiment 1. Between the enrichment experiments, experiment 2 results presented the widest variety of isoform types compared to experiment 4, possibly due to utilising only one cycle of PCR during the second cDNA synthesis step, where in contrast, experiment 4 utilised three cycles. One full-length alternatively spliced isoform (Δ 9-10,14) was observed in experiment 4 (**Figure 4.9**), and only 2 full-length alternatively spliced isoforms (Δ 5,9 & Δ 9-10) were observed in the Nanopore RNA Consortium data (**Table 4.3**). While experiment 1 (control experiment) showed no isoforms with exon-skipping, experiments 2 and 4 showed isoforms with exon-skipping events (**Tables 4.3-4.4**). Of note, none of the isoform types in experiment 4 were observed in the consortium data. Furthermore, short fragments such as exon 24-only transcripts (**Figure 4.10**) were observed in all experiments, except in experiment 4. Overall, counts and TPM for *BRCA1* were low across all datasets (**Table 4.3-4.4**).

Table 4. 3 Manual counts of *BRCA1* transcripts from all four experiments (NM_007294)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Total Data**
Full length (Δ 5,9)				1
Full length (Δ 9-10)				1
Full length (Δ 9-10,14)			1	
Exon 1-6		2	13	
Exon 1-6 (Δ 5)			1	
Exon 1-7			5	
Exon 1-7(Δ 2-3)			1	
Exon 1-7 (Δ 5)			1	
Exon 1-11		1	1	
Exon 1-11(Δ 9-10)			2	
Exon 2-24				5
Exon 2-24 (Δ 3,9)				1
Exon 2-24 (Δ 9)				1
Exon 2-24 (Δ 9-10)				1
Exon 2-24 (Δ 22)				1
Exon 6-24				1
Exon 11				
Exon 11-24				22
Exon 11-24 (Δ 14)		1		
Exon 11-24 (Δ 16-18,21-23)				1
Exon 13-24		1		1
Exon 13-24 (Δ 14)				3
Exon 14	12			
Exon 14-24 (Δ 16-18)				1
Exon 15-24				4
Exon 16-24				7
Exon 17	1			
Exon 17-24		1		3
Exon 18-24				4
Exon 20-24				2
Exon 22-24				2
Exon 23-24	1			
Exon 24 / UTR	9	3		45

Table 4. 4 Transcripts per million of *BRCA1* transcripts from all four experiments (NM_007294)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Total Data**
Full length (Δ 5,9)				< 1
Full length (Δ 9-10)				< 1
Full length (Δ 9-10,14)			1.54	
Exon 1-6		6.94	19.96	
Exon 1-4 (Δ 5)			1.54	
Exon 1-7			7.68	
Exon 1-7(Δ 2-3)			1.54	
Exon 1-7 (Δ 5)			1.54	
Exon 1-11		3.47	1.54	
Exon 1-11(Δ 9-10)			3.07	
Exon 2-24				0.49
Exon 2-24 (Δ 3,9)				< 1
Exon 2-24 (Δ 9)				< 1
Exon 2-24 (Δ 9-10)				< 1
Exon 2-24 (Δ 22)				< 1
Exon 6-24				< 1
Exon 11				
Exon 11-24				2.14
Exon 11-24 (Δ 14)		3.47		
Exon 11-24 (Δ 16-18,21-23)				< 1
Exon 13-24		3.47		< 1
Exon 13-24 (Δ 14)				< 1
Exon 14	22.23			
Exon 14-24 (Δ 16-18)				< 1
Exon 15-24				< 1
Exon 16-24				< 1
Exon 17	1.85			
Exon 17-24		3.47		< 1
Exon 18-24				< 1
Exon 20-24				< 1
Exon 22-24				< 1
Exon 23-24	1.85			
Exon 24 / UTR	16.67	10.41		4.38

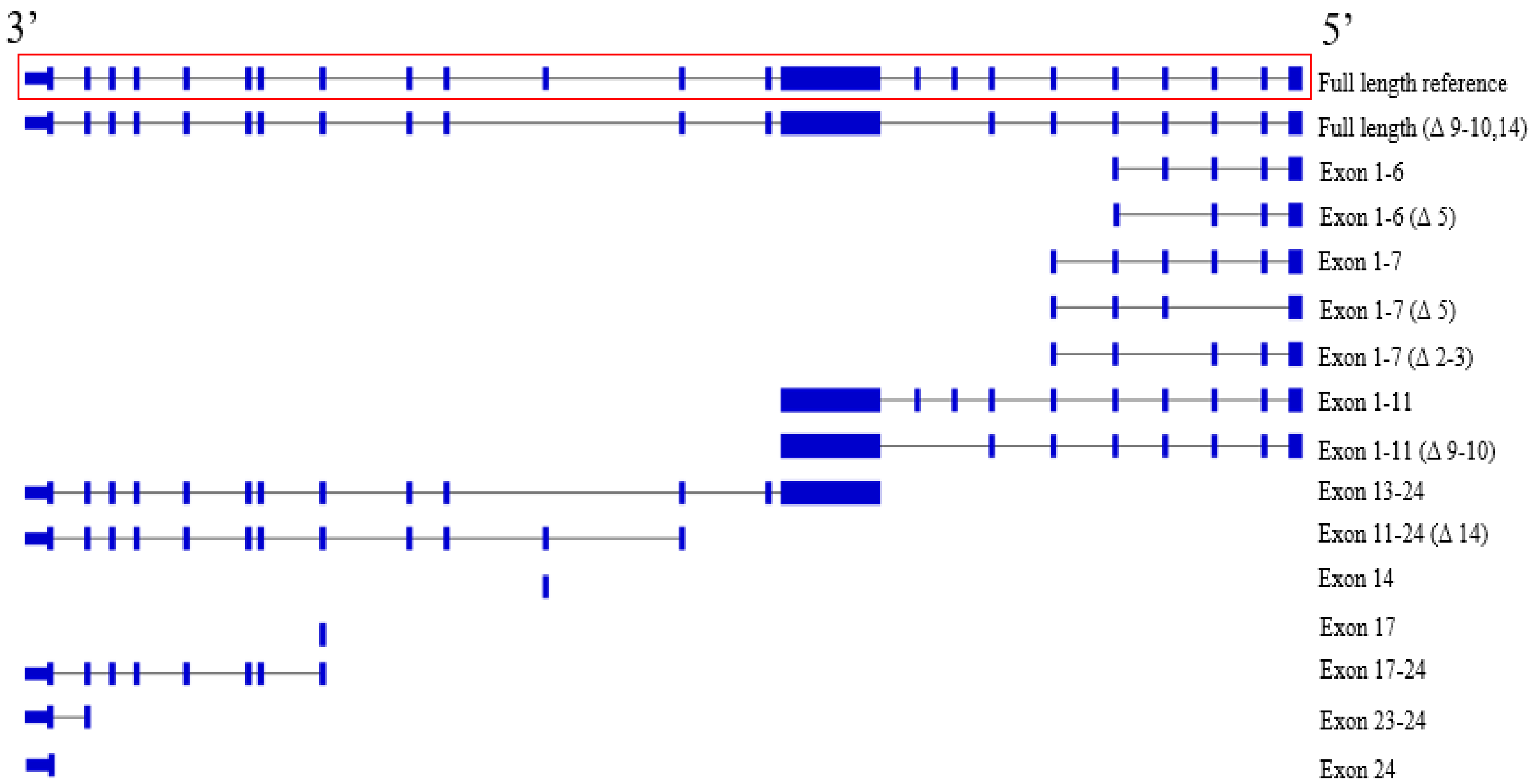


Figure 4. 10 A schematic of observed *BRCA1* isoform types across experiment 1-4

The reference of the *BRCA1* gene is enclosed in a red box. The isoform types observed in the experiments are directly below the full-length reference, with intron-exon information. Exons are indicated in blue blocks, and introns are indicated in black horizontal lines. No clear evidence of a 3’ or 5’ bias was present across experiments 1-4.

4.8.2.2 *CHEK2* transcript counts

Low numbers of *CHEK2* transcripts were observed across all datasets, along with 3' bias. However, there was lack of evidence to conclude a 3' bias for experiment 2. The absolute number of isoform counts was highest in experiment 4 (**Table 4.5**). Only two splice isoforms were found in all four experiments, with two Δ 3 isoforms being found in experiment 2 (**Table 4.5**). However, these transcripts were not full-length isoforms (exon 2-4, 2-16). The overall TPM for most transcript types was similar between experiments, however, the experimental TPMs were greater compared to the consortium dataset (**Table 4.5**). Only 3 types of transcripts (exon 2-16, 13-16 and 14-16) were found in both experiments 1 and 4 (**Table 4.5, Figure 4.11**). This suggests that these datasets may not be comparable due to the low number of overlapping transcript types between experiments.

Table 4. 5 Manual counts of *CHEK2* transcripts from all four experiments (NM_001257387)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Total Data**
Full length				3
Full length (▼2)				2
Exon 2-4	1			3
Exon 2-4 (▼2)				1
Exon 2-4 (∆ 3)		1		
Exon 2-15				1
Exon 2-16				33
Exon 2-16 (▼2)				1
Exon 2-16 (∆ 3)		1		
Exon 2-16 (∆ 6-7,11-15)				1
Exon 3-15				1
Exon 3-16				6
Exon 4-6				1
Exon 4-15				1
Exon 4-16				2
Exon 6-16				2
Exon 7-16		1		2
Exon 8-16	1			
Exon 9-16			3	1
Exon 10-16			2	4
Exon 11-12	1			
Exon 11-16	1			7
Exon 12-16	2		2	5
Exon 13-16	1		2	5
Exon 14-16	2		4	1
Exon 15-16			1	6
Exon 16			2	1

Table 4. 6 Transcripts per million of *CHEK2* transcripts from all four experiments (NM_001257387)

Description	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Total Data**
Full length				< 1
Full length (▼ 2)				< 1
Exon 2-4	1.85			< 1
Exon 2-4 (▼2)				< 1
Exon 2-4 (Δ 3)		3.47		
Exon 2-15				< 1
Exon 2-16				3.20
Exon 2-16 (▼ 2)				< 1
Exon 2-16 (Δ 3)		3.47		
Exon 2-16 (Δ 6-7,11-15)				< 1
Exon 3-15				< 1
Exon 3-16				< 1
Exon 4-6				< 1
Exon 4-15				< 1
Exon 4-16				< 1
Exon 6-16				< 1
Exon 7-16		3.47		< 1
Exon 8-16	1.85			
Exon 9-16			4.61	< 1
Exon 10-16			3.07	< 1
Exon 11-12	1.85			
Exon 11-16	1.85			< 1
Exon 12-16	3.71		3.070	< 1
Exon 13-16	1.85		3.070	< 1
Exon 14-16	3.71		6.14	< 1
Exon 15-16			1.54	< 1
Exon 16			3.07	< 1

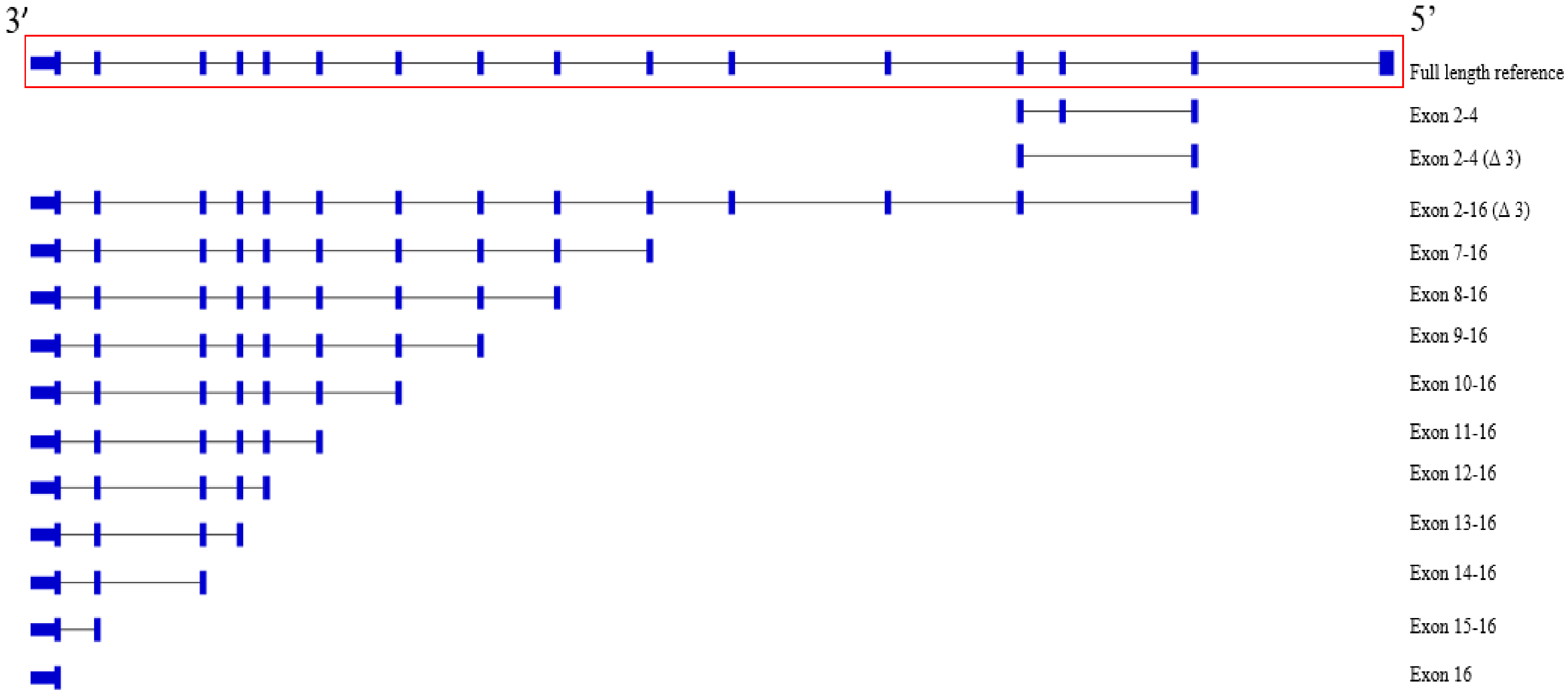


Figure 4. 11 A schematic of observed *CHEK2* isoform types across experiment 1-4

The reference of the *CHEK2* gene is enclosed in a red box. The isoform types observed in the experiments are directly below the full-length reference, with intron-exon information. Exons are indicated in blue blocks, and introns are indicated in black horizontal lines. 3' bias was observed across the four experiments.

4.8.2.3 *GAPDH* transcript counts

3’ bias was observed in all experiments (**Figure 4.12**), with clear enrichment in experiment 2 and 4 compared to experiment 1 (control experiment) (**Table 4.7**). The consortium subset data presented similar full-length transcript count levels as the enrichment experiments, however, all other types of isoforms presented greater counts in experiment 2 and 4 (**Table 4.7**). Spliced isoforms were rare across all experiments, with the most being found in experiment 4 (n=3) compared to experiment 2 (n=2) and experiment 1 (n=1) (**Table 4.7**). Higher TPMs were shown in enrichment experiments compared to experiment 1 (control experiment) (**Table 4.8**).

Table 4. 7 Manual counts of *GAPDH* transcripts from all four experiments (NM_001289745)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Subset Data**
Exon 1-9 (Full length)	36	108	101	92
Exon 1-9 (∆ 2)			1	
Exon 1-9 (∆ 4)			1	
Exon 1-9 (∆ 5)	1			
Exon 1-9 (∆ 7)			1	
Exon 2-9	91	241	197	
Exon 2-9 (∆ 5)		1		
Exon 3-9	48	199	240	109
Exon 4-9	28	344	193	74
Exon 5-9	29	189	141	34
Exon 5-9 (∆ 6)		1		
Exon 6-9	59	304	225	94
Exon 7-9	67	241	202	51
Exon 8-9	292	875	734	232
Exon 9	19	139	98	25

Table 4. 8 Transcripts per million of *GAPDH* transcripts from all four experiments (NM_001289745)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Subset Data**
Exon 1-9 (Full length)	66.69	374.87	155.08	252.92
Exon 1-9 (Δ 2)			1.54	
Exon 1-9 (Δ 4)			1.54	
Exon 1-9 (Δ 5)	1.85			
Exon 1-9 (Δ 7)			1.54	
Exon 2-9	168.59	836.52	302.48	
Exon 2-9 (Δ 5)		3.47		
Exon 3-9	88.93	690.73	368.51	299.65
Exon 4-9	51.87	1194.03	296.34	203.43
Exon 5-9	53.73	656.02	216.50	
Exon 5-9 (Δ 6)		3.47		
Exon 6-9	109.30	1055.19	345.48	258.41
Exon 7-9	124.13	836.52	310.16	140.20
Exon 8-9	540.97	3037.14	1127.02	637.79
Exon 9	35.20	482.47	150.47	68.73

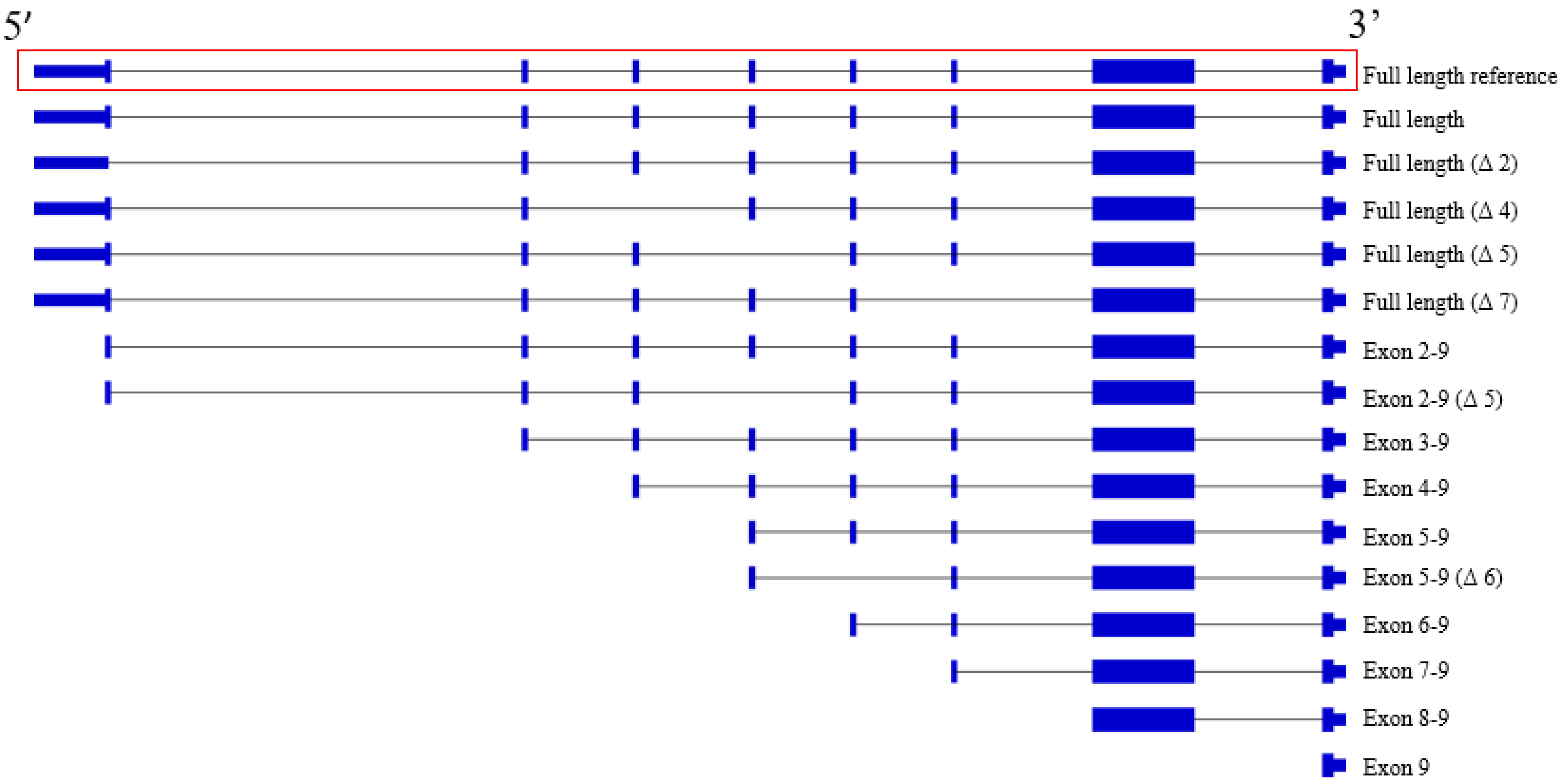


Figure 4. 12 A schematic of observed *GAPDH* isoform types across experiment 1-4

The reference of the *GAPDH* gene is enclosed in a red box. The isoforms types observed in the experiments are directly below the full-length reference, with intron-exon information. Intronic insertions are indicated in red blocks. 3' bias was observed across the four experiments.

4.8.2.4 *RAD51C* transcript counts

Full-length isoforms were observed for *RAD51C* in all experiments (**Table 4.9, Figure 4.13**). Experiment 1 showed 3' bias, whereas bias information was unable to be obtained with experiment 2 due to the low number of transcripts. In contrast, experiment 4 presented even coverage across most types of isoforms, and therefore 3' bias was less evident. This trend was also observed in the consortium data (**Table 4.9**). The absolute counts showed clear enrichment in experiment 4, when compared to experiment 1 (control experiment). Exon-skipping events were observed only for experiment 4, and the intronic insertions were only observed in experiment 1 (**Table 4.9**). All transcript types present in both experiment 1 and 4 show that there were higher TPMs in experiment 4 compared to experiment 1 across all transcripts. The TPM for full-length isoforms with no exon-skipping events were observed to be the highest in the Nanopore RNA Consortium data (**Table 4.10**).

Table 4. 9 Manual counts of *RAD51C* transcripts from all four experiments (NM_058216)

Isoform Type	**Exp 1 Counts**	**Exp 2 Counts**	**Exp 4 Counts**	**Consortium Total Data**
Exon 1-9 (full length)	2	2	6	126
Exon 1-9 (Δ 3)				1
Exon 1-9 (Δ 3, 5-6)			1	
Exon 1-9 (Δ 4)			1	1
Exon 1-9 (▼4)	1			
Exon 1-9 (Δ 4, 7)			2	
Exon 1-9 (Δ 6)			1	
Exon 1-9 (Δ 6-7)				1
Exon 1-9 (Δ 7)			4	4
Exon 1-9 (Δ 7-8)				2
Exon 1-2	1	1	96	
Exon 1-6				
Exon 2			118	2
Exon 2-9	1		28	20
Exon 2-9 (Δ 3)			1	1
Exon 2-9 (Δ 3-4)				1
Exon 2-9 (Δ 3,7)			1	
Exon 2-9 (Δ 3-4,7)				1
Exon 2-9 (Δ 4)			3	1
Exon 2-9 (Δ 7)			7	1
Exon 2-9 (Δ 7-8)	1			
Exon 2-3			1	
Exon 2-4			11	
Exon 2-5			1	
Exon 2-6			1	
Exon 3-9			13	15
Exon 3-8	1			
Exon 4-9	4	1	21	8
Exon 4-9 (▼4)	1			
Exon 4-9 (Δ 7)			1	
Exon 5-9			3	6
Exon 5-9 (Δ 7)			2	
Exon 5-9 (Δ 7-8)			1	
Exon 5-6			1	
Exon 6-9				2
Exon 6-9 (Δ 7)			1	
Exon 7-9			2	3
Exon 7-9 (▼7)	1			
Exon 8-9	1		6	4
Exon 9	3		7	5

Table 4. 10 Transcripts per million of *RAD51C* transcripts from all four experiments (NM_058216)

Isoform Type	Exp 1 Counts	Exp 2 Counts	Exp 4 Counts	Consortium Total Data
Exon 1-9 (full length)	3.71	6.94	9.21	12.23
Exon 1-9 (Δ 3)				< 1
Exon 1-9 (Δ 3, 5-6)			1.54	
Exon 1-9 (Δ 4)			1.54	< 1
Exon 1-9 (▼4)	1.85			
Exon 1-9 (Δ 4, 7)			3.07	
Exon 1-9 (Δ 6)			1.54	
Exon 1-9 (Δ 6-7)				< 1
Exon 1-9 (Δ 7)			6.14	< 1
Exon 1-9 (Δ 7-8)				< 1
Exon 1-2	1.85	3.47	147.40	
Exon 1-6				
Exon 2			181.18	< 1
Exon 2-9	1.85		42.99	1.94
Exon 2-9 (Δ 3)			1.54	< 1
Exon 2-9 (Δ 3-4)				< 1
Exon 2-9 (Δ 3-4,7)				< 1
Exon 2-9 (Δ 3,7)			1.54	
Exon 2-9 (Δ 4)			4.61	< 1
Exon 2-9 (Δ 7)			10.75	< 1
Exon 2-9 (Δ 7-8)	1.85			
Exon 2-3			1.54	
Exon 2-4			16.89	
Exon 2-5			1.54	
Exon 2-6			1.54	
Exon 3-9			19.96	1.46
Exon 3-8	1.85			
Exon 4-9	7.41	3.47	32.24	< 1
Exon 4-9 (▼4)	1.85			
Exon 4-9 (Δ 7)			1.54	
Exon 5-9			4.61	< 1
Exon 5-9 (Δ 7)			3.07	
Exon 5-9 (Δ 7-8)			1.54	
Exon 5-6			1.54	
Exon 6-9				< 1
Exon 6-9 (Δ 7)			1.54	
Exon 7-9			3.07	< 1
Exon 7-9 (▼7)	1.85			
Exon 8-9	1.85		9.21	< 1
Exon 9	5.56		10.75	< 1

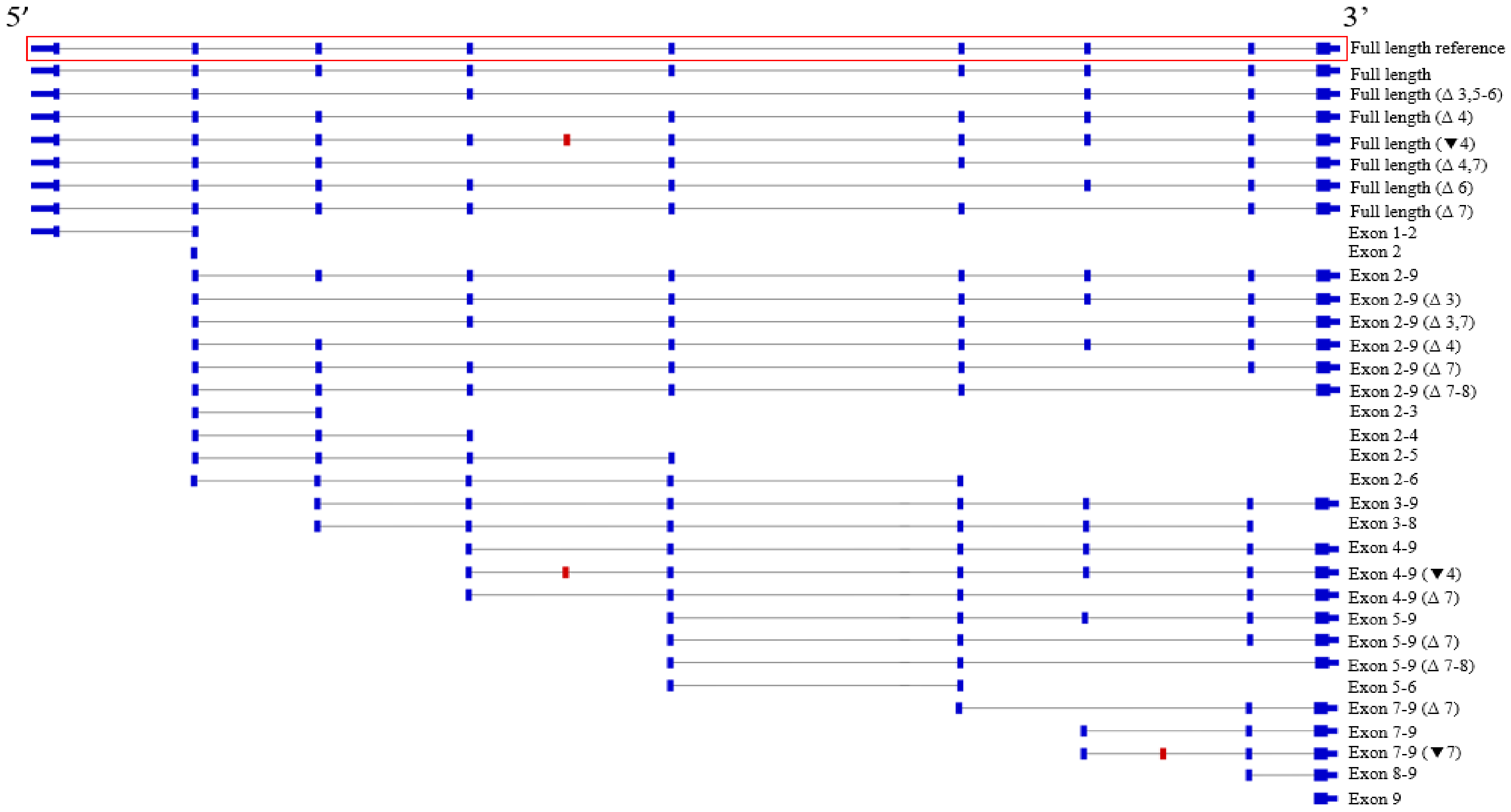


Figure 4. 13 A schematic of observed *RAD51C* isoform types across experiment 1-4

The reference of the *RAD51C* gene is enclosed in a red box. The isoform types observed in the experiments are directly below the full-length reference, with intron-exon information. Exons are indicated in blue blocks, and introns are indicated in black horizontal lines. Intronic insertions are indicated in red blocks. 3’ bias was observed across the four experiments.

4.8.2.5 TPM isoform ratios

To determine whether the enrichment experiments 2 and 4 were quantitative, the ratios of three *GAPDH* isoform types to full-length TPMs were examined across different experiments (**Table 4.11**). Results revealed that these ratios were similar between experiments 2 and 4 (enrichment experiments) and experiment 1 (control experiment), suggesting linear enrichment.

Table 4. 11 *GAPDH* isoform ratios of TPM

Ratio of TPM	**Exp 1 (Control)**	**Exp 2 (Enrichment)**	**Exp 4 (Enrichment)**
Exon 6-9 / Full length	1.64	2.81	2.23
Exon 7-9 / Full length	1.86	2.23	2.00
Exon 8-9 / Full length	8.11	8.10	7.27

The ratios between different *RAD51C* isoform types across different experiments revealed dissimilar ratios between control and enrichment experiments (**Table 4.12**). The ratios between experiments 1-2 (control experiments) were similar; however, experiment 4 presented greater ratios. This was particularly shown in exon 1-2/full length ratios in experiment 4, which were 32-fold greater than experiments 1-2 (control experiments). However, it is important to note that the accuracy of this comparison may be influenced by the very low number of reads in experiment 1 (control experiment) compared to experiment 4 (enrichment experiment).

Table 4. 12 *RAD51C* isoform ratios of TPM

Ratio of TPM	Exp 1 (Control)	Exp 2 (Control)*	Exp 4 (Enrichment)
Exon 4-9 / Full length	2	0.5	3.5
Exon 1-2 / Full length	0.5	0.5	16

*In experiment 2, only *BRCA1* and *GAPDH* were enriched, so that *RAD51C* was not targeted for enrichment. Whereas in experiment 4, *RAD51C* was targeted for enrichment. Therefore, for the purposes of comparing whether experiment 4 enriched for *RAD51C*, the results of experiment 2 was regarded as a control experiment along with experiment 1 only for this exercise.

4.8.2.6. Comparison of alternative splicing events to other datasets

All *BRCA1* single-exon alternative splicing events (Δ 5, Δ 14, Δ 22) were detected in both the DRS data and the MCRG data. Furthermore, they have all been previously described in the literature (**Table 4.13**). In contrast, the exon-skipping events which occurred across longer distances such as Δ 3, 9 were only detected in long-read DRS datasets. Other alternative splicing events, such as Δ 5, 9, Δ 16-18 and Δ 16-18, 21-23 were only observed in the Nanopore RNA Consortium data, and were not previously described in other external datasets which utilised RNA-seq techniques (Colombo et al., 2014; Lattimore et al., 2018).

Table 4. 13 Comparison of *BRCA1* alternative splicing events (NM_007294) with external datasets

Isoform Type	**Exp1 Counts**	**Exp2 Counts**	**Exp4 Counts**	**Consortium Total Counts**	**MCRG RNA-seq data***	**Previously described**
Δ 2-3			1			Colombo et al. (2014)
Δ 3, 9				1		de Jong et al. (2017)
Δ 5			2		8	Colombo et al. (2014) Lattimore et al. (2018)
Δ 5, 9				1		-
Δ 9				1		Colombo et al. (2014) Lattimore et al. (2018)
Δ 9-10			2	2	8	Colombo et al. (2014)
Δ 9-10, 14			1			de Jong et al. (2017)
Δ 14		1		3	2	Colombo et al. (2014)
Δ 16-18				1		-
Δ 16-18, 21-23				1		-
Δ 22				1	5	Colombo et al. (2014) Lattimore et al. (2018)

*Mackenzie Cancer Research Group, University of Otago, Christchurch

Only a small number of *CHEK2* alternative splicing events were observed across all datasets (**Table 4.14**). Δ 3 was observed in experiment 2 and was previously described (Davy et al., 2017). The Nanopore RNA Consortium dataset consisted of multiple ▼ 2 insertions.

Table 4. 14 Comparison of *CHEK2* alternative splicing events (NM_001257387) with external data

Isoform Type	**Exp1 Counts**	**Exp2 Counts**	**Exp4 Counts**	**Consortium Total Counts**	**MCRG RNA-seq data**	**Previously described**
▼2				4		-
Δ3		2				Davy et al (2017)
Δ 6-7,11-15				1		-

GAPDH mRNA splicing events Δ 2 and Δ 7 were both found in experiment 4 and MCRG RNA-seq data (**Table 4.15**). In contrast, Δ 4, Δ 5 and Δ 6 were only observed in experiments 1-4. None of these isoforms have been described in the literature to my knowledge. Exon-skipping co-occurrence was not observed in any single transcript, and alternative splicing events were not found in the Nanopore RNA Consortium subset data.

Table 4. 15 Comparison of *GAPDH* alternative splicing events (NM_001289745) with external data

Isoform Type	**Exp1 Counts**	**Exp2 Counts**	**Exp4 Counts**	**Consortium Subset Counts**	**MCRG RNA-seq counts**	**Previously described**
Δ2			1		33	-
Δ4			1			-
Δ5	1	2				-
Δ6		1				-
Δ7			1		8	-

Δ 3, Δ 4, Δ 6, Δ 7, and Δ 7-8 were observed in the *RAD51C* DRS enrichment data (**Table 4.16**) and MCRG RNA-seq data (unpublished), where they have all been previously described by Davy et al. (2017). However, alternative splicing events (Δ 3,5-6, Δ 3,7 and Δ 4,7) were only observed in experiment 4 and were not previously described (**Table 4.16**). Furthermore, insertion events were only found in experiment 1.

Table 4. 16 Comparison of *RAD51C* alternative splicing events (NM_058216) with external data

Isoform Type	**Exp1 Counts**	**Exp2 Counts**	**Exp4 Counts**	**Consortium Total Counts**	**MCRG RNA-seq data**	**Previously described**
Δ3			1	2	65	Davy et al. (2017)
Δ3-4				1		-
Δ3-4,7				1		-
Δ3,5-6			1			-
Δ3,7			1			-
Δ4			4		16	Davy et al. (2017)
▼4	2					-
Δ4,7			2			-
Δ6			1		7	Davy et al. (2017)
Δ6-7				1		Davy et al. (2017)
▼6	1					-
Δ7			15	5	50	Davy et al. (2017)
Δ7-8			1	2	5	Davy et al. (2017)

4.9 Discussion

4.9.1 Overview

Experiment 2 was the first attempt at enrichment using the T7 RNA polymerase binding site attached to a gene specific primer method, which targeted two genes (*BRCA1* and *GAPDH*). In experiment 2, the results on IGV showed considerable amplification of *GAPDH*. However, there was no clear evidence of clear amplification, nor full-length transcripts for *BRCA1*.

Experiment 3 was used as an attempt to enrich transcripts from 12 hereditary breast and ovarian cancer genes (**Table 1.1**), but experiment 3 results revealed less than the expected number of reads compared to all other experiments, and was omitted from the rest of the study. Reasons for this poor outcome may be due to the pooling of T7-GSP primers in one reaction during the second cDNA strand synthesis step, and utilising only one PCR cycle in the same step.

Experiment 4 consisted of a smaller panel of genes compared to experiment 3, which included *BRCA1*, *GAPDH*, *CHEK2* and *RAD51C*. This experiment proved to be the most successful enrichment experiment of all attempted experiments due to many factors. In experiment 4, the novel methodological technique conducted in this study gave rise to considerable enrichment of *GAPDH* and *RAD51C*, and less obvious amplification for *CHEK2* and *BRCA1*. Importantly, *GAPDH* results revealed evidence of linear enrichment, through calculating TPM ratios between experiments (**Table 4.11**).

Bioinformatics proved to be difficult, especially with quantifying long DRS reads. However, via manual counting, it appeared that the MinION DRS data was able to detect multiple exon-skipping events in the same transcript across long distances for *RAD51C*, which have not been previously described in literature (**Table 4.16)**. The results of these experiments will be fully discussed in chapter 5.

4.9.2 RNA source/cell culturing

LCLs are immortalized cells which have been derived from peripheral blood lymphocytes that have been infected by an Epstein-Bar virus (Neitzel, 1986). They have been routinely used as a continuous source of DNA, RNA and proteins, and is known to express *BRCA1* in relatively high quantities (https://gtexportal.org/home/gene/BRCA1). The particular control LCL (**Table 2.1**) was chosen to be implemented in this study due to previous work being done in this laboratory using the same cell line (de Jong et al., 2017). RNA was extracted from these LCL with an RNeasy® Mini Kit, and was checked for quality. Tapestation RINe scores are measures of RNA quality, which are derived from algorithms determining the quantitative measurement of RNA degradation (Agilent Technologies, 2016). The extracted total RNA showed RINe scores of above 9.3 across all extractions, which proved that the quality of starting RNA was satisfactory for downstream usage. Poly-A RNA (mRNA) was further isolated from total RNA in Experiment 1, which revealed a total amount of 151.2 ng of mRNA. Typically, mRNA contributes to 1-5% of the total eukaryotic RNA concentration (Rosenow, Saxena, Durst, & Gingeras, 2001). Compared to the total RNA input amount, the mRNA obtained from this mRNA isolation step was not satisfactory, therefore, the mRNA isolation step was omitted in successive experiments.

NMD inhibition was carried out for 4 hours before the process of RNA extraction in all experiments. The NMD inhibitor was used to prevent degradation of transcripts with premature protein truncations, which would normally be degraded by the NMD surveillance mechanism. Gene expression levels are dependent on many factors, including environmental stress (Murray et al., 2004). It has been suggested that using an NMD inhibitor may change the expression level of certain transcripts by overloading the endoplasmic reticulum with truncated misfolded proteins, which would cause an increase in cellular stress, although this remains to be proven (Hug, Longman, & Caceres, 2016). Furthermore, it is worth noting that expression data generated from cells exposed to NMD inhibitors may not be comparable to expression data from breast cancer tissues.

4.9.3 First strand cDNA synthesis

4.9.3.1 Standard long-range PCR vs alternative long-range PCR

Long-range PCR was carried out to confirm the presence of full-length *BRCA1* isoforms, as it was the primary gene of interest in the study. However, this was not carried out for other genes of interest, due to the availability of time and resources. In the results of the standard long-range PCR, one agarose gel showed presence of a prominent band around 2.4 KB (**Figure 4.2 A**), which was also observed in **Fig. 4.2 B.** These bands may be attributed to full-length isoforms which lack exon 11, the largest exon in the gene. Exon 11 is 3.4 KB long and contains the RAD51-interaction domain, which is involved in the double-stranded DNA break repair mechanism (Tammaro et al., 2012). The Δ 11 isoform has been previously described in the literature, where mouse embryonic fibroblast cells which were homozygous for the Δ 11 isoform showed a deficient G2-M checkpoint (X. Xu et al., 1999). Full-length *BRCA1* products without skipping were only observed in **Fig. 4.2 B**, and was represented by a faint, thin band of around 6 KB. This suggests that full-length isoforms were present in the sample, and the first cDNA strand synthesis proved to be successful. Given that full-length *BRCA1* isoforms were present in the sample, it can be inferred that other genes of interest with higher expression levels and shorter transcript lengths were also present in our sample, such as *RAD51C* and *GAPDH* (**Table 4.1**). The inconsistency between the two standard long-range PCR and the other failed 11 PCR results may have been generated due to the inconsistencies in aliquoting the *BRCA1* cDNA and different rates of degradation. In contrast, the results of the alternative PCR revealed improved outcomes, with full-length products of around 5.8 KB and what is assumed to be the Δ11 isoform of *BRCA1* with a length of approximately 2.4 KB (Raponi et al., 2014) (**Figure 4.3**). These bands were more prominent and brighter compared to the standard long-range PCR counterparts. This suggests that the alternative long-range PCR method incorporated better PCR cycling conditions, or improved ratios of reagents compared to the standard long-range PCR. The alternative long-range PCR involved a higher $MgCl_2$ and cDNA concentrations, and lower concentration of dNTP mix, with shorter denaturation and extension times compared to the standard long-range PCR. Higher $MgCl_2$ concentrations have been associated with a greater yield, but also a decrease in specificity (Lorenz, 2012). Furthermore, higher dNTP concentrations can inhibit the PCR reaction, which may explain the inferior results of the standard long-range PCR (Lorenz, 2012). As only one alternative long-range PCR was carried out, these comments are inconclusive, hence more reactions should be carried out in the future to validate this claim. The amount of input RNA was maintained at the maximum possible level of 5 µg per reaction to standardise our experiments during the first strand synthesis step.

4.9.4 Second strand cDNA synthesis

4.9.4.1 Temperature gradient PCR

The temperature gradient PCR was carried out in order to find the optimum annealing temperature for the second cDNA synthesis. The resulting gel revealed bright and prominent bands between 49.8 °C and 62.0 °C for *BRCA1* (**Figure 4.4**), and between 49.8 °C and 65.1 °C for *GAPDH* (**Figure 4.5**). The lowest integer temperature of the brightest bands was used in all second cDNA synthesis experiments, which was 50 °C for both *BRCA1* and *GAPDH*. If the annealing temperature is high, there is a chance of non-specific annealing, and if inadequate, it may cause decrease in yield (Rychlik, Spencer, & Rhoads, 1990). Ideally, this temperature gradient experiment should have been carried out for all genes studied in this project, however, lack of time and resources prevented this from occurring. Therefore, this may have limited the outcomes of the second cDNA strand synthesis step. One reason which could explain the poor outcomes of experiment 3, may be the use of one annealing temperature for many different T7-GSP primers, all with differing Tm values (**Table 2.2).** This also implies that experiment 4 may be improved by using optimal annealing temperatures in the future.

4.9.4.2 T7-only primer PCR

The T7-only primer PCR was carried out with T7-only forward and exon 7 reverse primers for *BRCA1* products, to assess the incorporation of the T7 RNA polymerase binding site at the 5’ of the second cDNA strand. The gel proved that the incorporation had taken place successfully as a clear band of the correct approximate size was visualised (**Figure 4.7**).

4.9.4.3 T7-oligo dT$_{(20)}$ (full-length) PCR

The purpose of this PCR was to determine if full-length transcription had taken place in the second cDNA strand synthesis step in experiment 4. It is clear that *GAPDH*, *RAD51C* and *CHEK2* showed full-length double stranded cDNA molecules in the sample, with approximate sizes of 1.2 KB to 1.9 KB (**Figure 4**.**9**). However, it was unclear whether there were full-length *BRCA1* double stranded cDNA molecules in the sample, as the gel did not show clear bands around 5.8 KB or greater. These results correspond to the relative expression levels of these genes (**Table 4.1**). *RAD51C* and *CHEK2* products presented numerous faint but distinct bands above 1.3 KB, which indicate that many types of isoforms of *RAD51C* and CHEK2 were present in the sample.

4.9.5 HighPrep™ magnetic bead clean-up

In experiment 2, Qubit™ dsDNA HS assays results revealed low and similar dsDNA concentrations for both *BRCA1* and *GAPDH* samples of 1.97 ng/µL and 1.96 ng/µL, respectively. This was an unexpected result, as *GAPDH* has a higher expression compared to *BRCA1* (**Table 4.1**). As the optimum annealing temperatures have been determined for both these genes (**Figures 4.4-4.5**), it appears that these results were not a consequence of suboptimal annealing temperatures during the second cDNA synthesis step. One possible explanation for these results, may be due to the inaccurate sampling of a small volume (1 µL) using a pipette during the Qubit™ assay step.

In experiment 3, an attempt was made to enrich a panel of 12 different hereditary breast and ovarian cancer genes. The T7-GSP for all 12 genes were pooled and the reaction was carried out in a single tube. The final Qubit™ dsDNA HS assay results showed a higher concentration (16.6 ng/µL) compared to the second experiment. However, assuming that each gene in the panel equally contributed to these results, the average dsDNA concentration of each gene would be 1.38 ng/µL, which is a poor outcome compared to the results of experiment 2. Although it is unclear why this experiment failed, it is possible the pooling step lead to suboptimal outcomes. For that reason, the pooling approach was not used in experiment 4.

In experiment 4, four different reaction volumes were carried separately through the second cDNA strand synthesis step, with one reaction for each gene of interest; *GAPDH*, *BRCA1*, *RAD51C*, *CHEK2*. *RAD51C* and *CHEK2* are hereditary breast and ovarian cancer genes with similar transcript lengths to *GAPDH* (**Table 4.1**). Through this experiment, the effect of gene expression on the efficacy of the novel enrichment method was tested. Three cycles of PCR were implemented only in this experiment in the second strand synthesis step, to increase the effect of enrichment. This increase in the number of cycles was expected to maintain linear amplification due to the use of unidirectional PCR. Furthermore, the Qubit™ dsDNA HS assay results exhibited approximately 6-fold increase of dsDNA concentration of *BRCA1* and *GAPDH* in this experiment compared to the results of experiment 2. The results provided evidence to suggest that the increase in the number of PCR cycles within the second strand synthesis step contributed to a higher yield. Interestingly, the dsDNA concentrations of all genes in this panel were similar, which reflects the results in experiment 2. Therefore, repeating these experiments may provide a clearer assessment of these results. Furthermore, there is a possibility that insufficient magnetic bead elution caused some loss of product, which would also confound the results.

4.9.6 *In vitro* transcription & RNAClean™ XP magnetic bead clean-up

The incubation times during *in vitro* transcription for the 2nd and 3rd experiments were 4 hours and 16 hours respectively, as the manufacturer's guidelines indicated that the incubation step should be carried out around 4-16 hours. Therefore, the two extremes of incubation time were tested in these experiments. However, as experiment 3 generated insufficient data to compare with experiment 2, the results were unable to be compared.

To determine whether further incubation time maximised the probability of detecting longer transcripts, 20 hours of incubation was trialled in the *in vitro* transcription step. It was evident that experiment 4 produced the most transcripts among all four experiments, and therefore this step may play a critical role within the enrichment process.

As with the HighPrep™ magnetic bead clean-up step, the RNAClean™ XP magnetic bead clean-up process increased the chances of losing *in vitro* transcribed RNA via insufficient eluting of beads. The results of the Qubit™ RNA HS assay revealed the greatest concentrations of RNA in experiment 4, and the lowest concentration in experiment 3. Therefore, the concentration of the *in vitro* RNA was not directly proportional to the incubation time, which may suggest that further testing needs to be carried out to validate this concept with experiments which only vary in the incubation time. In all experiments, the concentration of the input RNA levels was lower compared to the recommended amount of RNA (500 ng) described in the ONT protocol (**Appendix H**). This may have contributed some artifacts in the sequencing step. The dsDNA concentrations within the library preparation step followed the same trend, with the highest levels of dsDNA concentrations presented in experiment 4. The recommended recovery is up to 200 ng (**Appendix H**), which was achieved in all experiments.

Chapter 5: Discussion

5.1 Introduction

Initially, the main purpose of this project was to develop and test a novel method of linearly enriching *BRCA1* transcripts, and to make use of the new opportunities afforded by direct RNA sequencing on the MinION nanopore sequencer. The original motive for this enrichment was to enable the study of effects of VUS on alternative splicing patterns of *BRCA1*. Over the course of the project, the scope expanded to a wider range of genes, which included *GAPDH* and other associated hereditary breast and ovarian cancer genes **(Table 1.1).** The two major components of the project consisted of laboratory experiments and bioinformatic analyses.

5.2 Rationale

The first experiment was a proof-of-principle experiment whereby the efficacy of the MinION DRS method was tested. This experiment was set as the reference experiment for subsequent experiments. Experiments 2-4 were carried out to test a novel linear amplification method for targeted RNA transcripts. In the second experiment, *BRCA1* was the gene of interest, and *GAPDH* was used as a control gene. Building on the results of the second experiment, the third experiment involved testing a panel of hereditary breast and ovarian cancer genes, with a range of different expression levels and transcript lengths, to test the limits of the novel linear amplification method. The fourth experiment involved a panel of four genes; *GAPDH*, *RAD51C*, *BRCA1*, and *CHEK2*, with the latter three with biological relevance to breast cancer and wide range of expression levels. *RAD51C* and *CHEK2* have similar transcript lengths and different expression levels to *GAPDH* (**Table 4.1**) and were utilised to evaluate the effect of gene expression on this novel enrichment method. Below I discuss each key component of this project.

5.3. Direct RNA Sequencing (DRS)

As explained previously, DRS is a novel approach and its efficacy has not been validated as much as traditional RNA-seq. The first hypothesis of this project was that DRS was able to be carried out in a proof-of-principle experiment (experiment 1).

The results from this experiment revealed that DRS was successful with the RNA that was extracted in this laboratory (**Figures 3.2-3.9**). Therefore, it may be concluded that the relatively novel DRS kit is able to be carried out with relative ease. The outcome of 889,850 sequenced RNA reads in experiment 1 support this, as this is a comparable number to the Nanopore RNA Consortium data (**Table 2.4**).

5.4. Novel enrichment method

Through carrying out experiments 2-4, a novel enrichment approach was developed and tested. The enrichment method involved RNA extraction, first cDNA strand synthesis, second cDNA strand synthesis, *in vitro* transcription and MinION sequencing (**Figures 3.1, 4.1**). Experiment 2 and 4 proved to show positive outcomes of enrichment, whereas experiment 3 did not produce outcomes that were able to be compared with other datasets (**Tables 4.3-4.10**).

GAPDH and *RAD51C,* which have shorter transcript lengths and greater expression levels compared to *BRCA1* and *CHEK2* (**Table 4.1**), were able to be clearly enriched through this novel enrichment method. The specific details of these results are outlined below in sections 5.6-5.10.

5.5. Bioinformatic analyses

Bioinformatic analyses proved to be challenging due to the lack of availability of software specific for long-read DRS. This is almost certainly due to the novelty of this platform, and it is likely that more appropriate software will be rapidly developed due to the wide uptake of nanopore sequencing. In this thesis, software which was designed for short RNA-seq data or long RNA-seq data was trialled with the data generated from our DRS nanopore sequencing experiment.

Data acquisition with MinKNOW and Albacore was carried out with ease, as the software was provided by ONT. Basecalling with Albacore produced ‘pass’ and ‘fail’ reads, in which the number of ‘pass’ reads were greater than ‘fail’ reads in all experiments except experiment 3. Pass reads have a quality score of > Q7, and this result suggests that the quality of RNA in experiment 3 was suboptimal.

Databases and indices were successfully created with all four tested alignment programmes (Minimap2, BLAT, GMAP, LAST), however, alignment was only successfully carried out with Minimap2. Four quantification programmes (Kallisto, Salmon, Mandalorion and SQANTI) were tested, and the results revealed that the parameters used within these programmes were unable to quantify long DRS reads produced by the MinION (**Tables 3.9-3.14**).

The final pipeline used for analysing the reads generated from experiments 1-4 involved aligning the reads with Minimap2 with the given ONT DRS parameters, sorting and indexing with Samtools, and visualizing the file on IGV. The reads were manually counted and categorized. The specific details of the analyses are outlined below in sections 5.6-5.10.

5.6 *BRCA1* transcripts

BRCA1 was the initial focus of the thesis, and as expected, *BRCA1* was very poorly represented (**Tables 4.3-4.4**). Interestingly, many short transcripts located in the middle of the gene were observed in experiment 1 (**Figure 3.2**). Upon closer examination of these short mid-gene located reads, poly-A regions were found in the corresponding intronic region of *BRCA1*. This may have caused spurious alignment of gDNA contaminants or creation of truncated cDNA during first strand synthesis by the oligo-dT primer binding to these intronic poly-A region (length = 16 bps) (Nam et al., 2002). This has been additionally confirmed by studying these reads with the online UCSC Genome Browser BLAT tool. The short regions appeared to align to various genes, not only for *BRCA1*. Furthermore, this region was not recognised to be a retroviral element or retrotransposon by the UCSC Genome Browser repeat tracks.

There was lack of evidence to determine whether there was any 5' or 3' bias in experiments 1 and 2. In the second experiment, while reads from the 5' side of the RNA were present, other reads started from the middle of the gene in exons 11, 13 and 17 (**Table 4.3, Figure 4.10**). This phenomenon may also be explained by non-specific binding of the T7-GSP to the middle of the gene instead of the 5' end, or the premature interruption of sequencing. *BRCA1* appeared to be enriched in the 4th experiment, with a 5' bias, which was an unexpected result. The truncated 3' ends of these reads may indicate that the LongAmp® *Taq* polymerase may have not been able to fully transcribe to the 3'end of the gene, or the T7 RNA polymerase may have prematurely detached from the double stranded cDNA. However, these explanations do not justify the fact that reads without poly-A tails have been able to be sequenced with nanopore sequencing, which requires 3' poly-A tails for the attachment of adapters. This suggests that perhaps the nanopores are non-specific and are prone to drawing in molecules without adapters. Otherwise, the transcripts could have fused to the adapters spontaneously during sequencing.

The fourth experiment was also able to produce a full-length alternatively spliced *BRCA1* isoform (Δ 9-10, 14), whereas only two full-length alternatively spliced isoforms were found in the consortium dataset (Δ 5,9, Δ 9-10) (**Table 4.3**). Considering that the consortium dataset consisted of around 13 million reads from 30 MinION flow-cells which greatly exceeded the number of reads from the Experiment 4, this enrichment technique appears to be promising. Furthermore, longer transcripts were found in Experiment 4 compared to Experiments 1-2. This provides evidence that the protocol for the fourth experiment may be a good starting point for further development.

Exon-skipping events were only observed in the enrichment experiments (**Table 4.13**), which suggests that the enrichment method may enable a wider range of transcripts to be observed. However, this may also suggest that the increased number of manipulations have caused spurious alignment and introduced errors in the data. In other respects, as the Nanopore RNA Consortium dataset also contained *BRCA1* transcripts with exon-skipping events, this may suggest that in fact, the lack of data in the first experiment prevented the observation of these events. Furthermore, the isoform types observed in experiment 4 were not observed in the consortium data. This may be attributed to the fact that most of the 5' reads in experiment 4 were truncated (**Figure 4.10, Table 4.3**), whereas the Nanopore RNA Consortium dataset did not contain 3' truncations in general. Therefore, these differences may have occurred from using numerous manipulations in experiment 4, as opposed to a few manipulations used to generate the Nanopore RNA Consortium dataset.

5.7 *CHEK2* transcripts

CHEK2 is recognised as a gene with an even lower expression level compared to *BRCA1,* which is reflected by the low number of counts in **Table 4.5**. 3' bias was observed, which may be explained by the use of an oligo-dT primer, which may have contributed to the truncated reads. *CHEK2* transcripts in experiment 4 did not show a 5' bias such as with *BRCA1* transcripts, which may be attributed to the shorter transcript length of *CHEK2* compared to *BRCA1* transcripts. It can be hypothesised that shorter transcripts provide less opportunities for error and reduce the likelihood of fragmentation prior or during sequencing. It was difficult to observe a 5' or 3' bias with experiment 2, due to the lack of data (**Table 4.5**). Overall, low number of alternative splicing events were observed across all datasets. Furthermore, as the variety of transcript types were different between datasets (experiments 1-4), the datasets were not comparable. Therefore, it was difficult to determine the efficacy of the enrichment method with *CHEK2*. Moreover, a considerable number of *CHEK2* transcripts started from exon 2, as opposed to exon 1 in the consortium data. This may suggest that there is a presence of an alternative promoter within exon 2. Assuming that this is true, this would suggest that the T7-GSP primer has an efficacy for binding to the start of exon 2 as well as exon 1. This result was predictable as approximately 52% of RefSeq genes are subjected to putative alternative promoters (Kimura et al., 2006). Furthermore, 65 % of all promoters are within 500 bps of a CpG island, and 95 % of all promoters are within 10 KB of a CpG island (J. Wang, Ungar, Tseng, & Hannenhalli, 2007). Upon closer examination, a CpG island (CpG : 53) was found within 50 bps of exon 1, and revealed to show a high level of the H3K4Me3 mark, which is indicative of the presence of promoters. However, exon 2 did not show any considerable level of the mark, which indicates that exon 2 is unlikely to carry an alternative promoter site. A more feasible idea to explain the results (**Table 4.5**) may be that the *in vitro* RNA may have been fragmented at the junction of exon 1 and 2, or sequenced partially by the MinION.

5.8 *GAPDH* transcripts

GAPDH is a control housekeeping gene with high expression, and a short transcript length. Full-length *GAPDH* transcripts were abundant in various runs (**Table 4.7**). There was clear enrichment in both experiment 2 and 4 compared to the control experiment. In particular, the normalised ratios between different type of *GAPDH* isoforms across different experiments were similar (**Table 4.11**). This is potential evidence of linear amplification. These results collectively suggest that the enrichment method was successful and has the potential to deliver linear amplification for genes of interest with careful optimisation. Only a subset of the MinION consortium data was used for comparison with *GAPDH* due time restraints. This subset data was also normalised to minimise the variation in the datasets.

5.9 *RAD51C* transcripts

RAD51C has a moderate level of expression and has a similar transcript length to *GAPDH*. This was well reflected across all datasets, where numerous full-length isoforms were observed. Experiment 1 showed clear 3' bias, whereas this bias was less evident in the other datasets (**Table 4.9**). This was due to the lack of data from the second experiment, and a wide distribution of isoform types in experiment 4 and the consortium dataset (**Table 4.9**). The absolute counts and TPM suggest a clear enrichment in experiment 4 compared to all other datasets (**Tables 4.9-4.10**). This supports the fact that the protocol of the fourth experiment would be a good basis for future work. The enrichment outcomes appeared to be superior with *RAD51C* compared to the results of *BRCA1* and *CHEK2*, and this may be associated with the higher expression level and short transcript length of the gene (**Table 4.1**). Exon-skipping events were observed in experiment 4 and the MinION consortium dataset, and not in the smaller datasets of experiment 1 and 2. Intronic insertion events were observed only in experiment 1, which may be due to errors in alignment, or through an alternative splicing event (Cartegni et al., 2002). There was no evidence for linear enrichment of different isoforms in the fourth experiment, as the ratios between different types of isoforms across different experiments were inconsistent (**Table 4.12**).

5.10 Comparison of observed alternative splicing events with external datasets

As there were differences in the mRNA isolation step between the control and enrichment experiments, comparisons were made with external datasets to test the validity of the data. Each external dataset used different combinations of LCL, with and without NMD inhibition. The data published by Colombo et al. (2014) presented RNA-seq data from whole blood leukocytes, Ficoll-isolated peripheral blood mononuclear cells, primary cultures of stimulated peripheral blood lymphocytes and LCLs. Lattimore et al (2018) presented RNA-seq data on various LCLs with the same NMD inhibition utilised in experiments 1-4. The work of de Jong, et al. (2017), from this laboratory, revealed exon-skipping co-occurrence in *BRCA1* with long RNA-seq using the MinION. This data utilised the same LCL with NMD inhibition as experiments 1-4. Davy et al. (2017) data presented RNA-seq data from RNA extracted from various control and patient LCL without NMD inhibiton. The Nanopore RNA Consortium data, the only other study which used DRS, utilised a different LCL without NMD inhibition, and the MCRG RNA-seq data used a different LCL as experiments 1-4 with and without NMD inhibition. Due to these variations, these datasets may prove to be inefficient for comparing the data produced by this project. However, using them collectively provided a satisfactory comparison.

BRCA1 is a well-studied gene, and therefore most alternatively spliced isoforms detected in this project were described previously in the literature (Colombo et al., 2014; Davy et al., 2017; Lattimore et al., 2018). Non-described isoforms with multiple exon-skipping events along further distances were only detected in the Nanopore RNA Consortium total data (**Table 4.13**).

Alternative splicing events were scarce with *CHEK2* datasets, and insertion events in intron 2 and splicing of exons 6-7, 11-15 were observed in the consortium total data, but were not present in the published data mentioned above (**Table 4.14**). Novel *RAD51C* isoforms (Δ3,5-6, Δ3,7 and Δ4,7) not previously described in literature were found in my experiments (**Table 4.16**). The functional significance of alternative splicing within *CHEK2* and *RAD51C* transcripts has been documented (Janatova et al., 2015; Staalesen et al., 2004). Across all *GAPDH* datasets, only single exon-skipping events were observed. These events were only recognised in experiments 1-4 and the MCRG RNA-seq dataset, and do not appear to have been specifically reported elsewhere. However, they were able to be visualized in the ENCODE Caltech RNA-seq (Release 2) dataset on the UCSC genome browser. Interestingly, the consortium subset data did not show any splicing events in *GAPDH* (**Table 4.15**). These differences may be attributable to the use of NMD inhibitors in experiments 1-4 and the MCRG RNA-seq dataset. As NMD inhibitors prevent NMD mechanisms degrading truncated or aberrantly spliced transcripts, this may have caused the NMD inhibited dataset to present exon splicing events. The non-NMD inhibition data from the same MCRG dataset supports this claim as it contained fewer skipping events compared to the NMD inhibited dataset (**Appendix I**). Novel *GAPDH* isoforms in a mouse model have been reported, (Menon et al., 2009) however, no report describing human *GAPDH* isoforms was evident in the literature. Overall, these results confirm that the nanopore DRS method is able to provide useful information on novel splice isoforms and exon-skipping co-occurrence.

5.11 Limitations and future directions

Accuracy limits the feasibility of this enrichment method, in its current form, to be used for classifying VUS which impact on splicing. Firstly, the high number of manipulations carried out in the enrichment method was a potential source of artifacts within the results. These include sample preparation, library preparation, sequencing and downstream bioinformatics processing steps, which may be mitigated by implementing replications such as biological, technical, and cross-sequencing-platform replications (Robasky, Lewis, & Church, 2013). For instance, a single sequencing platform is not sufficient for deriving accurate variant calling of Single Nucleotide Polymorphisms (SNP) (Ratan et al., 2013). Therefore, using multiple sequencing platforms for comparison could be valuable, although this will increase costs. These experiments also employed 1D nanopore sequencing, which is known to be error prone (> 20 %) (Ip et al., 2015). Furthermore, counting and sorting isoforms manually may have introduced human error. Therefore, collectively, these sources of potential artifacts may have influenced the results obtained. Improving these aspects of the procedure would be important goals for future work.

Due to the limitations in time and resources, less than the optimum number of reactions were carried out through this project. Alternative long-range PCR, temperature gradient PCR, T7-primer PCR and T7 and oligo $dT_{(20)}$ primer PCR were not carried out in all relevant cases. Performing these quality control approaches will be imperative for future evaluation of the enrichment method.

Additionally, RNA is an unstable molecule compared to DNA and proteins. Therefore, there were many chances of RNA degradation throughout the experimentation process, which included freeze-thawing, and pipetting with fine tips. However, a study suggests that degradation of RNA derived from cancer tissue was less affected by freeze-thaw cycles compared to para-cancerous tissue RNA (Hu et al., 2017). This suggests that perhaps, the application of this method on biopsied breast cancer tissue may be less prone to freeze-thaw cycles compared to control LCLs (Hu et al., 2017).

Future experiments could include using this approach with other cell lines with known variants, which could be used to optimise the efficacy. These cells lines may be chosen from previous experiments described in the literature so that direct comparisons can be made (Colombo et al., 2014; Davy et al., 2017). Furthermore, this experiment could be tested with breast cancer tissue samples to further validate the approach. Carrying out the first experiment with total RNA instead of total mRNA may provide a better alternative for a reference experiment. Moreover, it would be beneficial to repeat the third experiment without pooling of T7-GSP and increasing the time of *in vitro* transcription incubation. Testing the effect of extra PCR cycles in the second strand cDNA synthesis step and incubation time in the *in vitro* transcription step may also provide beneficial information.

As RT-qPCR methods are able to provide a quantitative method for quantifying isoforms, the novel enrichment method should be compared with these methods. Using RT-qPCR probes which span exon-exon junctions for all query genes will be critical to avoid the need for DNAse-treatment, which proved to be detrimental (Barber, Harmer, Coleman, & Clark, 2005). While the datasets produced by this project were successfully aligned with Mimimap2, further development of quantification and alignment pipelines will be vital for future nanopore DRS research. As traditional RNA-seq data present with a length bias for longer genes (Oshlack & Wakefield, 2009), a normalisation approach for minimizing this bias may also be beneficial for comparing with DRS data in future studies. This could include utilising Reads Per Kilobase Million (RPKM) instead of TPM. It is unclear whether there is a length bias in DRS, as the coupling of a long-read sequencer such as the MinION may mitigate the length bias. Therefore, examining two genes with the same expression level and different transcript lengths may provide more evidence.

Additionally, the enrichment method incorporated a cDNA conversion and an *in vitro* transcription step, which lead to a loss of the ability to observe RNA modification. The ability to visualise RNA modifications is a key advantage of DRS, and therefore the loss of RNA modification information in the enrichment method was a limiting factor. Therefore, exploring further enrichment methods which do not involve RNA production will be profitable.

5.12 Conclusions

In this project, a novel RNA isoform enrichment method was developed and trialled with RNA extracted from a control LCL. The advantages of using this method lie in being able to visualise RNA isoform information for a set of specific genes of interest, and potentially to quantify these isoforms. If developed further, this method may be useful in clinical laboratory settings for providing more information to classify VUS that affect splicing and could be further utilised in many applications beyond cancer or the biomedical setting.

Through testing a panel of different genes, we found that *RAD51C* and *GAPDH* were able to be enriched with this novel method. It was not clear whether enrichment of *BRCA1* and *CHEK2* had occurred due to the lack of isoform type comparability between datasets. However, the absolute number of *BRCA1* and *CHEK2* counts were greater in experiment 4 compared to the control experiment, and the enrichment in the same experiments observed with *RAD51C* and *GAPDH* suggests that this method warrants further investigation. Whether this method is linear (i.e. maintains the isoform diversity that exists *in vivo*) is subject to debate, however, our preliminary bioinformatic evidence suggests that *GAPDH* isoforms were enriched in a linear fashion. This suggests that the method can be optimised for use in the future with genes such as *BRCA1* and may be used in the context of other diseases. Therefore, with successful optimisation, this novel enrichment method may be a valuable method to be used in studying VUS effects on alternative splicing patterns.

In summary, further research into this method may provide new tools for studying the effects of VUS on alternative splicing patterns, allowing an improved method of classifying VUS that impact on splicing. Furthermore, this method may be implemented in clinical settings and has the prospects of providing many individuals the opportunity to prevent the onset of breast and/or ovarian cancer. However, more refinement and optimisation of this method would be required before it could be applied in a clinical setting.

References

Agilent Technologies. (2016). *Comparison of RIN and RINe Algorithms for the Agilent 2100 Bioanalyzer and the Agilent 2200 TapeStation systems Technical Overview*. Retrieved from https://www.agilent.com/cs/library/technicaloverviews/public/5990-9613EN.pdf

Ahlborn, L. B., Dandanell, M., Steffensen, A. Y., Jonson, L., Nielsen, F. C., & Hansen, T. V. (2015). Splicing analysis of 14 BRCA1 missense variants classifies nine variants as pathogenic. *Breast Cancer Res Treat, 150*(2), 289-298. doi:10.1007/s10549-015-3313-7

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol, 12*(2), R18. doi:10.1186/gb-2011-12-2-r18

Albala, J. S., Thelen, M. P., Prange, C., Fan, W., Christensen, M., Thompson, L. H., & Lennon, G. G. (1997). Identification of a novel human RAD51 homolog, RAD51B. *Genomics, 46*(3), 476-479. doi:10.1006/geno.1997.5062

Ali, M. M., Li, F., Zhang, Z., Zhang, K., Kang, D. K., Ankrum, J. A., . . . Zhao, W. (2014). Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine. *Chem Soc Rev, 43*(10), 3324-3341. doi:10.1039/c3cs60439j

Au, K. F., Sebastiano, V., Afshar, P. T., Durruthy, J. D., Lee, L., Williams, B. A., . . . Wong, W. H. (2013). Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U S A, 110*(50), E4821-4830. doi:10.1073/pnas.1320101110

Baralle, D., & Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. *J Med Genet, 42*(10), 737-748. doi:10.1136/jmg.2004.029538

Barber, R. D., Harmer, D. W., Coleman, R. A., & Clark, B. J. (2005). GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol Genomics, 21*(3), 389-395. doi:10.1152/physiolgenomics.00025.2005

Beaudet, A. L. (2015). Global genetic carrier testing: a vision for the future. *Genome Med, 7*(1), 79. doi:10.1186/s13073-015-0204-9

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed, 98*(6), 236-238. doi:10.1136/archdischild-2013-304340

Behzadnia, N., Golas, M. M., Hartmuth, K., Sander, B., Kastner, B., Deckert, J., . . . Luhrmann, R. (2007). Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. *EMBO J, 26*(6), 1737-1748. doi:10.1038/sj.emboj.7601631

Belfiore, A., Frasca, F., Pandini, G., Sciacca, L., & Vigneri, R. (2009). Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease. *Endocr Rev, 30*(6), 586-623. doi:10.1210/er.2008-0047

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59. doi:10.1038/nature07517

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A, 74*(8), 3171-3175.

Berglund, J. A., Chua, K., Abovich, N., Reed, R., & Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell, 89*(5), 781-787.

Bochar, D. A., Wang, L., Beniya, H., Kinev, A., Xue, Y., Lane, W. S., . . . Shiekhattar, R. (2000). BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell, 102*(2), 257-265.

Bolisetty, M. T., Rajadinakaran, G., & Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol, 16*, 204. doi:10.1186/s13059-015-0777-z

Bork, P., Hofmann, K., Bucher, P., Neuwald, A. F., Altschul, S. F., & Koonin, E. V. (1997). A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J, 11*(1), 68-76.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. doi:10.3322/caac.21492

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol, 34*(5), 525-527. doi:10.1038/nbt.3519

Breitbart, R. E., & Nadal-Ginard, B. (1987). Developmentally induced, muscle-specific trans factors control the differential splicing of alternative and constitutive troponin T exons. *Cell, 49*(6), 793-803.

Brown, S. D., Nagaraju, S., Utturkar, S., De Tissera, S., Segovia, S., Mitchell, W., . . . Kopke, M. (2014). Comparison of single-molecule sequencing and hybrid approaches for finishing the genome of Clostridium autoethanogenum and analysis of CRISPR systems in industrial relevant Clostridia. *Biotechnol Biofuels, 7*, 40. doi:10.1186/1754-6834-7-40

Burnett, G., & Kennedy, E. P. (1954). The enzymatic phosphorylation of proteins. *J Biol Chem, 211*(2), 969-980.

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., . . . Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun, 8*, 16027. doi:10.1038/ncomms16027

Caldas, H., Fangusaro, J. R., Boue, D. R., Holloway, M. P., & Altura, R. A. (2007). Dissecting the role of endothelial SURVIVIN DeltaEx3 in angiogenesis. *Blood, 109*(4), 1479-1489. doi:10.1182/blood-2006-02-003749

Cantor, S., Drapkin, R., Zhang, F., Lin, Y., Han, J., Pamidi, S., & Livingston, D. M. (2004). The BRCA1-associated protein BACH1 is a DNA helicase targeted by clinically relevant inactivating mutations. *Proc Natl Acad Sci U S A, 101*(8), 2357-2362.

Cantor, S. B., Bell, D. W., Ganesan, S., Kass, E. M., Drapkin, R., Grossman, S., . . . Livingston, D. M. (2001). BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. *Cell, 105*(1), 149-160.

Cao, H., Wu, H., Luo, R., Huang, S., Sun, Y., Tong, X., . . . Wang, J. (2015). De novo assembly of a haplotype-resolved human genome. *Nat Biotechnol, 33*(6), 617-622. doi:10.1038/nbt.3200

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., & DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics, 13*, 375. doi:10.1186/1471-2164-13-375

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet, 3*(4), 285-298. doi:10.1038/nrg775

Casneuf, T., Van de Peer, Y., & Huber, W. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics, 8*, 461. doi:10.1186/1471-2105-8-461

Chang, S., Biswas, K., Martin, B. K., Stauffer, S., & Sharan, S. K. (2009). Expression of human BRCA1 variants in mouse ES cells allows functional analysis of BRCA1 mutations. *J Clin Invest, 119*(10), 3160-3171. doi:10.1172/JCI39836

Chen, X., Truong, T. T., Weaver, J., Bove, B. A., Cattie, K., Armstrong, B. A., . . . Godwin, A. K. (2006). Intronic alterations in BRCA1 and BRCA2: effect on mRNA splicing fidelity and expression. *Hum Mutat, 27*(5), 427-435. doi:10.1002/humu.20319

Cheon, J. Y., Mozersky, J., & Cook-Deegan, R. (2014). Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Med, 6*(12), 121. doi:10.1186/s13073-014-0121-3

Chu, Y., & Corey, D. R. (2012). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther, 22*(4), 271-274. doi:10.1089/nat.2012.0367

Cocquet, J., Chong, A., Zhang, G., & Veitia, R. A. (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics, 88*(1), 127-131. doi:10.1016/j.ygeno.2005.12.013

Colombo, M., Blok, M. J., Whiley, P., Santamarina, M., Gutierrez-Enriquez, S., Romero, A., . . . De La Hoya, M. (2014). Comprehensive annotation of splice junctions supports pervasive alternative splicing at the BRCA1 locus: a report from the ENIGMA consortium. *Hum Mol Genet, 23*(14), 3666-3680. doi:10.1093/hmg/ddu075

Cortez, D., Wang, Y., Qin, J., & Elledge, S. J. (1999). Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. *Science, 286*(5442), 1162-1166.

Crawford, B., Adams, S. B., Sittler, T., van den Akker, J., Chan, S., Leitner, O., . . . van 't Veer, L. (2017). Multi-gene panel testing for hereditary cancer predisposition in unsolved high-risk breast and ovarian cancer patients. *Breast Cancer Res Treat, 163*(2), 383-390. doi:10.1007/s10549-017-4181-0

Croville, G., Le Loc'h, G., Zanchetta, C., Manno, M., Camus-Bouclainville, C., Klopp, C., . . . Guerin, J. L. (2018). Rapid whole-genome based typing and surveillance of avipoxviruses using nanopore sequencing. *J Virol Methods, 261*, 34-39. doi:10.1016/j.jviromet.2018.08.003

Dacheux, E., Vincent, A., Nazaret, N., Combet, C., Wierinckx, A., Mazoyer, S., . . . Venezia, N. D. (2013). BRCA1-Dependent Translational Regulation in Breast Cancer Cells. *PLoS One, 8*(6), e67313. doi:10.1371/journal.pone.0067313

Davy, G., Rousselin, A., Goardon, N., Castera, L., Harter, V., Legros, A., . . . Krieger, S. (2017). Detecting splicing patterns in genes involved in hereditary breast and ovarian cancer. *Eur J Hum Genet, 25*(10), 1147-1154. doi:10.1038/ejhg.2017.116

de Jong, L. C., Cree, S., Lattimore, V., Wiggins, G. A. R., Spurdle, A. B., kConFab, I., . . . Walker, L. C. (2017). Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Res, 19*(1), 127. doi:10.1186/s13058-017-0919-1

de la Fuente, L., & Tardaguila, M. (2018). Welcome to SQANTI: Structural and Quality Annotation of Novel Transcript Isoforms.   Retrieved from https://bitbucket.org/ConesaLab/sqanti

Delsert, C., Morin, N., & Klessig, D. F. (1989). cis-acting elements and a trans-acting factor affecting alternative splicing of adenovirus L1 transcripts. *Mol Cell Biol, 9*(10), 4364-4371.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., . . . Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature, 412*(6849), 822-826. doi:10.1038/35090585

Dias Neto, E., Correa, R. G., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., da Silva, W., Jr., . . . Simpson, A. J. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc Natl Acad Sci U S A, 97*(7), 3491-3496.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., . . . French StatOmique, C. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform, 14*(6), 671-683. doi:10.1093/bib/bbs046

Dosanjh, M. K., Collins, D. W., Fan, W., Lennon, G. G., Albala, J. S., Shen, Z., & Schild, D. (1998). Isolation and characterization of RAD51C, a new human member of the RAD51 family of related genes. *Nucleic Acids Res, 26*(5), 1179-1184.

Duzkale, H., Shen, J., McLaughlin, H., Alfares, A., Kelly, M. A., Pugh, T. J., . . . Lebo, M. S. (2013). A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet, 84*(5), 453-463. doi:10.1111/cge.12257

Eggington, J. M., Bowles, K. R., Moyes, K., Manley, S., Esterling, L., Sizemore, S., . . . Wenstrup, R. J. (2014). A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. *Clin Genet, 86*(3), 229-237. doi:10.1111/cge.12315

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323*(5910), 133-138. doi:10.1126/science.1162986

Eklund, A. C., Turner, L. R., Chen, P., Jensen, R. V., deFeo, G., Kopf-Sill, A. R., & Szallasi, Z. (2006). Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat Biotechnol, 24*(9), 1071-1073. doi:10.1038/nbt0906-1071

Erlich, H. A., Gelfand, D., & Sninsky, J. J. (1991). Recent advances in the polymerase chain reaction. *Science, 252*(5013), 1643-1651.

Farber-Katz, S., Hsuan, V., Wu, S., Landrith, T., Vuong, H., Xu, D., . . . Karam, R. (2018). Quantitative Analysis of BRCA1 and BRCA2 Germline Splicing Variants Using a Novel RNA-Massively Parallel Sequencing Assay. *Front Oncol, 8*, 286. doi:10.3389/fonc.2018.00286

Faucon, F., Gaude, T., Dusfour, I., Navratil, V., Corbel, V., Juntarajumnong, W., . . . David, J. P. (2017). In the hunt for genomic markers of metabolic resistance to pyrethroids in the mosquito Aedes aegypti: An integrated next-generation sequencing approach. *PLoS Negl Trop Dis, 11*(4), e0005526. doi:10.1371/journal.pntd.0005526

Frith, M. (2018). Aligning long RNA and DNA reads to a genome. Retrieved from https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md

Frith, M. C., Wan, R., & Horton, P. (2010). Incorporating sequence quality data into alignment improves DNA read mapping. *Nucleic Acids Res, 38*(7), e100. doi:10.1093/nar/gkq010

Fu, X. Y., & Manley, J. L. (1987). Factors influencing alternative splice site utilization in vivo. *Mol Cell Biol, 7*(2), 738-748.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., . . . Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods, 15*(3), 201-206.

Ge, H., & Manley, J. L. (1990). A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro. *Cell, 62*(1), 25-34.

Ghosh, M., Tucker, D. E., Burchett, S. A., & Leslie, C. C. (2006). Properties of the Group IV phospholipase A2 family. *Prog Lipid Res, 45*(6), 487-510. doi:10.1016/j.plipres.2006.05.003

Goodson, M. L., Jonas, B. A., & Privalsky, M. L. (2005). Alternative mRNA splicing of SMRT creates functional diversity by generating corepressor isoforms with different affinities for different nuclear receptors. *J Biol Chem, 280*(9), 7493-7503. doi:10.1074/jbc.M411514200

Grindedal, E. M., Heramb, C., Karsrud, I., Ariansen, S. L., Maehle, L., Undlien, D. E., . . . Schlichting, E. (2017). Current guidelines for BRCA testing of breast cancer patients are insufficient to detect all mutation carriers. *BMC Cancer, 17*(1), 438. doi:10.1186/s12885-017-3422-2

Gueroussov, S., Gonatopoulos-Pournatzis, T., Irimia, M., Raj, B., Lin, Z. Y., Gingras, A. C., & Blencowe, B. J. (2015). An alternative splicing event amplifies evolutionary differences between vertebrates. *Science, 349*(6250), 868-873. doi:10.1126/science.aaa8381

Haddad, F., Qin, A. X., Giger, J. M., Guo, H., & Baldwin, K. M. (2007). Potential pitfalls in the accuracy of analysis of natural sense-antisense RNA pairs by reverse transcription-PCR. *BMC Biotechnol, 7*, 21. doi:10.1186/1472-6750-7-21

Hakem, R., de la Pompa, J. L., Sirard, C., Mo, R., Woo, M., Hakem, A., . . . Mak, T. W. (1996). The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. *Cell, 85*(7), 1009-1023.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science, 250*(4988), 1684-1689.

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., . . . Feldmann, H. (2016). Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg Infect Dis, 22*(2), 331-334. doi:10.3201/eid2202.151796

Holland, P. M., Abramson, R. D., Watson, R., & Gelfand, D. H. (1991). Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A, 88*(16), 7276-7280.

Howarth, D. R., Lum, S. S., Esquivel, P., Garberoglio, C. A., Senthil, M., & Solomon, N. L. (2015). Initial Results of Multigene Panel Testing for Hereditary Breast and Ovarian Cancer and Lynch Syndrome. *Am Surg, 81*(10), 941-944.

Hu, Y., Han, H., Wang, Y., Song, L., Cheng, X., Xing, X., . . . Ji, J. (2017). Influence of Freeze-Thaw Cycles on RNA Integrity of Gastrointestinal Cancer and Matched Adjacent Tissues. *Biopreserv Biobank, 15*(3), 241-247. doi:10.1089/bio.2016.0035

Hug, N., Longman, D., & Caceres, J. F. (2016). Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res, 44*(4), 1483-1495. doi:10.1093/nar/gkw010

Ip, C. L. C., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., . . . Reference, C. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Res, 4*, 1075. doi:10.12688/f1000research.7201.1

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., . . . Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods, 11*(2), 163-166.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol, 17*(1), 239. doi:10.1186/s13059-016-1103-0

Janatova, M., Soukupova, J., Stribrna, J., Kleiblova, P., Vocka, M., Boudova, P., . . . Pohlreich, P. (2015). Mutation Analysis of the RAD51C and RAD51D Genes in High-Risk Ovarian Cancer Patients and Families from the Czech Republic. *PLoS One, 10*(6), e0127711. doi:10.1371/journal.pone.0127711

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A, 93*(24), 13770-13773.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene, 514*(1), 1-30. doi:10.1016/j.gene.2012.07.083

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664. doi:10.1101/gr.229202

Kesselheim, A. S., & Mello, M. M. (2010). Gene patenting--is the pendulum swinging back? *N Engl J Med, 362*(20), 1855-1858. doi:10.1056/NEJMp1004026

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., . . . Sugano, S. (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res, 16*(1), 55-65. doi:10.1101/gr.4039406

Koonin, E. V., Altschul, S. F., & Bork, P. (1996). BRCA1 protein products ... Functional motifs. *Nat Genet, 13*(3), 266-268. doi:10.1038/ng0796-266

Kramer, A., & Utans, U. (1991). Three protein factors (SF1, SF3 and U2AF) function in pre-splicing complex formation in addition to snRNPs. *EMBO J, 10*(6), 1503-1509.

Lattimore, V. L., Currie, M., Lintott, C., Sullivan, J., Robinson, B. A., & Walker, L. C. (2015). Meeting the challenges of interpreting variants of unknown clinical significance in BRCA testing. *N Z Med J, 128*(1419), 56-61.

Lattimore, V. L., Pearson, J. F., Currie, M. J., Spurdle, A. B., kConFab, I., Robinson, B. A., & Walker, L. C. (2018). Investigation of Experimental Factors That Underlie BRCA1/2 mRNA Isoform Expression Variation: Recommendations for Utilizing Targeted RNA Sequencing to Evaluate Potential Spliceogenic Variants. *Front Oncol, 8*, 140. doi:10.3389/fonc.2018.00140

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif, 3*, 1-8. doi:10.1016/j.bdq.2015.02.001

Lee, J. S., Collins, K. M., Brown, A. L., Lee, C. H., & Chung, J. H. (2000). hCds1-mediated phosphorylation of BRCA1 regulates the DNA damage response. *Nature, 404*(6774), 201-204. doi:10.1038/35004614

Lee, S. C., & Abdel-Wahab, O. (2016). Therapeutic targeting of splicing in cancer. *Nat Med, 22*(9), 976.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., . . . Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods, 7*(9), 709-715. doi:10.1038/nmeth.1491

Li, H. (2017). Which human reference genome to use? Retrieved from http://lh3.github.io/2017/11/13/which-human-reference-genome-to-use

Li, H. (2018a). Minimap2. Retrieved from https://github.com/lh3/minimap2

Li, H. (2018b). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100. doi:10.1093/bioinformatics/bty191

Li, M., & Yu, X. (2013). Function of BRCA1 in the DNA damage response is mediated by ADP-ribosylation. *Cancer Cell, 23*(5), 693-704. doi:10.1016/j.ccr.2013.03.025

Li, S., Hu, Z., Zhao, Y., Huang, S., & He, X. (2018). Transcriptome-Wide Analysis Reveals the Landscape of Aberrant Alternative Splicing Events in Liver Cancer. *Hepatology*. doi:10.1002/hep.30158

Li, Z., Tian, Y., Tian, N., Zhao, X., Du, C., Han, L., & Zhang, H. (2015). Aberrant alternative splicing pattern of ADAR2 downregulates adenosine-to-inosine editing in glioma. *Oncol Rep, 33*(6), 2845-2852. doi:10.3892/or.2015.3907

Lindor, N. M., Guidugli, L., Wang, X., Vallee, M. P., Monteiro, A. N., Tavtigian, S., . . . Couch, F. J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum Mutat, 33*(1), 8-21. doi:10.1002/humu.21627

Link, S., Grund, S. E., & Diederichs, S. (2016). Alternative splicing affects the subcellular localization of Drosha. *Nucleic Acids Res, 44*(11), 5330-5343. doi:10.1093/nar/gkw400

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods, 12*(8), 733-735. doi:10.1038/nmeth.3444

Lorenz, T. C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *J Vis Exp*(63), e3998. doi:10.3791/3998

Lovf, M., Thomassen, G. O., Mertens, F., Cerveira, N., Teixeira, M. R., Lothe, R. A., & Skotheim, R. I. (2013). Assessment of fusion gene status in sarcomas using a custom made fusion gene microarray. *PLoS One, 8*(8), e70649. doi:10.1371/journal.pone.0070649

Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabl, M., Raschke, W., & Tonegawa, S. (1981). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. *Cell, 24*(2), 353-365.

Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., . . . Turner, D. J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods, 7*(2), 130-132. doi:10.1038/nmeth.1417

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics, 27*(6), 764-770. doi:10.1093/bioinformatics/btr011

Matsuoka, S., Huang, M., & Elledge, S. J. (1998). Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science, 282*(5395), 1893-1897.

McCabe, M. S., Cormican, P., Johnston, D., & Earley, B. (2018). Simultaneous detection of DNA and RNA virus species involved in bovine respiratory disease by PCR-free rapic tagmentation-based library preparation and MinION nanopore sequencing. *bioRxiv*, p269936.

Meetei, A. R., Medhurst, A. L., Ling, C., Xue, Y., Singh, T. R., Bier, P., . . . Wang, W. (2005). A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat Genet, 37*(9), 958-963. doi:10.1038/ng1626

Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F., & Dietz, H. C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet, 36*(10), 1073-1078. doi:10.1038/ng1429

Menon, R., & Omenn, G. S. (2010). Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res, 70*(9), 3440-3449. doi:10.1158/0008-5472.CAN-09-2631

Menon, R., Zhang, Q., Zhang, Y., Fermin, D., Bardeesy, N., DePinho, R. A., . . . States, D. J. (2009). Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. *Cancer Res, 69*(1), 300-309. doi:10.1158/0008-5472.CAN-08-2145

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., . . . et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science, 266*(5182), 66-71.

Millot, G. A., Carvalho, M. A., Caputo, S. M., Vreeswijk, M. P., Brown, M. A., Webb, M., . . . Group, E. C. F. A. W. (2012). A guide for functional analysis of BRCA1 variants of uncertain significance. *Hum Mutat, 33*(11), 1526-1537. doi:10.1002/humu.22150

Miyamoto, S., Nagamura, Y., Nakabo, A., Okabe, A., Yanagihara, K., Fukami, K., . . . Yamaguchi, H. (2018). Aberrant alternative splicing of RHOA is associated with loss of its expression and activity in diffuse-type gastric carcinoma cells. *Biochem Biophys Res Commun, 495*(2), 1942-1947. doi:10.1016/j.bbrc.2017.12.067

Morris, J. R., & Solomon, E. (2004). BRCA1 : BARD1 induces the formation of conjugated ubiquitin structures, dependent on K6 of ubiquitin, in cells during DNA replication and repair. *Hum Mol Genet, 13*(8), 807-817. doi:10.1093/hmg/ddh095

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol, 155*, 335-350.

Murray, J. I., Whitfield, M. L., Trinklein, N. D., Myers, R. M., Brown, P. O., & Botstein, D. (2004). Diverse and specific gene expression responses to stresses in cultured human cells. *Mol Biol Cell, 15*(5), 2361-2374. doi:10.1091/mbc.e03-11-0799

Nadeau, G., Boufaied, N., Moisan, A., Lemieux, K. M., Cayanan, C., Monteiro, A. N., & Gaudreau, L. (2000). BRCA1 can stimulate gene transcription by a unique mechanism. *EMBO Rep, 1*(3), 260-265. doi:10.1093/embo-reports/kvd059

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science, 320*(5881), 1344-1349. doi:10.1126/science.1158441

Nagy, E., & Maquat, L. E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem Sci, 23*(6), 198-199.

Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., . . . Wang, S. M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci U S A, 99*(9), 6152-6156. doi:10.1073/pnas.092140899

Narod, S. A., Feunteun, J., Lynch, H. T., Watson, P., Conway, T., Lynch, J., & Lenoir, G. M. (1991). Familial breast-ovarian cancer locus on chromosome 17q12-q23. *Lancet, 338*(8759), 82-83.

National Human Genome Research Institute. (05 May 2017). Breast Cancer Information Core BIC Retrieved from http://research.nhgri.nih.gov/bic/)

Neitzel, H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. *Hum Genet, 73*(4), 320-326.

New Zealand Ministry of Health. (2017). *New cancer registrations 2015*. Retrieved from: https://www.health.govt.nz/publication/new-cancer-registrations-2015

O'Neill, S. C., Rini, C., Goldsmith, R. E., Valdimarsdottir, H., Cohen, L. H., & Schwartz, M. D. (2009). Distress among women receiving uninformative BRCA1/2 results: 12-month outcomes. *Psychooncology, 18*(10), 1088-1096. doi:10.1002/pon.1467

Okoniewski, M. J., & Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics, 7*, 276. doi:10.1186/1471-2105-7-276

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct, 4*, 14. doi:10.1186/1745-6150-4-14

Ozsolak, F., & Milos, P. M. (2011a). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet, 12*(2), 87-98. doi:10.1038/nrg2934

Ozsolak, F., & Milos, P. M. (2011b). Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA, 2*(4), 565-570. doi:10.1002/wrna.84

Pachter, L. (2016). Kallisto. Retrieved from https://pachterlab.github.io/kallisto/manual

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet, 40*(12), 1413-1415. doi:10.1038/ng.259

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., . . . Blencowe, B. J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell, 16*(6), 929-941. doi:10.1016/j.molcel.2004.12.004

Park, J. H., Kim, I. J., Kang, H. C., Shin, Y., Park, H. W., Jang, S. G., . . . Park, J. G. (2004). Oligonucleotide microarray-based mutation detection of the K-ras gene in colorectal cancers with use of competitive DNA hybridization. *Clin Chem, 50*(9), 1688-1691. doi:10.1373/clinchem.2004.034017

Patel, K. J., Yu, V. P., Lee, H., Corcoran, A., Thistlethwaite, F. C., Evans, M. J., . . . Venkitaraman, A. R. (1998). Involvement of Brca2 in DNA repair. *Mol Cell, 1*(3), 347-357.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon. Retrieved from https://salmon.readthedocs.io/en/latest/salmon.html

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods, 14*(4), 417-419. doi:10.1038/nmeth.4197

Payne A, H. N., Rakyan V, Loose M. (2018). Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv*, p312256.

Petrini, J. H., & Stracker, T. H. (2003). The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol, 13*(9), 458-462.

Pittman, D. L., Weinberg, L. R., & Schimenti, J. C. (1998). Identification, characterization, and genetic mapping of Rad51d, a new mouse and human RAD51/RecA-related gene. *Genomics, 49*(1), 103-111. doi:10.1006/geno.1998.5226

Ponchel, F., Toomes, C., Bransfield, K., Leong, F. T., Douglas, S. H., Field, S. L., . . . Markham, A. F. (2003). Real-time PCR based on SYBR-Green I fluorescence: an alternative to the

TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. *BMC Biotechnol, 3*, 18. doi:10.1186/1472-6750-3-18

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., . . . Loman, N. J. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc, 12*(6), 1261-1276. doi:10.1038/nprot.2017.066

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., . . . Carroll, M. W. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature, 530*(7589), 228-232. doi:10.1038/nature16996

Rand, M. D., Vorojeikina, D., van Wijngaarden, E., Jackson, B. P., Scrimale, T., Zareba, G., . . . Watson, G. E. (2016). Methods for Individualized Determination of Methylmercury Elimination Rate and De-Methylation Status in Humans Following Fish Consumption. *Toxicol Sci, 149*(2), 385-395. doi:10.1093/toxsci/kfv241

Raponi, M., Douglas, A. G., Tammaro, C., Wilson, D. I., & Baralle, D. (2012). Evolutionary constraint helps unmask a splicing regulatory region in BRCA1 exon 11. *PLoS One, 7*(5), e37255. doi:10.1371/journal.pone.0037255

Raponi, M., Smith, L. D., Silipo, M., Stuani, C., Buratti, E., & Baralle, D. (2014). BRCA1 exon 11 a model of long exon splicing regulation. *RNA Biol, 11*(4), 351-359. doi:10.4161/rna.28458

Ratan, A., Miller, W., Guillory, J., Stinson, J., Seshagiri, S., & Schuster, S. C. (2013). Comparison of sequencing platforms for single nucleotide variant calls in a human sample. *PLoS One, 8*(2), e55089. doi:10.1371/journal.pone.0055089

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics, 13*(5), 278-289. doi:10.1016/j.gpb.2015.08.002

Robasky, K., Lewis, N. E., & Church, G. M. (2013). The role of replicates for error mitigation in next-generation sequencing. *Nat Rev Genet, 15*, 56-62.

Roberts, J. D., Preston, B. D., Johnston, L. A., Soni, A., Loeb, L. A., & Kunkel, T. A. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. *Mol Cell Biol, 9*(2), 469-476.

Rosenow, C., Saxena, R. M., Durst, M., & Gingeras, T. R. (2001). Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches. *Nucleic Acids Res, 29*(22), E112.

Ruskin, B., Zamore, P. D., & Green, M. R. (1988). A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell, 52*(2), 207-219.

Rychlik, W., Spencer, W. J., & Rhoads, R. E. (1990). Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res, 18*(21), 6409-6412.

Salzberg , A. C., Hu, J., Conroy, E. J., Cladel, N. M., Brucklacher, R. M., Bixler, G. V., & Kawasana, Y. I. (2017). Effects of duplicated mapped read PCR artifacts on RNA-seq differential expression analysis based on qRNA-seq. *bioRxiv*, p301259.

Schneider, G. F., & Dekker, C. (2012). DNA sequencing with nanopores. *Nat Biotechnol, 30*(4), 326-328. doi:10.1038/nbt.2181

Schneider, J., Buness, A., Huber, W., Volz, J., Kioschis, P., Hafner, M., . . . Sultmann, H. (2004). Systematic analysis of T7 RNA polymerase based in vitro linear RNA amplification for use in microarray experiments. *BMC Genomics, 5*(1), 29. doi:10.1186/1471-2164-5-29

Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res, 30*(12), e57.

Scully, R., Chen, J., Plug, A., Xiao, Y., Weaver, D., Feunteun, J., . . . Livingston, D. M. (1997). Association of BRCA1 with Rad51 in mitotic and meiotic cells. *Cell, 88*(2), 265-275.

Sharan, S. K., Morimatsu, M., Albrecht, U., Lim, D. S., Regel, E., Dinh, C., . . . Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature, 386*(6627), 804-810. doi:10.1038/386804a0

Sharon, D., Tilgner, H., Grubert, F., & Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol, 31*(11), 1009-1014. doi:10.1038/nbt.2705

Shiloh, Y. (2006). The ATM-mediated DNA-damage response: taking shape. *Trends Biochem Sci, 31*(7), 402-410. doi:10.1016/j.tibs.2006.05.004

Shinohara, A., Ogawa, H., & Ogawa, T. (1992). Rad51 protein involved in repair and recombination in S. cerevisiae is a RecA-like protein. *Cell, 69*(3), 457-470.

Shiovitz, S., & Korde, L. A. (2015). Genetics of breast cancer: a topic in evolution. *Ann Oncol, 26*(7), 1291-1299. doi:10.1093/annonc/mdv022

Siliciano, P. G., & Guthrie, C. (1988). 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev, 2*(10), 1258-1267.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods, 14*(4), 407-410.

Sinha, R., Nikolajewa, S., Szafranski, K., Hiller, M., Jahn, N., Huse, K., . . . Backofen, R. (2009). Accurate prediction of NAGNAG alternative splicing. *Nucleic Acids Res, 37*(11), 3569-3579. doi:10.1093/nar/gkp220

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics, 14*, 91. doi:10.1186/1471-2105-14-91

Spurdle, A. B., Lakhani, S. R., Healey, S., Parry, S., Da Silva, L. M., Brinkworth, R., . . . kConFab, I. (2008). Clinical classification of BRCA1 and BRCA2 DNA sequence variants: the value of cytokeratin profiles and evolutionary analysis--a report from the kConFab Investigators. *J Clin Oncol, 26*(10), 1657-1663. doi:10.1200/JCO.2007.13.2779

Staalesen, V., Falck, J., Geisler, S., Bartkova, J., Borresen-Dale, A. L., Lukas, J., . . . Lonning, P. E. (2004). Alternative splicing and mutation status of CHEK2 in stage III breast cancer. *Oncogene, 23*(52), 8535-8544. doi:10.1038/sj.onc.1207928

Stickeler, E., Kittrell, F., Medina, D., & Berget, S. M. (1999). Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene, 18*(24), 3574-3582. doi:10.1038/sj.onc.1202671

Tabor, S., & Richardson, C. C. (1985). A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. *Proc Natl Acad Sci U S A, 82*(4), 1074-1078.

Tabuchi, M., Tanaka, N., Nishida-Kitayama, J., Ohno, H., & Kishi, F. (2002). Alternative splicing regulates the subcellular localization of divalent metal transporter 1 isoforms. *Mol Biol Cell, 13*(12), 4371-4387. doi:10.1091/mbc.E02-03-0165

Tammaro, C., Raponi, M., Wilson, D. I., & Baralle, D. (2012). BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer. *Biochem Soc Trans, 40*(4), 768-772. doi:10.1042/BST20120140

Tammaro, C., Raponi, M., Wilson, D. I., & Baralle, D. (2014). BRCA1 EXON 11, a CERES (composite regulatory element of splicing) element involved in splice regulation. *Int J Mol Sci, 15*(7), 13045-13059. doi:10.3390/ijms150713045

Thomassen, M., Blanco, A., Montagna, M., Hansen, T. V., Pedersen, I. S., Gutierrez-Enriquez, S., . . . Vega, A. (2012). Characterization of BRCA1 and BRCA2 splicing variants: a collaborative report by ENIGMA consortium members. *Breast Cancer Res Treat, 132*(3), 1009-1023. doi:10.1007/s10549-011-1674-0

Toland, A. E., Forman, A., Couch, F. J., Culver, J. O., Eccles, D. M., Foulkes, W. D., . . . Committee, B. I. C. S. (2018). Clinical testing of BRCA1 and BRCA2: a worldwide snapshot of technological practices. *NPJ Genom Med, 3*, 7. doi:10.1038/s41525-018-0046-7

Tram, E., Savas, S., & Ozcelik, H. (2013). Missense variants of uncertain significance (VUS) altering the phosphorylation patterns of BRCA1 and BRCA2. *PLoS One, 8*(5), e62468. doi:10.1371/journal.pone.0062468

Vallejo-Illarramendi, A., Domercq, M., & Matute, C. (2005). A novel alternative splicing form of excitatory amino acid transporter 1 is a negative regulator of glutamate uptake. *J Neurochem, 95*(2), 341-348. doi:10.1111/j.1471-4159.2005.03370.x

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet, 30*(9), 418-426. doi:10.1016/j.tig.2014.07.001

van Dijk, S., van Asperen, C. J., Jacobi, C. E., Vink, G. R., Tibben, A., Breuning, M. H., & Otten, W. (2004). Variants of uncertain clinical significance as a result of BRCA1/2 testing: impact of an ambiguous breast cancer risk message. *Genet Test, 8*(3), 235-239. doi:10.1089/gte.2004.8.235

Vandenbroucke, I. I., Vandesompele, J., Paepe, A. D., & Messiaen, L. (2001). Quantification of splice variants using real-time PCR. *Nucleic Acids Res, 29*(13), E68-68.

Vegran, F., Boidot, R., Solary, E., & Lizard-Nacol, S. (2011). A short caspase-3 isoform inhibits chemotherapy-induced apoptosis by blocking apoptosome assembly. *PLoS One, 6*(12), e29058. doi:10.1371/journal.pone.0029058

Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., & Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proc Natl Acad Sci U S A, 115*(39), 9726-9731. doi:10.1073/pnas.1806447115

Volden, R., & Vollmers, C. (2018). Mandalorion-Episode-II. Retrieved from https://github.com/rvolden/Mandalorion-Episode-II

Walker, L. C., Whiley, P. J., Couch, F. J., Farrugia, D. J., Healey, S., Eccles, D. M., . . . Spurdle, A. B. (2010). Detection of splicing aberrations caused by BRCA1 and BRCA2 sequence variants encoding missense substitutions: implications for prediction of pathogenicity. *Hum Mutat, 31*(6), E1484-1505. doi:10.1002/humu.21267

Wang, J., Ungar, L. H., Tseng, H., & Hannenhalli, S. (2007). MetaProm: a neural network based meta-predictor for alternative human promoter prediction. *BMC Genomics, 8*, 374. doi:10.1186/1471-2164-8-374

Wang, Y., Liu, J., Huang, B. O., Xu, Y. M., Li, J., Huang, L. F., . . . Wang, X. Z. (2015). Mechanism of alternative splicing and its regulation. *Biomed Rep, 3*(2), 152-158. doi:10.3892/br.2014.407

Weitzel, J. N., McCaffrey, S. M., Nedelcu, R., MacDonald, D. J., Blazer, K. R., & Cullinane, C. A. (2003). Effect of genetic cancer risk assessment on surgical decisions at breast cancer diagnosis. *Arch Surg, 138*(12), 1323-1328; discussion 1329. doi:10.1001/archsurg.138.12.1323

West, S. C. (2003). Molecular views of recombination proteins and their control. *Nat Rev Mol Cell Biol, 4*(6), 435-445. doi:10.1038/nrm1127

Whiley, P. J., de la Hoya, M., Thomassen, M., Becker, A., Brandao, R., Pedersen, I. S., . . . consortium, E. (2014). Comparison of mRNA splicing assay protocols across multiple laboratories: recommendations for best practice in standardized clinical testing. *Clin Chem, 60*(2), 341-352. doi:10.1373/clinchem.2013.210658

Wolf, E., Kastner, B., Deckert, J., Merz, C., Stark, H., & Luhrmann, R. (2009). Exon, intron and splice site locations in the spliceosomal B complex. *EMBO J, 28*(15), 2283-2292. doi:10.1038/emboj.2009.171

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., . . . et al. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science, 265*(5181), 2088-2090.

Wu, T. (2018). GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences, and

GSNAP: Genomic Short-read Nucleotide Alignment Program. Retrieved from http://research-pub.gene.com/gmap/src/README

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics, 21*(9), 1859-1875. doi:10.1093/bioinformatics/bti310

Xia, B., Sheng, Q., Nakanishi, K., Ohashi, A., Wu, J., Christ, N., . . . Livingston, D. M. (2006). Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol Cell, 22*(6), 719-729. doi:10.1016/j.molcel.2006.05.022

Xia, F., Taghian, D. G., DeFrank, J. S., Zeng, Z. C., Willers, H., Iliakis, G., & Powell, S. N. (2001). Deficiency of human BRCA2 leads to impaired homologous recombination but maintains normal nonhomologous end joining. *Proc Natl Acad Sci U S A, 98*(15), 8644-8649. doi:10.1073/pnas.151253498

Xu, Q., Zhu, J., Zhao, S., Hou, Y., Li, F., Tai, Y., . . . Wei, C. (2017). Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing Approach for In-depth Understanding of Genes in Secondary Metabolism Pathways of Camellia sinensis. *Front Plant Sci, 8*, 1205. doi:10.3389/fpls.2017.01205

Xu, X., Weaver, Z., Linke, S. P., Li, C., Gotay, J., Wang, X. W., . . . Deng, C. X. (1999). Centrosome amplification and a defective G2-M cell cycle checkpoint induce genetic instability in BRCA1 exon 11 isoform-deficient cells. *Mol Cell, 3*(3), 389-395.

Yalaza, M., Inan, A., & Bozer, M. (2016). Male Breast Cancer. *J Breast Health, 12*(1), 1-8. doi:10.5152/tjbh.2015.2711

Yu, X., & Chen, J. (2004). DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. *Mol Cell Biol, 24*(21), 9478-9486. doi:10.1128/MCB.24.21.9478-9486.2004

Zhan, X., Yan, C., Zhang, X., Lei, J., & Shi, Y. (2018). Structure of a human catalytic step I spliceosome. *Science, 359*(6375), 537-545. doi:10.1126/science.aar6401

Zhang, J., Willers, H., Feng, Z., Ghosh, J. C., Kim, S., Weaver, D. T., . . . Xia, F. (2004). Chk2 phosphorylation of BRCA1 regulates DNA double-strand break repair. *Mol Cell Biol, 24*(2), 708-718.

Zhao, H., Hastie, T., Whitfield, M. L., Borresen-Dale, A. L., & Jeffrey, S. S. (2002). Optimization and evaluation of T7 based RNA linear amplification protocols for cDNA microarray analysis. *BMC Genomics*, *3*(1), 31.

Zhuo, B., Emerson, S., Chang, J. H., & Di, Y. (2016). Identifying stably expressed genes from multiple RNA-Seq data sets. *PeerJ*, *4*, e2791. doi:10.7717/peerj.2791

Appendix A: RNeasy® Mini Kit Protocol 4th Edition with Modifications

Materials:

- 13 mL Eppendorf Centrifuge Tubes
- RNeasy Mini Spin Columns
- Collection Tubes (1.5 mL)
- Collection Tubes (2 mL)
- Buffer RLT
- Buffer RW1
- Buffer RPE (concentrate)
- RNAse-Free Water
- Sterile, RNAse-free pipette tips
- Microcentrifuge
- 100% ethanol
- Fresh 70% ethanol
- 18-gauge needle
- 1mL syringe

Methods:

1. Pellet cells grown in suspension by centrifuging for 5 min at 300 x g in centrifuge tubes. Carefully remove all supernatant by aspiration.
2. Loosen the cell pellet thoroughly by flicking the tube. Add 600 µL of Buffer RLT and vortex to mix.
3. Pass the lysate at least 10 times through a 18-gauge needle fitted to an RNAse-free syringe.
4. Add 600 µL of 70% to the homogenised lysate and mix well by pipetting. Do not centrifuge.
5. Transfer 700 µL of the sample, including any precipitate that may have formed, to an RNeasy spin column placed in a 2mL collection tube. Close the lid gently, and centrifuge for 15 s at 8000 x g. Discard the flow-through.
6. Transfer the remaining the rest of the sample, including any precipitate that may have formed, to the same RNeasy spin column placed in a 2mL collection tube. Close the lid gently, and centrifuge for 15 s at 8000 x g. Discard the flow-through.
7. Add 700 µL Buffer RW1 to the RNeasy spin column. Close the lid gently, and centrifuge for 15 s at 8000 x g to wash the spin column membrane. Discard the flow-through.
8. Add 500 µL Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 15 s at 8000 x g to wash the spin column membrane. Discard the flow-through.
9. Add 500 µL Buffer RPE to the RNeasy spin column. Close the lid gently, and centrifuge for 2 min at 8000 x g to wash the spin column membrane. Discard the flow-through.

10. Place the RNeasy spin column in a new 2 mL collection tube and discard the old collection tube with the flow-through. Close the lid gently, and centrifuge at full 24000 x g for 1 min.
11. Place the RNeasy spin column in a new 1.5 mL collection tube. Add 50 µL RNAse-free water directly to the spin column membrane. Close the lid gently, and centrifuge at full speed for 1 min at 8000 x g to elute the RNA.

Appendix B: PolyA Spin™ mRNA Isolation Kit Protocol Version 1.1 with Modifications

Materials:

- Kit wash buffer
- Kit elution buffer
- Low salt buffer
- 5 M NaCl
- 3 M NaAc
- Glycogen solution
- Kit microcentrifuge tubes containing oligo (dT)25 –cellulose beads in storage buffer
- Microcentrifuge spin columns for isolation of poly(A)+ RNA
- Microcentrifuge tubes for the deposit of isolated poly(A)+ RNA
- 95% ethanol
- Sterile 13 x 100 mm disposable test tubes
- Microcentrifuge
- Rotary or other type of bidirectional shaker

Methods:

1. Add 50 µl of 5M NaCl per 450 µl of cell lysate or total RNA solution or dissolve total RNA sample in 450 µl of Elution Buffer then add 50 µl of 5M NaCl. Make sure that sample is totally dissolved. If not, microcentrifuge for 5 minutes to pellet insoluble material. Transfer RNA solution to clean microcentrifuge tube.

2. Heat at 65°C for 5 minutes and quickly cool in an ice bath for 3 minutes.

3. Apply total RNA solution to equilibrated oligo (dT)25 -cellulose, seal cap and mix thoroughly. Let stand at room temperature for 5 minutes agitating by hand.

4. Microcentrifuge for 10 seconds. Note: It is important to agitate beads during binding, washing and elution steps.

5. Pipette supernatant back into original microcentrifuge tube. Repeat steps 2-4.

6. Pipet supernatant back into original microcentrifuge tube for storage. It is recommended that no spin-column eluates be discarded until entire isolation procedure is completed and the results are evaluated. Eluates can be stored in sterile test tubes on ice.

7. Add 400 µl of Wash Buffer to oligo (dT)25 -cellulose beads. Agitate by hand to resuspend the cellulose beads. Using a 1 ml micropipette with sterile pipette tip, transfer Wash Buffer and beads to the column reservoir of a clean microcentrifuge spin column unit (provided with kit).

8. Let stand at room temperature for 2 minutes agitating by hand or place horizontally on rotary shaker. Microcentrifuge for 10 seconds. Remove column reservoir and transfer column eluent to a clean 13 x 100 cm test tube.

9. Add 400 µl of Wash Buffer to column reservoir and wash as in Step 8 three times.

10. Using the same method wash column once with 400 µl of Low Salt Buffer. *Note: This wash step with 0.1 M NaCl removes residual poly(A)– RNA which is bound to the cellulose column. This step can be omitted during a second round purification.

11. Remove spin-column reservoir and place in a clean microcentrifuge tube (provided with kit).

12. Add 200 µl of prewarmed Elution Buffer to column reservoir. Agitate by hand resuspending the cellulose beads. Let stand for 2 minutes agitating by hand or place on rotary shaker. Microcentrifuge for 10 seconds.

13. Repeat Step 12 using fresh prewarmed Elution Buffer.

14. Place Elution Buffer eluate on ice.

15. Second round purification is to be done immediately. Equilibrate fresh oligo (dT)25 -cellulose beads and repeat isolation procedure starting at step 1.

16. If poly(A)+ eluate is to be used at this point it can be ethanol precipitated: a) To the poly(A)+ eluent add 44 µl of 3 M Sodium Acetate, 20 µl of Glycogen Carrier and 1.0 ml of cold 95% ethanol. Allow to stand at –20°C for at least 30 minutes. At this point poly(A)+ material can be stored as an ethanol precipitate at –70°C until it is needed. To recover poly(A)+ material microcentrifuge for 15 minutes at 4°C. Carefully decant supernant, then wash the pellet (often not visible) with 70% ethanol. Recentrifuge briefly, decant supernant and allow pellet to air dry. Resuspend poly(A)+ material as required.

Appendix C: SuperScript™ III Reverse Transcriptase First-Strand cDNA Synthesis Protocol Revision Date 7 Dec 2004 with Modifications

Materials:

- SuperScript™ III Reverse Transcriptase (200 U/ µL)
- 5X First-Strand Buffer
- 0.1 M DTT
- Anchored Oligo(dT)20 (50 µM)
- Total RNA
- 10 mM dNTP mix (10 mM each dATP, dGTP, dCTP and dTTP at neutral pH)
- Sterile, distilled water
- RNaseOUT™ Recombinant RNAse Inhibitor (40 units/µL)
- RNaseH or RNAseA/T1 mix

Methods:

The following 20 µL reaction volume can be used for 10 pg-5 µg of total RNA or 10 pg-500 ng of mRNA.

1. Add the following components to a nuclease-free microcentrifuge tube.

1 µL of oligo(dT)20 (50µM)

10 pg-5 µg total RNA or 10 pg-500 ng mRNA

1 µL 10 mM dNTP Mix

Sterile, distilled water to 13 µL

2. Heat mixture to 65 °C for 5 minutes and incubate on ice for at least 1 minute.
3. Collect the contents of the tube by brief centrifugation and add:

4 µL 5X first-strand buffer

1 µL 0.1 M DTT
1 µL RNaseOUT™ Recombinant RNAse Inhibitor

1 µL SuperScript™ III Reverse Transcriptase

4. Mix by inversion and flicking.
5. Incubate at 50 °C for 45 minutes.
6. Inactivate the reaction by heating at 70 °C for 15 minutes.
7. Add 1 µL of RNAseH or RNAseA/T1 mix and incubate at 37 °C for 20 minutes.

Appendix D: LongAmp® *Taq* PCR Kit Second-Strand cDNA Synthesis Protocol with Modifications

Materials:

- NEB 5X LongAmp Taq Reaction buffer
- 10 mM dNTP mix
- 5 µM GSP-T7 primer
- LongAmp Taq
- First strand cDNA
- Nuclease-free water

Methods:

1. Preheat thermocycler to 94 °C and keep all reagents on ice.
2. In one PCR tube, place the following reagents;

10 µL 5X LongAmp Taq Reaction Buffer

1.5 µL 10 mM dNTPs

4 µL 5 µM GSP-T7 primer

2 µL LongAmp Taq

Nuclease-free water to 50 µL

3. Perform two PCR cycles using optimal annealing temperature determined via a temperature gradient experiment.

94 °C for 1 min

Annealing temperature for 1 min

65 °C for 30 min

4 °C hold

Appendix E: HighPrep™ Magnetic Bead Clean Up Protocol with Modifications

Materials:

- HighPrep magnetic beads
- Double-stranded cDNA
- Eppendorf DNA LoBind microcentrifuge tubes
- Magnetic rack
- Fresh 70% ethanol
- Nuclease-free water

Methods:

1. Add 90 µL of resuspended HighPrep beads to 50 µL of double-stranded cDNA at room temperature.
2. Incubate by inverting for 5 minutes, spin down and pellet on magnet. Discard the superntant.
3. Keep on magnet, wash with 150 µL fresh 70% ethanol, by twisting the tube in the magnetic rack 2x 180 °, waiting for the pellet to migrate to the other side of the tube each time.
4. Briefly spin down, replace on magnet, pipitte off residual ethanol.
5. Repeat steps 3-4.
6. Resuspend pellet in 20 µL nuclease-free water, incubate at room temperature for 5 minutes by inversion.
7. Pellet beads on a magnet, remove eluate of double-stranded cDNA and transfer to a fresh DNA LoBind tube.

Appendix F: MEGAscript® Kit Protocol Revision G with Modifications

Materials:

- Nuclease-free water
- PTRI-Xef, 0.5 mg/mL (Control Template)
- Double-stranded cDNA isolated via HighPrep magnetic beads
- T7 enzyme mix
- 10X reaction buffer
- ATP solution
- CTP solution
- GTP solution
- UTP solution

Methods:

1. Thaw the frozen reagents Place the RNA Polymerase Enzyme Mix on ice, it is stored in glycerol and will not be frozen at –20°C. Vortex the 10X Reaction Buffer and the 4 ribonucleotide solutions (ATP, CTP, GTP, and UTP) until they are completely in solution. Once thawed, store the ribonucleotides on ice, but keep the 10X Reaction Buffer at room temperature while assembling the reaction. All reagents should be microfuged briefly before opening to prevent loss and/or contamination of material that may be present around the rim of the tube.
2. Assemble transcription reaction at room temp. The spermidine in the 10X Reaction Buffer can coprecipitate the template DNA if the reaction is assembled on ice. Add the 10X Reaction Buffer after the water and the ribonucleotides are already in the tube. The following amounts are for a single 20 µL reaction. Reactions may be scaled up or down if desired.

2 µL ATP solution

2 µL CTP solution

2 µL GTP solution

2 µL UTP solution

2 µL 10X reaction buffer

0.1-1 µg double-stranded cDNA or 1 µg of control template

2 µL enzyme mix

Nuclease-free water to 20 µL

Note: For convenience, mix equal volumes of the four ribonucleotide solutions together and add 8 µL of the mixture to a standard 20 µL reaction instead of adding the ribonucleotides separately.

3. Mix thoroughly Gently flick the tube or pipette the mixture up and down gently, and then microfuge tube briefly to collect the reaction mixture at the bottom of the tube.
4. Incubate at 37°C for 4, 16 or 20 hours. The optimal incubation time for a given template will vary depending on the size and transcriptional efficiency of your template.

Appendix G: RNAClean™ XP Magnetic Bead Clean-Up Protocol 001298v001 with Modifications

Materials:

- Agencourt® RNACLEAN XP beads
- Eppendorf DNA LoBind microcentrifuge tubes
- Magnetic rack
- Fresh 70% ethanol
- Nuclease-free water
- MEGAscript kit RNA product

Methods:

1. Transfer the MEGAscript kit RNA product into an Eppendorf LoBind tube.
2. Gently shake the Agencourt RNAClean XP bottle to resuspend any magnetic particles that may have settled. Add 1.8 x the volume of the RNA product.
3. Mix the beads and sample thoroughly by inversion and flicking.
4. Let the tube incubate at room temperature for 5 minutes before proceeding to the next step.
5. Place the tube onto the magnetic rack for 5 minutes to separate the beads from solution.
6. Slowly aspirate the cleared solution from the tube and discard.
7. Dispense 1000 µL of 70% ethanol into the tube and incubate for 30 seconds at room temperature. Aspirate out the ethanol and discard. Repeat a total of three washes.
8. Let the reaction tube air-dry 10 minutes on the magnetic rack with the cap open.
9. Elute the purified product from the beads with nuclease-free water.

Appendix H: Oxford Nanopore Technologies Direct RNA Sequencing Protocol Version DRS_9026_v1_revN_15Dec2016 with Modifications

Materials:

- Direct RNA sequencing kit (SQK-RNA001)
- Magnetic rack
- 1.5 mL Eppendorf DNA LoBind tubes
- 0.2 mL thin-walled PCR tubes
- Nuclease-free water
- Fresh 70% ethanol
- SuperScript III reverse transcriptase
- 10 mM dNTP mix
- NEBNext quick ligation reaction buffer
- Concentrated T4 DNA ligase 2M U/mL
- Agencourt RNAClean XP beads
- Qubit RNA HS assay kit
- Qubit dsDNA HS assay kit
- MEGAscript RNA

Methods:

1. Set up the MinION, flow-cell and host computer.
2. Open the MinKNOW GUI from the desktop icon and establish a local connection.
3. Plug the MinION into the computer. When the connection name appears under the Local tab, click Connect.
4. Enter the Sample ID and FlowcellID being used, and click Submit.
5. Click into the Sample ID box and name your sample using free text in alphanumeric format
6. Check the number of active pores available for the experiment, reported in the message panel or in notifications when the check is complete. Once the check is complete, the software will return to the Connection page. To see the active pore report, click on notifications. If the total number of pores reported is lower than 800, please re-run the flow-cell check using the instructions in this FAQ.

1. In a 0.2 ml thin-walled PCR tube, mix the reagents in the following order:

NEBNext Quick Ligation Reaction Buffer 3.0 µl

RNA 9.0 µl

RNA CS 0.5 µl

RT Adapter (RTA) 1.0 µl

T4 DNA Ligase 1.5 µl

Total 15 µl

Mix by pipetting and spin down.

2. Incubate the reaction for 10 minutes at room temperature.
3. Mix the following reagents together to make the reverse transcription master mix:

Nuclease-free water 9.0 µl

10 mM dNTPs 2.0 µl

5x first-strand buffer 8.0 µl

0.1 M DTT 4.0 µl

Total 23.0 µl

Add the master mix to the 0.2 ml PCR tube containing the RT adapter ligated RNA from the "RT Adapter ligation" step. Mix by pipetting.

4. Add 2 µl of SuperScript III reverse transcriptase to the reaction and mix by pipetting. Place the tube in a thermal cycler and incubate at 50° C for 50 min, then 70° C for 10 min, and bring the sample to 4° C before proceeding to the next step.
5. Transfer the sample to a 1.5 ml DNA LoBind Eppendorf tube. Resuspend the stock of Agencourt RNAClean XP beads by vortexing.
6. Add 72 µl of resuspended RNAClean XP beads to the reverse transcription reaction and mix by pipetting. Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature.
7. Prepare 200 µl of fresh 70% ethanol in nuclease-free water. Spin down the sample and pellet on a magnet. Keep the tube on the magnet, and pipette off the supernatant.
8. Keep the tube on magnet, and wash the beads with 150 µl of freshly prepared 70% ethanol without disturbing the pellet as described below.

   Keeping the magnetic rack on the benchtop, rotate the bead-containing tube by 180°. Wait for the beads to migrate towards the magnet and form a pellet.

   Rotate the tube 180° again (back to the starting position), and wait for the beads to pellet.

   Remove the 70% ethanol using a pipette, and discard. Spin down and place the tube back on the magnet. Pipette off any residual 70% ethanol.

Remove the tube from the magnetic rack and resuspend pellet in 20 µl nuclease-free water. Incubate for 5 minutes at room temperature. Pellet beads on magnet until the eluate is clear and colourless.

Pipette 20 µl of eluate into a clean 1.5 ml Eppendorf DNA LoBind tube.

9. In a clean 1.5 ml Eppendorf DNA LoBind tube, mix the reagents in the following order:

Reverse-transcribed RNA from the "Reverse Transcription" step 20.0 µl

NEBNext Quick Ligation Reaction Buffer 8.0 µl

RNA Adapter (RMX) 6.0 µl

Nuclease-free water 3.0 µl

T4 DNA Ligase 3.0 µl

Total 40 µl Mix by pipetting.

10. Incubate the reaction for 10 minutes at room temperature. Resuspend the stock of Agencourt RNAClean XP beads by vortexing. Add 40 µl of resuspended RNAClean XP beads to the adapter ligation reaction and mix by pipetting. Incubate on a Hula mixer (rotator mixer) for 5 minutes at room temperature. Spin down the sample and pellet on a magnet. Keep the tube on the magnet, and pipette off the supernatant.

11. Add 150 µl of the Wash Buffer (WSB) to the beads. Close the tube lid, and resuspend the beads by flicking the tube. Return the tube to the magnetic rack, allow beads to pellet and pipette off the supernatant. Repeat. IMPORTANT Agitating the beads by flicking results in a more efficient removal of free adapter, compared to adding the wash buffer and immediately aspirating. Remove the tube from the magnetic rack and resuspend pellet in 21 µl Elution Buffer.
12. Incubate for 10 minutes at room temperature. Pellet beads on magnet until the eluate is clear and colourless.
13. Remove and retain 21 µl of eluate into a clean 1.5 ml Eppendorf DNA LoBind tube. Quantify 1 µl of reverse-transcribed and adapted RNA using the Qubit fluorometer DNA HS assay - recovery aim ~200 ng. END OF STEP The reverse-transcribed and adapted RNA is now ready for loading into the MinION Flow-Cell.
14. Direct RNA sequencing Priming and loading the SpotON Flow-Cell ~15 minutes IMPORTANT
15. Thoroughly mix the contents of the RRB tube by vortexing or pipetting, and spin down briefly.
16. Flip back the MinION lid and slide the sample port cover clockwise to that the sample port is visible. Priming and loading the SpotON Flow-Cell Priming and loading: The steps for priming and loading the SpotON Flow-Cell. Written instructions are given below. The library is loaded dropwise without putting the pipette tip firmly into the port. Take care to avoid introducing any air during pipetting. IMPORTANT Care must be taken when drawing back buffer from the flow-cell. The array of pores must be covered by buffer at all times.

17. Removing more than 20-30 µl risks damaging the pores in the array. After opening the priming port, check for small bubble under the cover. Draw back a small volume to remove any bubble (a few µls):

    Set a P1000 pipette to 200 µl

    Insert the tip into the priming port

    Turn the wheel until the dial shows 220-230 µl, or until you can see a small volume of buffer entering the pipette tip Visually check that there is continuous buffer from the priming port across the sensor array.

18. Prepare the Flow-Cell priming mix in a clean 1.5 ml Eppendorf DNA LoBind tube. Reagent Volume RRB 600 µl Nuclease-free water 600 µl Total 1200 µl Load 800 µl of the priming mix into the flow-cell via the priming port, avoiding the introduction of air bubbles. Wait for 5 minutes. IMPORTANT Thoroughly mix the contents of the RRB tube by vortexing or pipetting, and spin down briefly. Take 20 µl of the prepared RNA library and mix it with 17.5 µl of nuclease-free water.
19. In a new tube, prepare the library for loading as follows: Reagent Volume RRB 37.5 µl RNA library in nuclease-free water 37.5 µl Total 75 µl Complete the flow-cell priming:

    Gently lift the SpotON sample port cover to make the SpotON sample port accessible.

    Load 200 µl of the priming mix into the flow-cell via the priming port (not the SpotON sample port), avoiding the introduction of air bubbles. Mix the prepared library gently by pipetting up and down just prior to loading.

20. Add 75 µl of sample to the Flow-Cell via the SpotON sample port in a dropwise fashion. Ensure each drop flows into the port before adding the next. Gently replace the SpotON sample port cover, making sure the bung enters the SpotON port, close the priming port and replace the MinION lid.

Appendix I: RNA-seq data from Makenzie Cancer Research Group (MCRG) with RNA from a Control Lymphoblastoid Cell Line (08.111.0064)

Chromosome	Start	Stop	Splicing event	Gene	Total reads NMD treated	Total reads NMD untreated
12	6,643,736	6,643,975	FL 1-2	*GAPDH*	90053	38445
12	6,643,736	6,645,659	Delta 2	*GAPDH*	33	16
12	6,643,828	6,643,975	FL 1b-2	*GAPDH*	327	120
12	6,644,028	6,645,659	FL 2-3	*GAPDH*	174336	86509
12	6,644,028	6,646,085	Delta 3_4	*GAPDH*	9	4
12	6,645,721	6,645,849	Delta 3q (50nt)	*GAPDH*	631	240
12	6,645,760	6,645,849	FL 3-4	*GAPDH*	170204	94922
12	6,645,957	6,646,085	FL 4-5	*GAPDH*	156870	111090
12	6,646,177	6,646,266	FL 5-6	*GAPDH*	119868	98803
12	6,646,177	6,646,474	Delta 6	*GAPDH*	2	0
12	6,646,356	6,647,266	delta 6q delta 7_8	*GAPDH*	1474	1545
12	6,646,383	6,646,474	FL 6-7	*GAPDH*	179721	209243
12	6,646,383	6,646,503	exon 6 - mid exon 7	*GAPDH*	49	56
12	6,646,383	6,646,749	Delta 7	*GAPDH*	8	0
12	6,646,383	6,647,266	Delta 7_8	*GAPDH*	1	5
12	6,646,518	6,646,881	Mid exon 7- mid exon 8	*GAPDH*	8693	12273
12	6,646,557	6,646,749	FL 7-8	*GAPDH*	203336	274910
12	6,646,557	6,647,266	Delta 8	*GAPDH*	6	36
12	6,646,873	6,647,004	Delta mid exon 8	*GAPDH*	7095	10236
12	6,647,163	6,647,266	FL 8-9	*GAPDH*	152837	517299

17	41,197,820	41,199,659	FL 23-24	*BRCA1*	376	64
17	41,197,820	41,201,137	Delta 23	*BRCA1*	2	1
17	41,199,721	41,201,137	FL 22-23	*BRCA1*	345	55
17	41,199,721	41,203,079	Delta 22	*BRCA1*	5	1
17	41,201,212	41,203,079	FL 21-22	*BRCA1*	494	67
17	41,201,212	41,209,068	Delta 21	*BRCA1*	5	0
17	41,203,135	41,209,068	FL 20-21	*BRCA1*	504	63
17	41,209,153	41,215,349	FL 19-20	*BRCA1*	490	41
17	41,215,391	41,215,890	FL 18-19	*BRCA1*	509	40
17	41,215,969	41,219,624	FL 17-18	*BRCA1*	486	35
17	41,215,969	41,222,944	Delta 17	*BRCA1*	4	0
17	41,219,713	41,222,944	FL 16-17	*BRCA1*	466	46
17	41,223,256	41,226,347	FL 15-16	*BRCA1*	579	60
17	41,223,256	41,228,504	Delta 15	*BRCA1*	5	0
17	41,226,539	41,228,504	FL 14-15	*BRCA1*	522	57
17	41,226,539	41,234,420	Delta 14	*BRCA1*	1	2
17	41,228,632	41,234,420	FL 13-14	*BRCA1*	338	29
17	41,228,632	41,242,960	Delta 13	*BRCA1*	1	0
17	41,234,593	41,242,960	FL 12-13	*BRCA1*	446	28
17	41,243,050	41,243,451	FL 11-12	*BRCA1*	427	24
17	41,243,050	41,247,862	Delta 11	*BRCA1*	4	0
17	41,243,050	41,251,791	Delta 9_11	*BRCA1*	5	2
17	41,246,878	41,247,862	FL 10-11	*BRCA1*	287	6
17	41,246,878	41,249,260	Delta 10	*BRCA1*	2	0
17	41,246,878	41,251,791	Delta 9_10	*BRCA1*	158	8
17	41,246,878	41,256,138	Delta 8_20	*BRCA1*	1	0
17	41,247,940	41,249,260	FL 9-10	*BRCA1*	222	3

17	41,247,940	41,251,791	Delta 9	*BRCA1*	57	0
17	41,249,307	41,251,791	FL 8-9	*BRCA1*	226	3
17	41,251,898	41,256,138	FL 7-8	*BRCA1*	405	18
17	41,256,279	41,256,884	FL 6-7	*BRCA1*	452	13
17	41,256,974	41,258,472	FL 5-6	*BRCA1*	382	10
17	41,256,974	41,267,742	Delta 5	*BRCA1*	8	0
17	41,256,974	41,276,033	Delta 3_5	*BRCA1*	1	0
17	41,258,551	41,267,742	FL 3-5	*BRCA1*	404	11
17	41,258,551	41,276,033	Delta 3	*BRCA1*	8	0
17	41,267,797	41,276,033	FL 2-3	*BRCA1*	344	11
17	41,267,797	41,277,287	Delta 2	*BRCA1*	1	0
17	41,276,133	41,277,287	FL 1-2	*BRCA1*	86	3
17	56,770,150	56,772,291	FL 1-2	*RAD51C*	1147	336
17	56,770,150	56,774,053	Delta 2	*RAD51C*	1	0
17	56,772,551	56,774,053	FL 2-3	*RAD51C*	1229	189
17	56,772,551	56,780,556	Delta 3	*RAD51C*	65	0
17	56,772,551	56,811,478	Delta 3_8	*RAD51C*	1	0
17	56,774,221	56,780,556	FL 3-4	*RAD51C*	904	152
17	56,774,221	56,787,219	Delta 4	*RAD51C*	16	2
17	56,780,691	56,787,219	FL 4-5	*RAD51C*	1004	208
17	56,787,352	56,798,106	FL 5-6	*RAD51C*	1318	383
17	56,787,352	56,801,400	Delta 6	*RAD51C*	7	1
17	56,787,352	56,809,844	Delta 6_7	*RAD51C*	1	0
17	56,787,352	56,811,478	Delta 6_8	*RAD51C*	3	3
17	56,798,174	56,801,400	FL 6-7	*RAD51C*	1235	328
17	56,798,174	56,809,844	Delta 7	*RAD51C*	50	28
17	56,798,174	56,811,478	Delta 7_8	*RAD51C*	5	19

17	56,801,462	56,809,844	FL 7-8	*RAD51C*	1076	362
17	56,801,462	56,811,478	Delta 8	*RAD51C*	4	0
22	29,083,975	29,085,122	FL 14-15	*CHEK2*	461	756
22	29,085,204	29,090,019	FL 13-14	*CHEK2*	411	554
22	29,090,106	29,091,114	FL 12-13	*CHEK2*	547	666
22	29,091,231	29,091,697	FL 11-12	*CHEK2*	394	455
22	29,091,231	29,095,825	Delta 10_11	*CHEK2*	7	0
22	29,091,862	29,092,888	FL 10-11	*CHEK2*	449	291
22	29,091,862	29,095,825	Delta 10	*CHEK2*	21	17
22	29,092,976	29,095,825	FL 9-10	*CHEK2*	412	240
22	29,095,926	29,099,492	FL 8-9	*CHEK2*	287	177
22	29,095,926	29,105,993	Delta 8	*CHEK2*	109	34
22	29,099,555	29,105,993	FL 7-8	*CHEK2*	231	148
22	29,106,048	29,107,896	FL 6-7	*CHEK2*	404	245
22	29,108,006	29,115,382	FL 5-6	*CHEK2*	355	166
22	29,108,006	29,120,964	Delta 5	*CHEK2*	28	15
22	29,108,006	29,121,230	Delta 4_5	*CHEK2*	12	0
22	29,115,474	29,120,964	FL 4-5	*CHEK2*	232	139
22	29,121,113	29,121,230	FL 3-4	*CHEK2*	418	188
22	29,121,356	29,130,390	FL 2-3	*CHEK2*	420	128