Analysis of hereditary and putative breast cancer susceptibility genes in *BRCA1* and *BRCA2* mutation-negative individuals

Chloé Anne Louise Thompson-Peach
BSc(Hons), BMedSc

Department of Molecular Medicine and Pathology

College of Medicine and Public Health

Flinders University

March 2020

Thesis is submitted to the fulfil the requirements for the degree of

Doctor of Philosophy

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and

2. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed: Chloé Anne Louise Thompson-Peach Date: 16 October 2019

Table of Contents

List of Figures

List of Tables

Acknowledgements

The work described in this thesis was funded by the Flinders Foundation Seeding Grant, Lyn Wrigley Research and Development Fund and a Flinders University Research Scholarship.

First and foremost, I would like to express my extreme gratitude to my primary supervisor Associate Professor Karen Lower. Without your constant support, guidance, advice and encouragement, this work would not have been possible. Thank you for sparking my love for genetics and encouraging me to pursue it. Thank you for your continual enthusiasm and always pushing me to look beyond what the computer has said and for a deeper understanding behind the results. Thank you for your belief in my abilities and the enormous amounts of time and effort you put into me, my project and my thesis. I cannot thank you enough. To my co-supervisors, Associate Professor Michael Michael and Dr. Scott Grist, thank you for the help you provided in supervising me. To Michael, thank you for welcoming me into your lab, encouraging me to try new things and venture into the world of CRISPR. To Scott, thank you for your help in obtaining patient samples, assistance in troubleshooting issues with my project and for the constant supply of polymerase.

Generous thanks to all the patients who consented to the generous use of their DNA for this study. Without you, this work would not have been possible.

Thank you to everyone in the Department of Molecular Medicine and Pathology for your support and encouragement. In particular, many thanks to our fearless leader, Associate Professor Bryone Kuss. Thank you for welcoming me into the department and guiding me in the right direction. Thank you for your encouragement and support in meetings and pushing me to think outside the scope of my project to the larger applicability of the work I was doing and finding time to read drafts of work. Many thanks to Di Luke for her constant cheerful demeanour, encouragement and help with all things ordering and admin related, but also just general life advice. Thank you to those in the MRD research group, particularly Brad Budgen and Dr. Paul Bartley for their welcoming nature, help whenever I was stuck and their assistance in all things PCR and DNA integrity related.

Many thanks to Dr. Binoy Appukutan for making every day in the lab an interesting experience, your stories with multiple tangents and your constant encouragement and enthusiasm for science. Thanks for helping me troubleshoot my experiments, taking every possible variable into

consideration until we had 'Binoy-ed it' and made it work! Thanks for the many chats we've had over a gin or two and your constant reassurance when I was feeling defeated and lost throughout my PhD. Many thanks to Dr. Stephen Gregory for your rapid and insightful feedback on thesis chapters. I would also like to thank Dr. Lauren Thurgood for teaching me all things proteomics, being a constant source of support and teaching me what a life in science is really like. Your passion for science, ability to smash out experiments and love for educating the next generation of scientists amazes me. Thank you for making time to read chapter drafts and teach me new skills amongst your many students and complex experiments. Your friendship and encouragement has been invaluable.

To my PhD Wife, (almost Dr.) Lara Escane. Thank you for your endless support throughout this time. Thank you for listening to my rants, helping me through my frustrations, helping me troubleshoot my problems or just being a general sounding board. Thank you for being stuck in this with me. Without you, I don't know if I would have been able to get through. Thank you for the encouragement and helping me believe in myself when my Imposter syndrome was at an all-time high and for teaching me all things cell related! To the other PhD students that have shared in this journey with me, especially Dr. Cuc Do, Dr. Anya Hotinski, Dr. Katherine Morel and Saira Ali thanks for the reassurance that it does get better and the experience was just as difficult for you too.

The work detailed within this thesis would not have been possible without significant help from a number of people. I would like to thank the Genetic Pathology team at SA Pathology, particularly Dr. Andrew Dubowsky, Kristy Nichol, Dr. Melanie Hayes and Duncan Holds for the patient samples and the provision of all PCR reagents required throughout this project. Special thanks to Oliver Van Wageningen for all of your help with Sanger sequencing and fragment analysis throughout this project. To the researchers in the gene expression lab, particularly Marie Lowe, Kym McNicholas and Saira Ali, thank you for your training and assistance with all the bacterial and plasmid components of my project. I would like to thank Dr. Renee Smith and Letitia Pimlott from the Flinders Genomics Facility for their support and assistance throughout my project. Thank you for your expertise, patience and friendship. Thank you to Isabell Bastian for taking time out of your busy schedule to irradiate my cells and many thanks to Dr. Shari Javadiyan for all of your assistance and guidance with all things MPS and Ion Torrent related. Thank you for making time for me and answering my many questions. Many thanks to the Life Technologies Field application specialists, specifically Dr. Bennett Shum , Dr. Fabrice Odifrey and Dr. Dale Watkins. Thanks for your expertise,

help and support in troubleshooting multiple library preparations and data analysis issues. Many thanks to Associate Professor Richard Allcock from the LotteryWest State Biomedical Genomics Facility, University of Western Australia for your assistance in sequencing the remaining required samples when all other Ion Torrent Machines were decommissioned. Thanks to Dr. Lesley Ann Gray from Australian Genome Research Facility (Melbourne) for carrying out the data analysis for the Tri-Pool-Seq work. Finally, I would like to thank the Medical Science placement students that carried out the *UIMC1* fragment analysis work, Leah Cameron, Bronte Hyams and Simon Mandel. Thank you for your assistance in this section of my project, sparking my love for teaching and your passion for science.

To the teaching and technical staff in Biology, particularly Dr. Masha Smallhorn, Narelle Hunter, Dr. Jeanne Young, Dr. Jess Clayton, Dr. Lui Fei Tan, Dr. Lucy Clive, Dr. Sam Davies and Tania Neville. Thank you for your constant support, encouragement and friendship. Thank you for all the chats we've had about the PhD experience, life in general and just being there when I needed someone. Thank you for encouraging me, helping spark a passion for education I never knew existed, teaching me new skills and assisting me to become the best educator that I could be.

My immense gratitude to my friends and family for their support and understanding throughout this whole process. I love you all so much and I cannot thank you enough. To Mum, thank you for all the encouragement, unwavering support and belief in me. To Ashley, thank you for constantly building me up when I was feeling down and defeated by everything. To my Med Sci friends and honours students past and present, particularly (almost Dr.) Madi Oprea and Dr. Stuart Denham, thank you for your friendship and supporting me through this trying time. Thank you for reassuring me, helping me believe in myself and telling me that I can get through it. To my Hereford Family, especially Erica Langley, Sam Oldfield, Lauren Angeli, Alan Ryan and Kimberly Arnold, thank you for being so understanding and supportive. Thank you for the encouragement and the many brunches, steaks, gins and wines that we have shared to help get me through this process. To my amazing husband Sam, I couldn't have done this without you, especially in that last year. Thank you for keeping me sane and being the person to motivate, help me and believe in me when I needed it most. I cannot thank you enough for your patience, understanding, encouragement and support. Finally, to the resilient and inspiring people in my life who have struggled with and still fight cancer every day, Liam and Dee, this is for you.

Publications and presentations derived from this thesis

Posters:

Thompson-Peach, CAL, Michael, MZ, Grist, SA, Kuss, BJ and Lower, KM. (2018) Analysis of hereditary and putative breast cancer susceptibility genes in *BRCA1* and *BRCA2* mutation-negative individuals. FCIC research week conference, Adelaide, September 2018

Thompson-Peach, CAL, Michael, MZ, Grist, SA, Kuss, BJ and Lower, KM. (2017) Next generation sequencing analysis of 51 genes of interest in *BRCA1* and *BRCA2* mutation-negative individuals with a familial history of breast cancer. ComBio, Adelaide Convention Centre, Adelaide October 2017

Thompson-Peach CAL, Michael, MZ, Grist, SA, Kuss, BJ and Lower KM. (2015) NGS analysis of 51 genes in *BRCA1* and *BRCA2* mutation-negative individuals with a familial history of breast and ovarian cancer. Australian Society for Medical Research (ASMR), SA Branch, Adelaide Convention Centre, Adelaide, June 2015.

Thompson-Peach CAL, Michael, MZ, Grist, SA, Kuss, BJ and Lower KM. (2014) NGS analysis of 51 genes in *BRCA1* and *BRCA2* mutation-negative individuals with a familial history of breast and ovarian cancer. Australian Society for Medical Research (ASMR), SA Branch, Adelaide Convention Centre, Adelaide, June 2014.

Presentations:

Thompson-Peach, CAL, Lower KM (2019), Analysis of breast cancer susceptibility genes in *BRCA1* and *BRCA2* mutation-negative individuals with familial breast cancer. College of Medicine and Public Health, Medicine and Bioscience seminar series, Flinders Centre for Innovation in Cancer, April 2019

Thompson-Peach, CAL, Michael MZ, Grist SA and Lower KM. (2018), Analysis of *BRCA1/2* related breast cancer genes in mutation-negative individuals; a South Australian perspective. Human Genetics Society of Australasia, International Convention Centre, Sydney, August 2018

Publications:

Thompson-Peach CAL, Braun, SE, Grist SA, Michael, MZ, Lower, KM, Tri-Pool-Seq analysis in inherited breast cancer *(in preparation)*

Thompson-Peach CAL, Braun, SE, Grist SA, Michael, MZ, Lower, KM, Analysis of *BRCA1/2* related breast cancer genes in mutation-negative individuals; a South Australian perspective ***(in preparation)***

Abstract

Breast cancer is the most common cancer affecting Australian women, with many affected individuals exhibiting a strong family history of the disease. Whilst inherited mutations in *BRCA1*, *BRCA2* and additional susceptibility genes account for approximately 30% of familial breast cancer cases, the underlying cause in the remaining 70% is unknown, suggesting that additional breast cancer susceptibility genes exist. We hypothesised that mutations within genes that play a role in the DNA damage repair and checkpoint control pathways may be involved in predisposing families to inherited breast cancer.

In order to test our hypothesis Ion Torrent Massively Parallel Sequencing and a custom targeted panel were used to sequence 51 genes of interest in a cohort of *BRCA1/2* mutation-negative individuals with familial breast cancer. The gene panel consisted of 19 known breast cancer susceptibility genes (diagnostic genes) and 32 genes which play integral roles in the DNA damage repair and cell cycle control pathways and therefore are potentially involved in the development of breast cancer (discovery genes). For this study, a bespoke bioinformatics pipeline was developed for the analysis of Ion Torrent data generated from a cohort of *BRCA1/2* mutation-negative individuals from South Australia. A novel three-dimensional pooling strategy (Tri-Pool-Seq) was piloted for the identification of rare variants within the patient cohort, however this failed to identify known sequence changes and therefore was not extended to the full cohort.

From the individual sequencing and analysis of patients, an average of 125 variants were identified in each sample, with rare variants analysed further. In total 166 rare variants were identified which were predicted to alter gene transcription or translation; of these 82 variants were identified as being potentially pathogenic. Moreover, a known pathogenic truncation mutation was identified in *PALB2* in 2 individuals.

CRISPR/Cas9 was used to functionally validate a *UIMC1* polymorphism identified in 2 patients in an attempt to establish the role of this gene in cancer development. This study indicated that cells lacking functional UIMC1 demonstrated an increased sensitivity to ionising radiation, resulting in an increase in cell death and a reduced capacity to repair DNA double stranded breaks. These results indicate that a loss of UIMC1 may play an important role in the development of hereditary breast cancer, through the loss of vital DNA damage repair capabilities and dysregulation of cell growth.

Overall, this research has the potential to provide much needed diagnostic information for the identification of mutations resulting in familial breast cancer, and to identify novel breast cancer genes.

The work carried out within this thesis provides further evidence that additional genes are involved in the development of hereditary breast cancer. This South Australian population-based analysis of *BRCA1* and *BRCA2* mutation-negative individuals has resulted in the identification of both pathogenic disease causative mutations, and a potentially novel gene involved in cancer development.

Abbreviations

°C	degrees Celsius
γH2AX	histone H2AX phosphorylated on serine 139
µg	microgram/micrograms
µL	microlitre/microlitres
µM	micromolar/micromole
3'UTR	3 prime untranslated region
3D	3 dimensional
5'UTR	5 prime untranslated region
A	ampere
AA	amino acid
ATF1	Activating Transcription Factor 1
ATM	Ataxia telangiectasia mutated
ATP	adenosine triphosphate
BARD1	BRCA1 associated RING domain 1
BASC	BRCA1-associated genome surveillance complex
BED	browser extensible data
BLAST	basic local alignment search tool
BLAT	BLAST-like alignment tool
BOADICEA	breast and ovarian analysis of disease incidence and carrier estimation algorithm
bp	base pair
BPE	bovine pituitary extract
BRCA1	Breast Cancer 1 (early onset)
BRCA2	Breast Cancer 2 (early onset)
BRCC3	BRCA1/BRCA2 containing complex, subunit 3
BRCT	BRCA1 C Terminus
BRIP1	BRCA1 interacting protein 1
BSA	bovine serum albumin
Cas9	CRISPR associated 9
CDH1	Cadherin 1
CDKN1A	Cyclin dependent kinase inhibitor 1A
CDKN2A	Cyclin dependent kinase inhibitor 2A
cDNA	complementary DNA
CHEK1	Cell cycle checkpoint Kinase 1
CHEK2	Checkpoint Kinase 2
CKS1B	CDC28 Protein Kinase 1B
cm	centimetre/centimetres
CO_2	Carbon Dioxide
COSMIC	Catalogue of Somatic Mutations in Cancer
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNAs
C_q	cycle quantification
DAPI	Diamidino-2-phenylindole dihydrochloride
dbSNP	Single nucleotide polymorphism database
DCIS	Ductal carcinoma *in situ*
DEPC	diethyl pyrocarbonate
DMEM	Dulbecco's modified medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DOX	Doxorubicin Hydrochloride

DSB	double stranded break
dsDNA	double stranded DNA
DTT	dithiothreitol
E2F1-6	E2F Transcription Factor (1-6)
EDF	extended depth of field
EDTA	ethylenediaminetetraacetic acid
EP300	E1A-Binding Protein, 300-KD
ER	oestrogen receptor
ESE	exon splicing enhancer
ESS	exon splicing silencer
FAM	fluorescein amidite
FAM175A	Family with sequence similarity 175, member A
FCS	foetal calf serum
FGF	Flinders genomics facility
FH9	Flinders Hospital 9 immortalised B cell line
FMC	Flinders Medical Centre
FS	frameshift
g	gram/grams
G1 Phase	Gap 1 Phase (Cell cycle)
G2 Phase	Gap 2 Phase (Cell cycle)
GA-100	gentamicin/amphotericin solution
GADD45A	Growth arrest and DNA damage-inducible gene, alpha
GD	Grantham Deviation
gDNA	genomic DNA
GFP	green fluorescent protein
gnomAD	Genome Aggregation Database
GRCh37	Genome Reference Consortium Human Build 37
GV	Grantham Variation
GWAS	genome wide association studies
Gy	gray
H2AX	H2A histone family member X
HDR	homology directed repair
hEGF	human epidermal growth factor
HEK293	Human Embryonic Kidney 293 immortalised cell line
HER2	human epidermal growth factor receptor 2
HEX	hexachloro-fluorescein
hg19	Homo sapiens reference genome assembly (February 2009 build)
HGMD	human gene mutation database
HLTF	Helicase-like transcription factor
HMMR	Hyaluronan-mediated motility receptor
HPLC	High performance liquid chromatography
HR	homologous recombination
IDC	invasive ductal carcinoma
IDT	Integrated DNA Technologies
IGV	Integrative genomics viewer
Indel	insertion and/or deletion
IR	Ion Reporter
IR	ionising radiation
ISP	Ion sphere particle
ISX	Image Stream X
IVA	Ingenuity variant analysis
KAT2B	K(Lysine) Acetyltransferase 2B

kDa	kilodaltons
KeV	kiloelectronvolts
kg	kilogram
KO	knockout
L	Litre/litres
LB	Luria Broth/Lysogeny Broth
lncRNA	Long non-coding RNA
M Phase	Mitosis phase (Cell cycle)
mA	milliampere
MAF	minimum allele frequency
MEBM	Mammary Epithelial Cell Growth Basal medium
MIM	Mendelian inheritance in Man
min	minute/minutes
miRNA	microRNA
mL	millilitre/millilitres
MLPA	Multiple ligation-dependent probe amplification
mM	millimole/millimolar
mm	Millimetre/millimetres
MNP	multiple nucleotide polymorphism
MNV	multiple nucleotide variant
MPS	massively parallel sequencing
MRE11A	Meiotic recombination 11, Homolog A (*S.Cerevisiae*)
MRI	magnetic resonance imaging
mRNA	messenger RNA
MSS	Manchester scoring system
MW	molecular weight
NBN	Nibrin
NCBI	National Centre for Biotechnology Information
NEB	New England Biolabs
ng	Nanogram/nanograms
NGS	next generation sequencing
NHEJ	non-homologous end joining
NLS	nuclear localisation signal
nm	nanometre/nanometres
nM	nanomole/nanomolar
NQO2	Nicotinamide adenine dinucleotide phosphate (NADPH) dehydrogenase, Quinone 2
O/N	overnight
OB	oligosaccharide binding fold
OCCR	ovarian cancer cluster region
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PALB2	Partner & localiser of BRCA2
PAM	Protospacer-adjacent motif
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PGM	personal genome machine
PKMYT1	Protein Kinase Membrane Associated Tyrosine/Threonine 1
pM	Picomole/picomolar
PNK	polynucleotide kinase
PolyPhen-2	Polymorphism Phenotyping-2
PR	progesterone receptor
PRKDC	Protein Kinase, DNA Activated Catalytic Subunit

PROVEAN	protein variation effect analyser
PTEN	Phosphatase and tensin homolog
Q20	quality score; incorrect base call probability of 1 in 100
QC	quality control
qPCR	quantitative PCR
RAD50	RAD50 Homolog (*S.Cerevisiae*)
RAD51	RAD51 Homolog (*S.Cerevisiae*)
RAD51C	RAD51 Homolog , Paralog C (*S.Cerevisiae*)
RAD51D	RAD51 Homolog , Paralog D (*S.Cerevisiae*)
RBL1	Retinoblastoma-like 1
RBL2	Retinoblastoma-like 2
RFC(2-5)	Replication Factor C, Subunits 2-5
RING Domain	really interesting new gene domain
RNA	ribonucleic acid
RNP	ribonucleoprotein complex
RPA1	Replication Protein A1
rpm	revolutions per minute
RPRM	Reprimo
RPS6KA1	Ribosomal Protein S6 Kinase 1
RT	room temperature
RT-qPCR	reverse transcription quantitative polymerase chain reaction
S Phase	DNA Replication (cell cycle)
SA	South Australia
SABC	South Australian Breast Cancer Patient, de-identified
SAP	shrimp alkaline phosphatase
SD	standard deviation
SDS	sodium dodecyl sulphate
SFN	Stratifin
sgRNA	single guide RNA
SIFT	sorting intolerant from tolerant
SLC19A1	Solute carrier family 19 (folate transporter) member 1
SMARCD2	SWI/SNF-Related matrix-associated, actin-dependent regulator of chromatin, Subfamily D, Member 2.
SNP	single nucleotide polymorphism
SOC	super optimal broth with catabolite repression
spCas9	*Streptococcus pyogenes* Cas9
SS	Sanger sequencing
ssDNA	single stranded DNA
ssODN	single stranded oligodeoxynucleotide
STS	sequence-tagged site
SUMO	Small Ubiquitin-like Modifiers
TAE	tris-acetate-EDTA
TBS	tris buffered saline
TE	tris-EDTA buffer
TP53	Tumour Protein p53
tracrRNA	trans-activating crRNA
TSG	tumour suppressor gene
U	unit/units
UCSC	University of California Santa Cruz
UDG	uracil-DNA glycosylase
UIM	ubiquitin interaction motifs
UIMC1	Ubiquitin Interaction Motif-Containing Protein 1

UV	ultra-violet
UWA	University of Western Australia
V	volts
v/v, w/v	volume/volume, weight/volume
VUS	variant of uncertain significance
WEE1	WEE1 Tyrosine Kinase
WES	whole exome sequencing
WGS	whole genome sequencing
X	times
x *g*	times gravity
XRCC2	X-ray repair complementing defective repair in Chinese hamster cells 2
ZFN/ZFD	Zinc Finger/Zinc Finger Domain

Chapter 1:
Introduction

1.1 Breast Cancer

Cancer is a disease where uncontrolled cell growth arises secondary to key underlying driver events. These events result in the cell's ability to evade cell death, sustain chronic proliferation, enable replicative immortality, elude growth suppressors, obtain sustenance and vascularisation through angiogenesis and the ability to invade and metastasise (Hanahan and Weinberg, 2011). Often, cancer arises due to genetic aberrations within vital genes required for cell cycle regulation, controlled cell growth and DNA damage repair. Mutations within these vital genes often results in genomic instability and therefore high mutability, but also leads to uncontrolled cell proliferation, destruction of healthy neighbouring cells and invasion of surrounding tissues and organs (Hanahan and Weinberg, 2011). Mutations can arise sporadically from errors in DNA replication or from external factors such as exposure to carcinogens and lifestyle factors (i.e. diet and alcohol consumption), or mutations can be inherited.

Cancer affects many different organs and body systems, with breast cancer identified as the second most common cancer affecting women worldwide (Torre *et al.*, 2015). Cancers originating from the epithelium are termed carcinomas (specifically adenocarcinomas), with those arising from the breast epithelium comprising a highly heterogeneous group of tumours, which can differ significantly based on age of onset, clinical features, and the histological characteristics. Additionally, the genetic context associated with breast cancer development plays a significant role in treatment options and prognosis. The majority of genetic changes identified in cancer, including breast cancers tend to fall into two categories: loss of function mutations within tumour suppressor genes (TSGs) and gain of function mutations within proto-oncogenes.

Mutations in both TSGs and proto-oncogenes have both been shown to play a pivotal role in carcinogenesis, with mutations often initiating tumour development and further driving tumour progression (Hanahan and Weinberg, 2011). TSGs function to negatively regulate cell growth and proliferation and maintain homeostasis. Therefore, the loss of function of TSGs within these pivotal pathways enables cancers to sustain cell growth, evade growth suppression and resist cell death. Proto-oncogenes function to control cell growth and proliferation, and when mutated, can function as a cancer promoting oncogene due to dysregulation of vital cellular control. This control is maintained by various negative feedback loops that function to diminish various types of cell signalling and maintain homeostatic regulation, which has been demonstrated by numerous studies (Amit *et al.*, 2007, Mosesson *et al.*, 2008, Wertz and Dixit, 2010, Hanahan and Weinberg, 2011). It

has been shown that defects within these proto-oncogenes can enhance proliferative signalling. Mutations within genes of an oncogenic nature have been shown to correspond to an increase in proliferation of cancer cells and therefore, an increase in tumour growth and progression. Due to the vital function of both TSGs and proto-oncogenes in the control of cell growth, cell death and maintenance of genomic integrity, mutations within both of these classes of genes are likely to play a role in breast cancer development.

Breast cancer is the most common cancer in women and the second most common malignancy in Australia with 16,753 new cases being diagnosed in 2014 (Australian Institue of Health and Welfare, 2016). Breast cancer affects 1 in 8 women by age 85 (Australian Institute of Health and Welfare, 2013) and, in 2014, accounted for 25.4% of all cancers in women worldwide (International Agency for Research on Cancer, 2014). On a positive note, the mean five-year survival rate for those diagnosed with breast cancer has increased from 72% in 1987 to 89% in 2014 (Cancer Australia, 2014, National Breast Cancer Foundation, 2014). This improvement in survival rate is in part attributed to early detection of breast cancer through the implementation of regular mammogram screening. However, the incidence of breast cancer is increasing and it is anticipated that by 2020 approximately 20,000 women will be diagnosed with breast cancer annually in Australia (Australian Institue of Health and Welfare, 2016). This has been attributed to several factors, one of which is the introduction of the BreastScreen Australia Program (Australian Government, 2015). This service provides increased breast cancer surveillance and therefore detection of breast cancer at earlier stages and has contributed to the apparent increase in cancer incidence. Additionally, lifestyle factors such as hormone replacement therapy and alcohol consumption, in addition to the aging population, are also playing a role in the observed increase in breast cancer incidence.

Symptoms of breast cancer commonly include physical changes in the breast such as the development of a lump, breast pain, or changes to the nipple or skin (Australian Institue of Health and Welfare, 2016). Most breast cancers originate in the cells lining the ducts or within the lobules (Sharma *et al.*, 2010). Cancers also develop in other mammary tissues or progress to form tumours within lymph nodes following metastasis (**Figure 1.1**). Breast cancers are most commonly identified as ductal carcinoma *in situ* (DCIS) or invasive ductal carcinoma (IDC), both of which originate from the milk ducts (Tamimi *et al.*, 2008). DCIS is the most common form of non-invasive cancer, and whilst these tumours are initially benign, they have the potential to become invasive and malignant if not treated (Sharma *et al.*, 2010). IDC is the infiltrative and malignant proliferation of cells from

within the milk ducts into the surrounding breast tissue. These carcinomas can invade the lymph nodes and spread to other regions of the body. IDC is the most commonly identified type of breast cancer, accounting for 80% of all breast cancer diagnoses. Similarly, malignancies within the milk producing lobules are also identified as either invasive or *in situ* (Wellings and Jensen, 1973, Cristofanilli *et al.*, 2005)

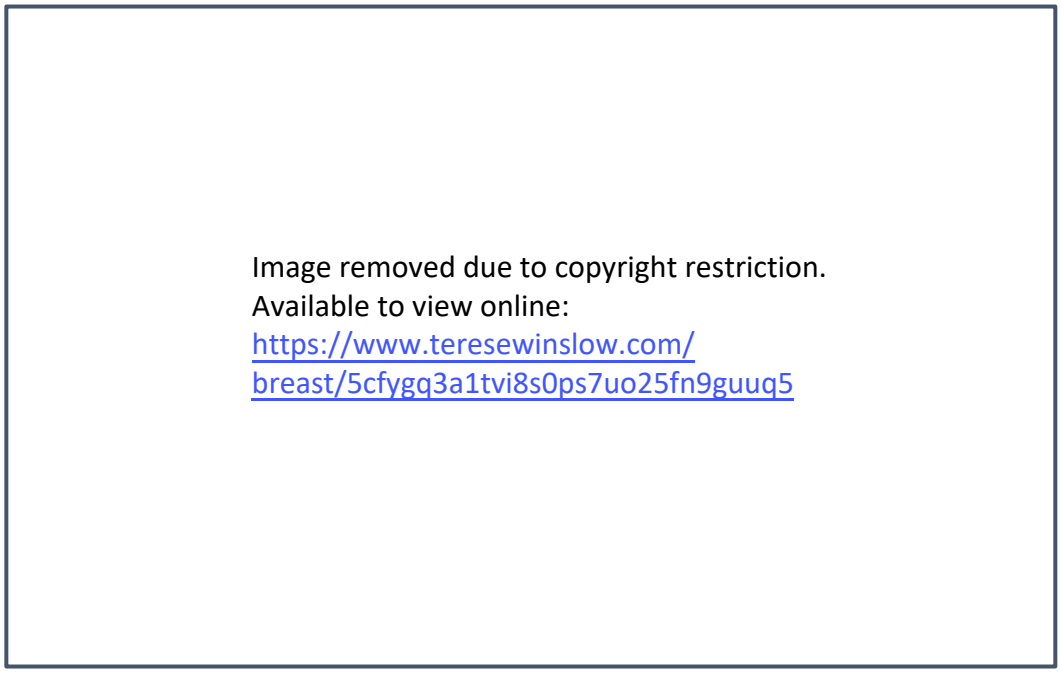


Figure 1.1: Anatomy of the female breast. Majority of breast cancers originate in the milk ducts or the lobules (Winslow, 2011).

Diagnosis of breast cancer typically involves a clinical breast examination and breast imaging including mammogram, MRI or ultrasound (Fuller *et al.*, 2015). Biopsies may also be taken for histological examination and are frequently performed under ultrasound guidance. If breast cancer is confirmed, treatment options can involve surgery (whether it be a wide local excision, mastectomy or prophylactic total mastectomy), radiotherapy, chemotherapy, hormone therapy or targeted therapies. Treatment approach is determined by various factors that include the stage, type and location of the breast cancer, severity of symptoms and the general health of the affected individual (Miller *et al.*, 2016). The genetic basis of breast cancer development also plays a role in determining possible treatment options (as discussed further in **Section 1.5**).

1.1.1 Sporadic Breast Cancer

Sporadic cases of breast cancer account for approximately 90% of all reported Australian breast cancer cases per year (van der Groep *et al.*, 2006). Sporadic cancers typically have no known

hereditary links but are likely a consequence of cumulative acquired mutations within somatic cells. Whilst sporadic cases of breast cancer are not the focus of this research project, understanding the pathways and proteins that play a role in sporadic breast cancer development remain relevant to understanding the pathology of breast cancer. Often, sporadic cancers are attributed to an accumulation of acquired mutations within key regulatory genes. Most frequently, sporadic tumours are attributed to the activation and/or over-expression of oncogenes which play a role in cell proliferation and tumour growth. The activation of oncogenes such as *MYC* (MYC-proto oncogene, BHLH Transcription factor), *CCND1* (Cyclin D1) and *ERBB2* (Erb-B2 receptor tyrosine kinase 2) have been shown to be crucial events in breast carcinogenesis. The resulting increase in cell proliferation and uncontrolled cell growth leads to subsequent tumour formation (Mitrunen and Hirvonen, 2003, Kenemans *et al.*, 2008)

There is a significant difference in the key mutagenic events that occur within sporadic and familial breast cancers. While sporadic cancers are most commonly attributed to an acquired accumulation of mutations within oncogenes, familial cancers are often the result of germline mutations within TSGs. Often, cancers arising from mutations within TSGs require both alleles to be mutated or deleted, which is not commonly observed in individuals with sporadic cancer. However, individuals who have already inherited a mutated allele (as observed with familial cancers) are more likely to acquire a second mutation, which has the ability to knockout function of the key TSG (discussed in more detail in **section 1.4**)

1.1.2 Familial Breast Cancer

Familial cases of breast cancer are associated with the germline inheritance of a pathogenic variant affecting the function of a gene or multiple genes involved in pathways such as cell checkpoint control (Xu *et al.*, 1999), DNA damage response (Wang *et al.*, 2000) and transcriptional regulation (Starita and Parvin, 2003). Familial breast cancers, which typically involve mutations within TSGs in these pathways, accounts for approximately 10% of all breast malignancies (Liebens *et al.*, 2007).

The two most common TSGs found to be mutated in familial breast cancer are the breast cancer susceptibility genes *BRCA1* and *BRCA2* (Miki *et al.*, 1994, Wooster *et al.*, 1995). An inherited pathogenic mutation within either of these genes results in a significantly increased lifetime risk of breast cancer; 55-85% for *BRCA1* mutations and 35-60% for *BRCA2* mutations, compared with a population risk of approximately 10% (Brose *et al.*, 2002, Thompson and Easton, 2002, Antoniou *et*

al., 2003, King *et al.*, 2003). Individuals from families with a strong history of breast cancer that meet the diagnostic criteria can undergo mutational screening within the *BRCA1/2* genes. Identification of a pathogenic mutation in individuals within these families is important as it enables access to a range of additional surveillance opportunities, and, if chosen, prophylactic surgical interventions. Unfortunately mutations in these genes only account for approximately 20% of familial breast cancer cases (Turnbull and Rahman, 2008, Shiovitz and Korde, 2015). Whilst several other moderate-risk breast cancer predisposition genes have been identified, the underlying cause of more than 70% of familial breast cancer cases is still unknown (Turnbull and Rahman, 2008, Shiovitz and Korde, 2015). Whilst these cases may contain undetected pathogenic mutations within the *BRCA* genes, it is also highly likely that additional breast cancer susceptibility genes may exist.

1.2 Ovarian Cancer

In addition to familial breast cancer, mutations within *BRCA1* and *BRCA2* have also been identified to play a role in the development of ovarian cancer. After uterine cancer, ovarian cancer is the second most commonly diagnosed gynaecological cancer in Australian women. However, it is the deadliest in terms of mortality rate (Sankaranarayanan and Ferlay, 2006). Ovarian cancer is the fourth most common cause of cancer mortality in women and although less frequent than breast cancer, it is rapidly fatal, a characteristic which is often attributed to poor detection rates (Lengyel, 2010). Unfortunately, due to the lack of signs and symptoms in the early stages of disease, combined with an absence of screening tests, most ovarian cancer cases remain undiagnosed until the advanced stages. The majority (>90%) of malignant ovarian cancers are epithelial (Ramus and Gayther, 2009). These ovarian cancers shed epithelial cells into the fluid of the abdominal cavity, facilitating the implantation of the tumour cells within other peritoneal structures, including the uterus, bladder and bowel. More than 60% of women presenting with ovarian cancer are diagnosed at stage III or IV, indicating that the cancer has already spread beyond the ovaries (Sood *et al.*, 2001). Mortality in these women is high, with a 5-year survival rate of approximately 43% (Ramus and Gayther, 2009).

Germline mutations within *BRCA1* and *BRCA2* confer a high lifetime risk of ovarian cancer and mutations within these genes represent the most significant and well characterised risk factors for ovarian cancer (Ramus and Gayther, 2009). The risk of developing ovarian cancer is 40-53% with *BRCA1* mutations and 20-30% with *BRCA2* mutations (Ford *et al.*, 1998, Antoniou *et al.*, 2002). Other genes such as the mismatch repair genes *MLH1* (MutL homolog 1) and *MSH2* (MutS homolog 2), are

associated with an increased risk of ovarian cancer, but not breast cancer (Bonadona *et al.*, 2011). Additionally, mutations in *MSH6* (MutS homolog 6) and *PMS2* (PMS1 homolog 2, mismatch repair system component) are associated with the development of Lynch syndrome which is associated with an increased risk of hereditary cancers, including both ovarian and breast cancers (Roberts *et al.*, 2018).

1.3 Genetic Risk Prediction

Whilst breast cancer can cluster with other phenotypic features when part of a syndrome, such as in Cowden syndrome, there is no phenotype associated with carrying a pathogenic *BRCA* mutation until the onset of breast and/or ovarian cancer. Therefore, the likelihood of identifying an individual carrying a mutation within one of these predisposition genes is based largely on family history. Over the past two decades, there have been several statistical and empirical models designed and validated for the assessment of breast cancer risk. The cohort selected for this study have all been assessed through the Manchester scoring system (MSS).

The MSS is used to determine the likelihood of identifying a *BRCA1/2* mutation in a given individual (Evans *et al.*, 2004, Evans *et al.*, 2005). The MSS involves assessing both the maternal and paternal lineages and assigning scores for each affected individual within the family (score criteria is outlined in **Section 2.1.1**). The system takes into account types of cancers in the family, including breast (both male and female), ovarian, prostate and pancreatic, and the age of onset. This model has been validated in multiple datasets and has been shown to perform well in comparison to other established models (Evans *et al.*, 2004, Amir *et al.*, 2010). The main advantage associated with the MSS is its simplicity. While some of the manual and other tabular models (eg. Couch Model, Myriad tables) are also relatively easy to use, they often ignore important familial information. Conversely, the computer models (such as BOADICEA and BRCAPRO) are very time consuming to carry out and can be difficult to manipulate (Antoniou *et al.*, 2004, Antoniou *et al.*, 2008). Therefore the MSS is most often utilised by clinicians to determine if an individual would benefit from *BRCA1/2* analysis.

1.4 Tumour Suppressor Genes

Mutations within TSGs, such as *BRCA1, BRCA2, TP53* (tumour protein 53) and *ATM* (Ataxia telangiectasia mutated*)* have been shown to be involved in breast cancer. TSGs regulate the proliferation of normal cells and play an important role in cell cycle arrest (Suter and Marcum, 2007). Loss of TSG function results in uncontrolled cell proliferation and often results in tumour formation.

TSGs have been divided into two major categories; gatekeeper and caretaker genes (Kinzler and Vogelstein, 1997). Gatekeeper genes identified as TSGs are responsible for the control or promotion of cell death, such as *TP53* and *PTEN* (Oliveira *et al.*, 2005). These genes directly inhibit tumour growth or promote cell death and as such, the inactivation of these genes may directly contribute to the formation of cancers and their progression. Caretaker genes, such as *MLH1* and *MSH2*, encode products necessary for genome stabilisation (Hickson, 2003). The inactivation of a caretaker gene leads to genetic instability, resulting in an accumulation of uncorrected mutations throughout the genome. The onset of tumourigenesis as a result of mutation within a caretaker gene can progress rapidly due to an accelerated rate of mutation in other genes directly involved in regulating cell proliferation or apoptosis. Both *BRCA1* and *BRCA2* have been categorised as caretaker genes (Oliveira *et al.*, 2005).

1.4.1 BRCA1

The existence of the *BRCA1* gene (Breast Cancer 1, early onset; MIM 113705) was first identified through linkage studies which found that mutations within 17q12-21 were associated with inherited breast and ovarian cancer (Hall *et al.*, 1990). Further analysis through positional cloning resulted in the identification of the *BRCA1* gene within this region (Miki *et al.*, 1994). *BRCA1* encodes an 1863 amino acid protein and is comprised of 24 exons, located on chromosome 17q21 (**Figure 1.2**). The exon boundaries as annotated from GenBank (U14690.1) demonstrate that exon 4 is missing, due to a correction made after the initial description of the gene (Fackenthal and Olopade, 2007).

BRCA1 acts as a nuclear phospho-protein which shuttles between the nuclear and cytoplasmic compartments of the cell and has been identified as a critical protein within the DNA damage response and cell cycle control pathways (Friedenson, 2007). While BRCA1 itself is a signalling protein, it is often found co-localised with other tumour suppressor proteins, DNA damage sensors and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC), which enables multiple repair functions within the DNA damage repair pathway. This complex facilitates both the recognition of a break in the DNA (single stranded or double stranded) and the recruitment of further proteins and enzymes for the repair of these sites (Wang *et al.*, 2000). BRCA1 has also been shown to have various regulatory roles within G_2/M checkpoint control (Moynahan *et al.*, 1999).

These vital roles that BRCA1 plays in DNA damage repair indicates its fundamental significance in maintaining genomic stability. Additionally, it has been shown that BRCA1 acts as a regulatory protein in the apoptotic pathway, further illustrating the function of BRCA1 in response to cellular stress and damage (Thangaraju *et al.*, 2000, Venkitaraman, 2002). BRCA1 protein expression peaks during the S and G_1 phases of the cell cycle (during which DNA replication occurs), further emphasising the role of BRCA1 in genomic integrity (Vaughn *et al.*, 1996). Considering the multitude of roles BRCA1 plays in DNA damage repair and cell cycle control, it is not surprising that a loss of *BRCA1* expression would lead to the development of cancer.

1.4.2 BRCA2

BRCA2 (Breast Cancer 2, early onset; MIM 600185) was identified in 1995 through linkage studies carried out on families with multiple cases of breast cancer that were not associated with mutations within *BRCA1* (Wooster *et al.*, 1995). *BRCA2* consists of 27 exons (**Figure 1.3**) located on the long arm of chromosome 13 and encodes a protein of 3148 amino acids (Tavtigian *et al.*, 1996, Tonin *et al.*, 1996). Mutations within *BRCA2* are not only associated with breast and ovarian cancer, but are also associated with prostate, pancreatic and peritoneal cancers in addition to melanomas (The Breast Cancer Linkage Consortium, 1999).

BRCA2 is essential for multiple DNA repair pathways and in cell cycle control. *BRCA2* expression is tightly regulated during cell proliferation and evidence supports it being co-regulated with the expression of *BRCA1*. The expression of *BRCA2* is induced in rapidly proliferating cells and is regulated in a cell cycle dependant manner, peaking around the G_1/S phases when DNA replication occurs (Rajan *et al.*, 1996).

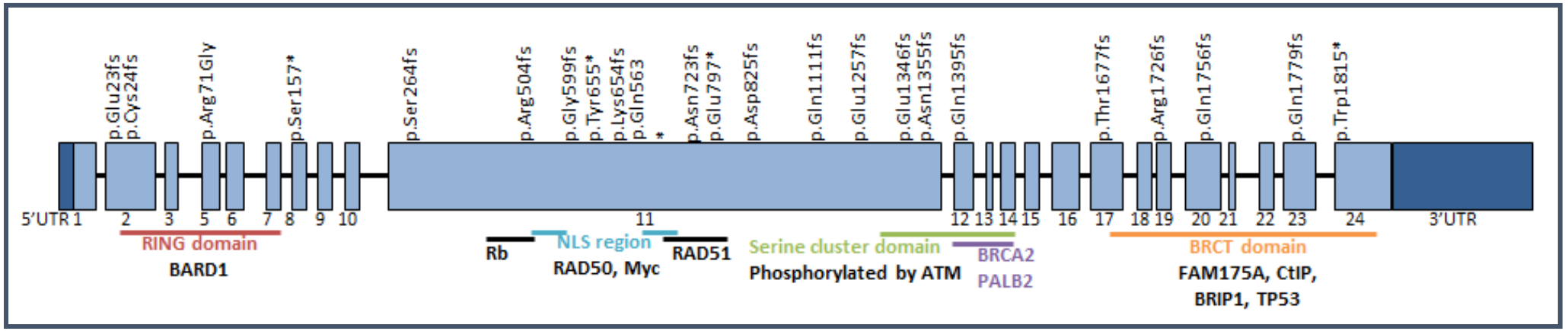


Figure 1.2: Gene structure of *BRCA1* including functional domain, interacting proteins and common mutations. Contains 23 exons, with exon 4 missing, and exon 11 being the largest. Exons are indicated by light blue boxes with untranslated regions (UTRs) shown in dark blue. Common pathogenic germline mutations associated with development of breast cancer are shown above exons. Exon numbers indicated under corresponding exons. Functional domains of BRCA1 are shown along with interacting proteins under the exons. RING; RING-type Zinc finger, NLS; nuclear localisation sequences, BRCT; BRCA1 C-terminus

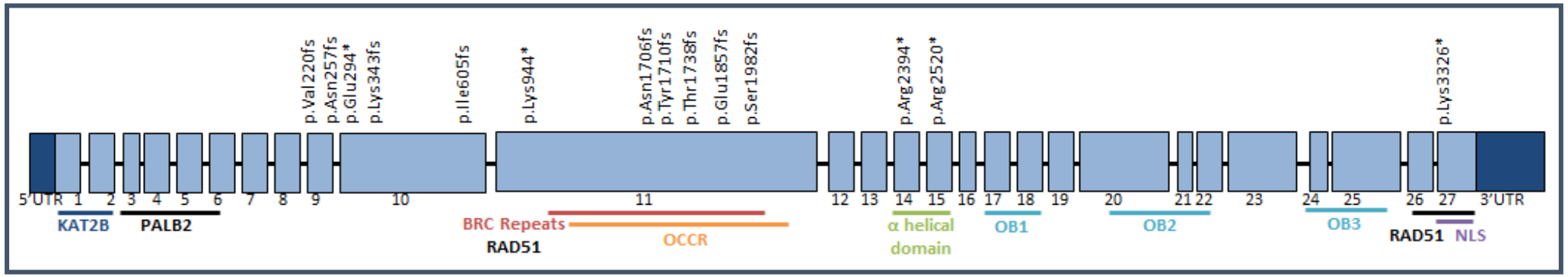


Figure 1.3: Gene structure of *BRCA2* including functional domains, interacting proteins and common mutations. Contains 27 exons, with exon 11 being the largest. Exons are indicated by light blue boxes with untranslated regions (UTRs) shown in dark blue. Common pathogenic germline mutations associated with development of breast cancer are shown above exons. Exon numbers indicated under corresponding exons. The functional domains of BRCA2 are shown along with the interacting proteins under the exons. BRC (BRCA C-terminal) repeats are eight highly conserved motifs contained within exon 11. OCCR; ovarian cancer cluster region, OB; oligonucleotide/oligosaccharide-binding fold, NLS; nuclear localisation sequences.

BRCA2 is a key protein involved in homologous recombination, a crucial pathway required for the repair of DNA double stranded breaks (DSBs) (Xia *et al.*, 2001). BRCA2 is required for the specific regulation of homologous recombination which functions to maintain genomic integrity and suppress tumorigenesis in proliferating cells. BRCA2 also regulates the activity of RAD51, an additional protein necessary for DNA repair via homologous recombination (Sharan *et al.*, 1997). It has been demonstrated that the transport of RAD51 into the nucleus is defective in cells with a known pathogenic *BRCA2* mutation (Sharan *et al.*, 1997). This suggests a direct role of BRCA2 in both the intracellular localisation and DNA binding of RAD51. Moreover, cells lacking BRCA2 result in spontaneous aberrations to chromosome structure that accumulates during the process of cell division. These abnormalities not only include broken chromosomes and chromatids, but also gross chromosomal rearrangements including translocations, deletions, and the fusion of multiple non-homologous chromosomes (Patel *et al.*, 1998). It is evident that BRCA2 plays a significant role in DNA damage repair and genomic integrity. Hence, a loss of function within these pivotal cellular pathways may be a key event in tumorigenesis.

1.5 Role of *BRCA1* and *BRCA2* mutations in cancer predisposition

The most commonly identified mutations associated with the pathogenesis of *BRCA1* and *BRCA2* mutations are loss of function mutations. These may result from point mutations, small insertions or deletions, large coding deletions or exon duplication (Carvalho *et al.*, 2007). These mutations have the ability to alter the reading frame of RNA sequences, resulting in the formation of a truncated or non-functional protein, and therefore affect protein expression or function (Gayther *et al.*, 1997). The presence of these inherited germline mutations within one allele is not sufficient to result in the onset of breast cancer, but rather acts as the 'first hit' of Knudson's two-hit hypothesis, leading to a cancer predisposition. A second mutation must occur within the remaining *BRCA* allele within the cell, leading to a complete loss of function of this important tumour suppressor.

There are approximately 2,000 identified mutations and sequence variants within *BRCA1* and *BRCA2* that are associated with a predisposition to breast cancer (Lin *et al.*, 2009). Variants associated with the development of breast cancer have been shown to be spread throughout the entirety of the *BRCA1* and *BRCA2* genes (**Figure 1.2** and **Figure 1.3**). These sequence variants are located within both the coding and non-coding regions of these genes and are classified as either pathogenic, non-pathogenic or sequence variants of unknown functional effect (Plon *et al.*, 2008). Pathogenic

mutations are genetic abberations resulting in the complete or partial loss of expression of the protein, or lead to the production of a non-functional protein (Whittemore *et al.*, 2004). These mutations usually result in the production of a premature stop codon, causing protein truncation. As illustrated **Figure 1.2** and **Figure 1.3**, frameshift and nonsense mutations are the most commonly identified mutations in *BRCA1/2* that are involved in breast cancer development. Additonal mutations that can lead to cancer development include translocations, inversions and large exon deletions.

Understanding the genetic basis of breast cancer development can be important for determining possible treatment options for affected individuals. Classification of biological markers in breast cancers is predominately based on the presence or absence of three receptors; oestrogen (ER+/ER), progesterone (PR+/PR-) and human epidermal growth factor receptor 2 (HER2+/HER2-) (Kittaneh *et al.*, 2013). Familial *BRCA1*-mutated breast cancers present as triple-negative tumours in approximately 80% of cases, whilst *BRCA2*-mutated cancers are most commonly ER+ and HER2- (Bayraktar and Gluck, 2012, Mavaddat *et al.*, 2012). Women with *BRCA* mutations have been found to be more likely to develop a secondary cancer, either within the same or opposite breast, and as a result, bilateral mastectomies are recommended to these women (Rebbeck *et al.*, 2004). Furthermore, women with *BRCA1*-associated cancers have shown an increased sensitivity to platinum agents such as cisplatin and other drugs that result in DNA DSBs (Silver *et al.*, 2010), whilst Poly ADP-ribose polymerase (PARP) inhibitors are highly effective in both *BRCA1* and *BRCA2* mutant cancers (Lord *et al.*, 2015). PARP inhibitors block the repair of DNA damage, resulting in instability, cell cycle arrest and leading to eventual apoptosis through synthetic lethality (Lord *et al.*, 2015). These inhibitors prevent repair of single stranded breaks, which are then converted to DSBs during replication. Due to the defective DNA DSB pathways in *BRCA*-mutant cells, these breaks are unable to be repaired and as such leads to apoptosis of the cancer cells (Livraghi and Garber, 2015). Therefore, it is clear that understanding of the genetic basis of the tumours is not only important for determining cancer risk but can also guide prophylactic and post-diagnosis treatment decisions.

1.5.1 Tissue-specific carcinogenesis observed in BRCA mutation carriers

Due to its imperative role in the fundamental processes of DNA damage repair and transcriptional regulation, both *BRCA1* and *BRCA2* are ubiquitously expressed. However, a tissue specific cancer predisposition is observed with *BRCA1/2* mutations. Unlike mutations in *TP53*, which can lead to widespread cancers throughout, mutations in *BRCA1/2* are primarily associated with breast and

ovarian cancers (and prostate cancer in males). However, the reason behind this tissue specificity is widely unknown (Welcsh and King, 2001, Venkitaraman, 2019).

Recent evidence suggests that R-loop accumulation is largely regulated by BRCA1 and BRCA2 and plays an important role in the tissue-specific nature of cancer development. Both BRCA1 and BRCA2 are required for the turnover of R-loops which are physiological intermediates of gene transcription and are a hybrid of RNA and single stranded DNA (ssDNA) (Aguilera and Garcia-Muse, 2012). Cells lacking BRCA1 or BRCA2 have shown an increase in R-loop accumulation (Bhatia *et al.*, 2014, Hill *et al.*, 2014), and it has been demonstrated that BRCA1 interacts with multiple proteins required for transcription, repair of transcriptional arrest and R-loop resolution (Hill *et al.*, 2014). An accumulation of R-loops has been observed at sites of unscheduled transcriptional termination, leading to ssDNA breaks and genomic instability (Hatchi *et al.*, 2015). These findings illustrate that unscheduled R-loop accumulation may result in significant endogenous DNA damage and subsequent chromosomal fragility following inactivation of BRCA1 or BRCA2 (Venkitaraman, 2019). Recent work carried out by Zhang *et al.* (2017) has shown that R-loops accumulate preferentially at promoter-proximal RNA Polymerase II pausing sites in luminal epithelial cells of BRCA1-mutant mammary tissues. This study identified that this Polymerase II pausing is an important contributor to R-loop accumulation, DNA damage and subsequent cancer development within breast luminal epithelial cells (Zhang *et al.*, 2017). This highlights the tissue-specific nature of cancer development observed in BRCA mutant individuals.

Additionally, the genotoxic nature of tissue-specific hormones such as oestrogen has been implicated in the tissue-specific nature of these cancers. The ability of oestrogen metabolites to result in DNA damage is well documented (Liehr, 1990, Montano *et al.*, 2012). However, recent evidence has further illustrated that oestrogen stimulation not only results in a rapid increase of hormonally-regulated gene transcription but also an increase in R-loop formation (Stork *et al.*, 2016). These oestrogen induced R-loops have been found to result in genomic instability, particularly at oestrogen responsive loci, playing an important role in DNA damage susceptibility at these sites and within these hormonally driven tissues. As oestrogen is particularly prevalent in the breast and ovary, this provides further understanding surrounding the tissue-specific nature of carcinogenesis associated with BRCA mutations.

While *BRCA1* and *BRCA2* are the most common genes known to be associated with the development of breast cancers, there are several other genes less frequently implicated in hereditary breast cancer cases. Furthermore, large genome wide association studies (GWAS) are still identifying novel loci associated with cancer predisposition (Michailidou *et al.*, 2015, Michailidou *et al.*, 2017).

1.6 Additional mechanisms of breast cancer susceptibility

Despite the knowledge of these two well-defined, high-penetrance breast cancer susceptibility genes, there are a significant proportion (approximately 80%) of familial breast cancers that are not found to be associated with mutations within *BRCA1* or *BRCA2.* These cancers may be due to undetected *BRCA1/2* mutations that are missed due to the current screening methods. Futhermore, an increasing number of *BRCA1/2* sequence variants of ambiguous functional signficance have been identified in a large number of families, constituting an increasing clinical challenge (Easton *et al.*, 2007, Larsen *et al.*, 2013). There are a considerable number of pathogenic mutations that are not localised to any one particular region with the *BRCA* genes, but rather are spread throughout the genes. As a result, analysis of the *BRCA* genes requires sequencing the entire coding region, which invariably results in the identification of hundreds of sequence variants in a single individual. A significant proportion of these variants have an unknown effect, and are therefore termed variants of uncertain significance (VUS) (Fernald *et al.*, 2011).

VUS can include missense mutations, in-frame insertion and deletions (indels) and splice site mutations which are often annotated as single nucleotide polymorphisms (SNPs) or multiple nucleotide polymorphisms (MNPs). These polymorphisms result in amino acid changes in the produced protein, the functional effect of which is often unclear as the severity of the effect can vary significantly (Wooster and Weber, 2003). The clinical significance of these VUS is unknown, resulting in an increased diagnostic challenge. With the implementation of massively parallel sequencing methods in diagnostic laboratories, the number of VUS identified within *BRCA1/2* and other susceptibility genes has increased significantly. As a result, new methods are required for the intepretation of VUS and increase detection rate of pathogenic germline mutations within *BRCA1/2*. Other techniques, such as RNA profiling and the use of gene signatures, have been shown to be beneficial for the identification of *BRCA*-associated breast tumours in individuals that were previously identified as BRCA mutation-negative (Larsen *et al.*, 2013).

1.6.1 Syndromic breast cancers

In addition to *BRCA1* and *BRCA2*, breast cancer can be associated with several inherited genetic syndromes (Antoniou and Easton, 2006). Germline mutations in *TP53* have been implicated in the development of Li-Fraumeni syndrome, an autosomal dominant disorder characterised by increased risk of tumour formation (Malkin *et al.*, 1990). Breast cancers are associated with this syndrome, and carriers of *TP53* mutations are at a high risk of developing early-onset breast cancer (Garber *et al.*, 1991). However, in individuals with a *TP53* mutation, it is far more likely for one of the first cancers identified to be a leukemia, melanoma, brain or soft tissue tumour, rather than a breast tumour (Olivier *et al.*, 2010). Additionally, these individuals often present with multiple cancers quite early in life (under 45 years of age) and demonstrate a strong familial history of cancer.

Breast cancer is also a feature of Cowden Syndrome, which occurs as a result of mutations within the *PTEN* gene (Starink *et al.*, 1986, Eng, 1998). Cowden syndrome is associated with a distinct phenotype of benign growth hamartomas on the surface of the skin and mouth and polyps within the gastrointestinal tract. This is observed in around 99% of individuals by their late 20s, and as a result, breast cancer is just one of the multiple cancers which present as secondary symptoms of the disease (Eng, 2003). Due to these observations, mutations within genes that are associated with a clear phenotype beyond breast cancer were excluded from this study.

1.6.2 Additional inherited breast cancer susceptibility genes

Given the large number of familial breast cancer cases that are not attributed to mutations within *BRCA1* and *BRCA2*, it has long been hypothesised that other breast cancer susceptibility genes exist. Family and linkage studies, candidate gene sequencing, genome wide association studies and case control association studies have been utilised to identify other breast susceptibility genes (Lalloo and Evans, 2012, Bogdanova *et al.*, 2013).

PALB2 (partner and localiser of BRCA2) has been shown to be another highly penetrant breast cancer susceptibility gene, with a 30-60% risk of developing breast cancer associated with a *PALB2* loss of function mutation (Antoniou *et al.*, 2014). Additionally, moderate-penetrance genes have been identified to play a role in breast cancer susceptibility. Mutations within *ATM, BRIP1* (BRCA1-interacting Protein 1) and *CHEK2* (checkpoint kinase 2) are associated with the formation of breast tumours (Olsen *et al.*, 2001, The CHEK2 Cancer Consortium, 2004). These genes are known to play a role in DNA repair, and pathogenic mutations within these genes confer an increased risk of breast

cancer by 2-fold (Seal *et al.*, 2006, Rahman *et al.*, 2007). The evidence for additional susceptibility genes has been determined through population-based screening of breast cancer affected families. These studies have revealed that only a proportion of breast cancer cases are attributed to mutations within *BRCA1* and *BRCA2* and other known breast cancer genes. It has been shown that mutations within known genes other than *BRCA1/2* only accounts for an additional 10% of hereditary breast cancers; therefore, the genetic predispositions underlying more than 70% of familial breast cancers remain unexplained (Turnbull and Rahman, 2008). This suggests that additional breast cancer susceptibility genes must exist.

Despite multiple genetic linkage studies, the identification of a *BRCA1* or *BRCA2*-like highly penetrant breast cancer susceptibility gene(s) has not been successful (Easton *et al.*, 1993, Kerangueven *et al.*, 1995, Seitz *et al.*, 1997, Smith *et al.*, 2006). These observations suggest that the majority of inherited breast cancer susceptibility may be polygenic in nature, implicating the involvement of a large number of low-penetrance genes (Pharoah *et al.*, 2002). The breast cancer risk associated with each low-penetrance locus is expected to be minor, however the cumulative effect of additional susceptibility alleles and environmental factors may explain the increased susceptibility risk and familial aggregation of cancer. Genetic polymorphisms identified within familial clusters in low-penetrance genes are often defined as “disease associated polymorphisms” or “functionally relevant polymorphisms”. This polygenic model of susceptibility is consistent with the observed familial aggregation patterns of inherited breast cancers and the overall risks observed are similar to those identified through epidemiological studies (Antoniou *et al.*, 2004). The clinical significance and association between moderate to low-penetrance alleles and cause of disease are difficult to establish due to inability to distinguish between genetic and environmental factors. Despite these difficulties, several additional breast cancer susceptibility genes have been identified to date, although they are not commonly offered as part of the diagnostic screening process (Refer to **Table 1.1**).

1.7 Current breast cancer screening

Hereditary cases of breast cancer are most commonly associated with a wide variety of pathogenic mutations within *BRCA1* and *BRCA2.* While there are several mutations which have been more frequently identified within specific genetic populations, including Ashkenazi Jews, African Americans and Hispanics (Mefford *et al.*, 1999, Weitzel *et al.*, 2007), mutations are typically located throughout the entirety of the gene. Due to this, it is not feasible to focus screening for causative

mutations to any particular mutation hotspot within the *BRCA1/2* genes, but rather full sequencing of the exons and flanking introns is required.

Breast cancer screening is currently offered by the South Australian Familial Cancer service to the 'at risk' population in South Australia, as determined by the Manchester scoring system (as discussed in **Section 1.3**). At the commencement of this project in 2014, screening methods consisted of Sanger sequencing of the coding and flanking intronic regions of *BRCA1* and *BRCA2*. In conjunction with Sanger sequencing, multiple ligation-dependant probe amplification (MLPA) is used to quantitatively analyse the genomic DNA for copy number variations, allowing for detection of duplications, inversions or deletions of whole exons or alleles (Schouten *et al.*, 2002, Sellner and Taylor, 2004).

Furthermore, screening is limited to the exons and surrounding introns of the *BRCA* genes, in which germline mutations only account for a small proportion of affected families (Apostolou and Fostira, 2013). Additionally, this protocol covers minimal regions of the non-coding regions (intronic and regulatory sequences) and hence, has a limited ability to detect variations within these non-coding regions. This is an issue as variants located within the regulatory and intronic regions have been shown to affect protein regulation, expression and/or function (Bogdanova *et al.*, 2013). Extending this screening methodology to include these regions could therefore be useful in investigating the role of regulatory and intronic variants in the predisposition to breast and ovarian cancer (Arnold *et al.*, 2002). However, this adds complexity to the identification of definitively pathogenic mutations, as the effect of sequence variants within these regions are still poorly understood. Additionally, this increases the cost and time associated with the sequencing of each individual and as a result, is not routinely carried out.

The BRCA screening protocol is both labour-and cost-intensive and is therefore limited to only the high-risk individuals with significant familial history (Trujillano *et al.*, 2015). This hinders the development of a widespread *BRCA* screening program for personalised risk assessment of hereditary breast and ovarian cancer to those who do not definitively meet the criteria for genetic testing. The development of such a program would play a crucial role in the early detection and prevention of hereditary breast cancer, as it has been estimated that approximately 50% of the clinical cases carrying a BRCA mutation remain undetected due to the current restrictive access to BRCA screening (Trujillano *et al.*, 2015)

As previously mentioned, hereditary breast cancer is not only attributed to mutations within *BRCA1* and *BRCA2*. There are a number of moderate and low-penetrance genes which play a role in familial breast cancer, yet the current sequencing regime is limited to *BRCA1* and *BRCA2.* Mutations within any one of these previously recognised breast cancer susceptibility genes are often rare and testing all of these genes by Sanger sequencing is both inefficient and expensive (Tung *et al.*, 2015). Therefore, there is clear utility in changing to a more cost and time effective method, such as massively parallel sequencing (MPS), in which simultaneous sequencing of multiple cancer susceptibility genes can be achieved through multiplexed gene panels.

Since the commencement of this study, the South Australian Familial Cancer service, like most diagnostic labs, have moved toward an MPS based approach (Refer to **Section 1.8**) as it significantly reduces the time associated with screening the referred individuals. These include a panel of genes implicated in hereditary breast cancer, often ranging from 5 – 25 genes. Genes most commonly screened include *BRCA1, BRCA2, ATM, CHEK2, BARD1, PALB2* and *TP53* (Easton *et al.*, 2015, Winship and Southey, 2016). The panel offered by the Genetic Pathology service in Adelaide is a 5 gene panel including *BRCA1, BRCA2, TP53, PALB2* and *PTEN* and is now routinely used for analysis of all individuals referred for genetic screening over the Sanger-based approach.

1.8 Massively Parallel Sequencing

As emphasised above, the limitations associated with the BRCA screening protocol at the commencement of this study illustrate the need for a more high throughput and cost-effective screening approach which could reduce the turn-around time, labour intensiveness and costs associated with BRCA genetic screening (Trujillano *et al.*, 2015). While Sanger sequencing is the 'gold standard' of sequencing technologies, there have been remarkable advances in DNA sequencing platforms with the emergence and evolution of MPS.

MPS (also known as Next Generation Sequencing; NGS) is a high throughput approach to DNA sequencing. MPS technologies utilise miniaturised platforms, which allows sequencing of 1 million to 43 billion short reads (usually 50-400bp) in a single run (Tucker *et al.*, 2009). These platforms often differ in their sequencing chemistries but share the technical paradigm of massively parallel sequencing via spatially separated, clonally amplified DNA templates. The demand for high throughput, low cost sequencing has driven the development of multiple platforms that produce thousands of sequences concurrently, with some platforms having the potential to run as many as

500,000 sequencing by synthesis reactions in parallel (ten Bosch and Grody, 2008). MPS has a multitude of applications including genome sequencing, transcriptome profiling, DNA–protein interactions and epigenome characterisation (de Magalhaes *et al.*, 2010).

1.8.1 MPS comparison to Sanger sequencing

MPS has revolutionised genomic and genetic research, with validated advantages over Sanger sequencing including the ability to generate massive amounts of data as a result of the huge parallel sequencing capacity (Metzker, 2010). These massively parallel runs allow thousands of reads to be generated concurrently, whilst Sanger sequencing is limited by a 96 well capillary array, allowing for approximately 70bp/capillary/hour (Hert *et al.*, 2008).

However, the increased throughput of MPS comes at the expense of read length. The majority of the available sequencing platforms offer shorter average read lengths (30-400bp) in comparison to the conventional Sanger sequencing of approximately 700bp (Hert *et al.*, 2008, Rizzo and Buck, 2012). Shorter read length restricts the types of experiments that MPS can be used for. Additionally, shorter read lengths may not map back to the reference genome uniquely, resulting in repetitive regions of the genome unable to be mapped (Nagarajan and Pop, 2010). Sequence alignment of MPS generated data is often difficult for regions with high levels of diversity (in comparison to a reference genome) due to the presence of structural variants such as insertions, deletions and translocations

Sanger sequencing is the most readily available and oldest sequencing technology, with well-defined chemistry that makes it the most accurate method for sequencing to date (Rizzo and Buck, 2012). As previously mentioned, Sanger sequencing is capable of reading DNA fragments much larger than the input limitations of MPS templates, and is still considered to be the gold standard in the clinical setting (Kingsmore and Saunders, 2011). However, Sanger sequencing has restricted applications due to technical limitations of the workflow; with the main factor being throughput, with the number of sequencing reactions that can be run in parallel failing in comparison to MPS platforms. The methodology associated with Sanger sequencing is the primary bottleneck, resulting in an increase in turnaround time. Due to this, many diagnostic laboratories have shifted from Sanger sequencing to high throughput MPS platforms (Costa *et al.*, 2013).

1.8.2 Comparison of sequencing technologies

Sequencing technologies are evolving rapidly and during the early 2010s several new sequencing platforms were released. While there are a wide range of sequencing platforms available, at the commencement of this study the two main technologies dominating the market were the Ion Torrent Personal Genome Machine (PGM) and the Illumina MiSeq. Due to the resources available within the Flinders Genomics Facility at the time of commencing this project, the Ion Torrent PGM was selected for the analysis of patients within this thesis (discussed further in **Section 1.9).**

1.8.2.1 Ion Torrent Sequencing

The Ion Torrent sequencing technology, utilises semi-conductor technology and a sequencing by synthesis approach, detecting protons that are released as nucleotides are incorporated during synthesis (Rothberg *et al.*, 2011). DNA fragments with specific adapter sequences are linked to and then clonally amplified by emulsion PCR on the surface of 3-micron diameter beads, known as Ion Sphere Particles (**Figure 1.4**)(Quail *et al.*, 2012). The templated beads are loaded into proton sensing wells of a semiconductor sequencing chip. As the sequencing reaction proceeds, each of the 4 nucleotides are introduced sequentially. As bases are incorporated, protons are released, and a signal is detected, which is proportional to the number of bases incorporated (Rothberg *et al.*, 2011). Ion Torrent generates an abundance of short reads (200bp fragments) which can be mapped back to a reference genome for assembly.

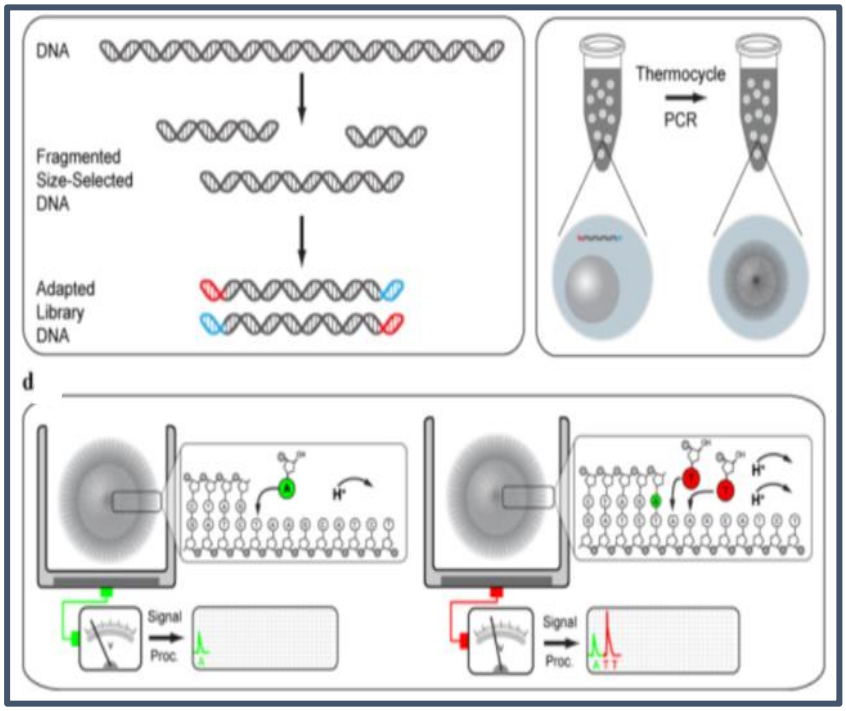


Figure 1.4: Ion Torrent Sequencing Workflow. DNA template is fragmented and ligated to adapters. Adapted DNA undergoes clonal amplification via emulsion PCR on an Ion Sphere Particle (ISP). ISPs are placed into a single well on a slide and the slide is flooded with single species of dNTPs, buffers and polymerase. When a nucleotide is incorporated the release of a H^+ ion results in a decrease in pH. The change in pH is used to determine the sequence (Adapted from BGI Platform, 2016)

1.8.3 Errors associated with the Ion Torrent sequencing chemistry

A well-known issue with Ion Torrent sequencing chemistry is the high error rate within homopolymer regions, i.e. three or more consecutive identical DNA bases (Quail *et al.*, 2012). These errors are a consequence of inaccurate flow-values, resulting in over or under calling of the length of homopolymeric regions (Bragg *et al.*, 2013). Additionally, Ion Torrent data is known to reduce in quality towards the end of sequencing reads, and within GC rich regions (Loman *et al.*, 2012). These errors were addressed during data analysis (see **Section 3.4**). Sequencing errors are a major challenge associated with the analysis of MPS generated data, especially for SNPs and indels. It is imperative to reduce the number of false positive calls for each technology

1.8.4 MPS applications to *BRCA1/2* screening

Many distinct loss of function mutations have been identified for both *BRCA1* and *BRCA2*. As these loss of function mutations are scattered throughout the *BRCA* genes, it is necessary to screen all coding regions of *BRCA1* and *BRCA2* for a genetic diagnosis of hereditary breast cancer. In addition

to Sanger sequencing, this can be carried out though targeted MPS panels which utilise multiplexed PCR reactions to sequence the entire coding regions of *BRCA1* and *BRCA2* in parallel which are available from several vendors. The efficacy and accuracy of the various MPS workflows has been compared by several studies in order to validate the use of MPS in the diagnostic setting (Chan *et al.*, 2012, Tarabeux *et al.*, 2014). These studies have sequenced entire *BRCA1* and *BRCA2* genes using MPS technologies and verified the sensitivity and specificity of MPS. In addition to accuracy and easy incorporation of the MPS workflow into the diagnostic setting, these studies emphasised that MPS improved turnaround time and increased sensitivity so that previously undetected variants were identified. These studies highlight the potential for mutation screening of clinically important gene targets in the diagnostic setting.

Other studies have also evaluated the efficacy of the Ion AmpliSeq *BRCA1* and *BRCA2* panel (Life Technologies; Carlsbad, CA) in conjunction with the Ion Torrent PGM (Life Technologies). Trujillano *et al.* (2015) validated this methodology as an accurate, comprehensive and cost-effective alternative to the conventional *BRCA* screening protocol. Utilising a validation cohort of individuals that had previously been subjected to Sanger sequencing, a comparison of the mutations and SNPs identified by both methodologies was carried out. Subsequently, patients with unknown *BRCA1* and *BRCA2* mutational status were analysed, and identified mutations in 51% of individuals, all of which were subjected to Sanger sequencing for confirmation (Trujillano *et al.*, 2015). This study highlights the sensitivity and specificity of the MPS methodology and its effectiveness in the diagnostic setting, illustrating that is it more cost and time effective, but it also offers higher throughput and scalability than the Sanger alternative.

1.8.5 MPS applications to breast cancer susceptibility genes

Advances in sequencing technologies have made multi-gene analysis a practical option when seeking to identify variants associated with a disease phenotype. This methodology is more efficient as it allows the simultaneous analysis of multiple genes in one sequencing reaction. Through this, it is possible to not only screen the high-penetrance susceptibility genes, but also the mid- to low-penetrance genes for the identification of cancer associated variants. This process relies on multiplexed sample preparation and in-depth bioinformatics analysis; however, data generation is still faster than multiple Sanger sequencing reactions (Judkins *et al.*, 2015). Custom gene panels or pre-designed gene panels which target known susceptibility genes are commercially available for a multitude of cancers and diseases in the general population.

Recent studies have utilised commercial panels for the analysis of germline mutations in cancer susceptibility genes other than *BRCA1* and *BRCA2* in a large cohort of *BRCA* mutation-negative individuals with a familial history of breast cancer. The majority of genes included on these panels are selected based on their established role in the development of inherited cancers. A large cohort study carried out by Tung *et al.* (2015) utilised a multiplexed gene panel analysis of 25 high- to low-penetrance breast cancer susceptibility genes. This study demonstrated that screening additional breast cancer susceptibility genes can identify mutations in approximately 5% of patients which had tested negative for mutations within *BRCA1/2*, with mutations most commonly identified in *CHEK2*, *ATM* and *PALB2* (Tung *et al.*, 2015). This supports the utility in sequencing not only *BRCA1/2* but also other susceptibility genes that have previously been shown to confer an increased risk of breast cancer. In addition, this methodological approach may also lend itself to the identification of novel breast cancer predisposition genes.

1.9 Experimental outline

As previously discussed, both *BRCA1* and *BRCA2* play pivotal roles in maintaining genome integrity by their involvement in DNA damage repair, homologous recombination and G_2/M cell cycle control (pathways illustrated in **Appendix** Error! Reference source not found.). Therefore, it is biologically feasible that mutations within genes in these pathways or genes that are acted on directly by *BRCA1/2* may also be implicated in the development of inherited breast cancer. As mentioned, there are several other known susceptibility genes which are also implicated in hereditary breast cancer cases, in addition to new causative genes that are still being identified. Recently, whole exome sequencing in inherited breast cancer individuals identified a novel cancer susceptibility gene *RECQL* (ATP dependent DNA helicase Q1), which plays a role in resolving stalled DNA replication forks to prevent DNA DSBs (Cybulski *et al.*, 2015). The function of this protein is related to that of other known susceptibility genes, illustrating the validity of the hypothesis that additional genes with similar roles to *BRCA1* and *BRCA2* may play a role in cancer predisposition. It is this principle which formed the basis for this study, consisting of designing a custom gene panel to be used for the analysis of not only known breast cancer susceptibility genes but also putative breast cancer genes.

In depth literature searches and pathway analysis was carried out in our department resulting in the curation of a custom gene panel comprised of genes that were predicted to contribute to the development of breast or ovarian cancer (Braun *et al.*, 2013). This gene panel consisted of previously identified breast and ovarian cancer susceptibility genes in addition to an array of genes that may

potentially be implicated in breast and ovarian cancer development (**Table 1.1** and **Table 1.2**). Custom gene panels have previously been shown to have benefit in the analysis of targeted pathways and have identified genes with an integral role in specific pathways in cancer development. For example, targeted MPS technologies were utilised for the identification of a novel breast cancer susceptibility gene *XRCC2* (Park *et al.*, 2012). This evidence illustrates that targeted gene panels are an appropriate approach to utilise for the analysis of genes within specific pathways of interest, allowing for further elucidation in their role in cancer predisposition.

Table 1.1: Genes included in the custom AmpliSeq diagnostic panel comprising of known breast cancer susceptibility genes. Cytogenetic location and mendelian inheritance in man (OMIM) ID indicated.

Acronym	Gene Name	Location/ OMIM ID	Function in relation to *BRCA1*/*BRCA2*
ATM	Ataxia telangiectasia mutated	11q22.3 607585	Cell cycle checkpoint kinase that regulates of a variety of downstream TSGs including *TP53* and *BRCA1* (Banin *et al.*, 1998, Cortez *et al.*, 1999). Master control protein in cell cycle checkpoint signalling (Savitsky *et al.*, 1995). *ATM* mutations have been identified in *BRCA* mutation-negative familial breast cancer (Thorstenson *et al.*, 2003, Thompson *et al.*, 2005, Renwick *et al.*, 2006).
BARD1	BRCA1 associated RING domain 1	2q35 601593	Shares homology with the 2 most conserved regions of *BRCA1* – The RING motif and BRCT domain (Refer to **Figure 1.2**) BARD1/BRCA1 interaction is disrupted by tumorigenic amino acid substitutions in *BRCA1* (Wu *et al.*, 1996). Pathogenic mutations have been associated with breast cancer predisposition (Karppinen *et al.*, 2004).
BRCA1	Breast Cancer 1, early onset	17q21.31 113705	Previously established breast cancer susceptibility gene (Miki *et al.*, 1994).
BRCA2	Breast Cancer 2, early onset	13q13.1 600185	Previously established breast cancer susceptibility gene (Wooster *et al.*, 1995).
BRIP1	BRCA1 interacting protein 1	17q23.2 605882	Interacts with the BRCT repeats of *BRCA1* and plays a role in dsDNA break repair. Identified as a low-penetrance breast cancer susceptibility gene (Seal *et al.*, 2006).
CDH1	Cadherin 1	17q23.2 192090	Loss of function is thought to contribute to cancer progression by increasing proliferation, invasion and/or metastases (Hiraguri *et al.*, 1998). Mutations in *CDH1* are associated with multiple cancers, including breast cancer (Guilford *et al.*, 1998, Masciari *et al.*, 2007, Chang *et al.*, 2014).
CHEK2	Checkpoint Kinase 2	22q12.1 604373	Involved in regulation of cell cycle checkpoints and tumour suppression (Matsuoka *et al.*, 1998). Interacts with BRCA1, enabling survival post DNA damage (Lee *et al.*, 2000). Identified as low-penetrance breast cancer gene in *BRCA* mutation-negative families (Meijers-Heijboer *et al.*, 2002).
FAM175A	Family with sequence similarity 175, member A	4q21.23 611143	Binds directly to the BRCT domain of BRCA1, targeting it to the sites of DNA damage. Required for G_2/M checkpoint control and DNA damage repair (Wang *et al.*, 2007). Demonstrated to be a low-penetrance breast cancer susceptibility gene (Solyom *et al.*, 2012).
HMMR	Hyaluronan-mediated motility receptor	5q34 600936	Expressed in breast tissue and forms a complex with other proteins including BRCA1 and BRCA2 and thus is associated with a higher risk of breast cancer (Pujana *et al.*, 2007).

Acronym	Gene Name	Location/ OMIM ID	Function in relation to *BRCA1*/*BRCA2*
MRE11A	Meiotic recombination 11, Homolog A (*S.Cerevisiae*)	11q21 600814	Involved in Homologous Recombination (HR) and dsDNA break repair (Paull and Gellert, 1998). Forms a complex with RAD50 and NBN which mediates the response of *BRCA1* to cellular damage and dsDNA break repair (Carney *et al.*, 1998, Zhong *et al.*, 1999). Identified as a moderately penetrant breast cancer susceptibility gene (Bartkova *et al.*, 2008, Yuan *et al.*, 2012).
NBN	Nibrin	8q21.3 602667	Forms part of the double stranded break repair complex (Carney *et al.*, 1998). Polymorphisms have been associated with increased risk of breast cancer (Gorski *et al.*, 2003, Zhang *et al.*, 2013a).
NQO2	NAD(P)H Dehydrogenase, Quinone 2	6p25.2 160998	Mutations ion *NQO2* lead to TP53 instability and are associated with the development of breast cancers (Yu *et al.*, 2009).
PALB2	Partner & localiser of BRCA2	16p12.2 610355	Binds to and co-localises with BRCA2, resulting in the stable intranuclear localisation and accumulation of BRCA2 at sites of DNA DSBs (Xia *et al.*, 2006). Mutations often result in protein truncation, resulting in decreased BRCA2-binding capacity, and deficiencies in homologous recombination (Erkko *et al.*, 2007). Mutations have been linked to hereditary breast and ovarian cancer(Rahman *et al.*, 2007, Teo *et al.*, 2013)
RAD50	RAD50 Homolog (*S.Cerevisiae*)	5q31.1 604040	Part of the dsDNA break repair complex (Carney *et al.*, 1998). Component of the BRCA1-associated genome surveillance complex (Wang *et al.*, 2000).
RAD51	RAD51 Homolog (*S.Cerevisiae*)	15q15.1 179617	Interacts with BRCA1 and BRCA2 (Jensen *et al.*, 2010). Intracellular localisation and DNA-binding ability is regulated by BRCA2 (Yang *et al.*, 2005), loss of which is thought to be a key event leading to genomic instability and tumourigenesis (Akisik *et al.*, 2011).
RAD51C	RAD51 Homolog Paralog C (*S.Cerevisiae*)	17q22 602774	Involved in homologous recombination and DNA repair (Dosanjh *et al.*, 1998). Germline mutations confer high ovarian cancer risk (Coulet *et al.*, 2013).
RAD51D	RAD51 Homolog Paralog D (*S.Cerevisiae*)	17q12 602954	Involved in HR and DNA repair (Hinz *et al.*, 2006). Loss of function mutations confer high risk of ovarian cancer (Thompson *et al.*, 2013b).
TP53	Tumour Protein p53	17p13.1 191170	Pivotal tumour suppressor protein (Vogelstein and Kinzler, 1994, Yin *et al.*, 2002). Responds to cellular stress in order to regulate gene expression, inducing cell cycle arrest, apoptosis, senescence, and DNA repair (Toledo and Wahl, 2006, Bourdon, 2007). Mutations are associated with a variety of cancers and disorders including breast cancer (Malkin *et al.*, 1990, Hollstein *et al.*, 1991, Petitjean *et al.*, 2007).
XRCC2	X-ray repair complementing defective repair in Chinese hamster cells 2	7q36.1 600375	Involved in homologous recombination to maintain chromosome stability and repair DNA damage (Tambini *et al.*, 1997, Johnson *et al.*, 1999). Rare variants have been associated with increased breast cancer susceptibility (Hilbers *et al.*, 2012, Park *et al.*, 2012).

Table 1.2: Genes included in the custom AmpliSeq discovery panel, comprising of potential breast cancer susceptibility genes. Cytogenetic location and mendelian inheritance in man (OMIM) ID indicated.

Acronym	Gene Name	Location/ OMIM ID	Potential Role in Breast Cancer Susceptibility
ATF1	Activating Transcription Factor 1	12q13.12 123803	BRCA1 directly acts on ATF1; and is required for activation of *ATF1* and its target genes (Houvras *et al.*, 2000) Involved in cell growth, survival and DNA damage response
BRCC3	BRCA1/BRCA2 containing complex, subunit 3	Xq28 300617	Component of the BRCA1-and BRCA2-containing complex. BRCC3 binds directly with BRCA1 and is responsible for BRCA1 accumulation at sites of DNA damage (Dong *et al.*, 2003).
CDKN1A	Cyclin dependent kinase inhibitor 1A	6p21.2 116899	Expression and function of CDKN1A is regulated by TP53 (el-Deiry *et al.*, 1993). Overexpression of CDKN1A acts as a mediator of cell cycle arrest in response to DNA damage (Bendjennat *et al.*, 2003). Protein levels have been shown to be affected in multiple types of cancer (Huang *et al.*, 2014, Zhang *et al.*, 2014).
CDKN2A	Cyclin dependent kinase inhibitor 2A	9p21.3 600160	Regulates both the TP53 and RBL pathways involved in cell cycle regulation (Robertson and Jones, 1999). *CDKN2A* is often mutated/deleted in many tumour types (Kamb *et al.*, 1994). Identified as a low-penetrance breast cancer susceptibility gene (Borg *et al.*, 2000, Debniak *et al.*, 2005a)
CHEK1	Cell cycle checkpoint Kinase 1	11q24.2 603078	Binds directly to BRCA1. Required for cell proliferation and survival (Tang *et al.*, 2006) and cell cycle mediated repair in response dsDNA breaks (Zhao *et al.*, 2002).
CKS1B	CDC28 Protein Kinase 1B	1q21.3 116900	Promotes mitosis through modulation of protein kinases (Morris *et al.*, 2003). Overexpression of CKS1B has been observed in multiple cancers, including breast cancer (Martin-Ezquerra *et al.*, 2011, Liberal *et al.*, 2012, Wang *et al.*, 2013).
E2F1	E2F Transcription Factor 1	20q11.22 189971	E2F (E2F1, E2F2, E2F3) factors act as transcriptional activators for progression through the cell cycle (Wu *et al.*, 2001). Activated in response to DNA damage and drives the expression of pro-apoptotic genes (Morris *et al.*, 2008).
E2F2	E2F Transcription Factor 2	1p36.12 600426	Refer to entry for E2F1
E2F3	E2F Transcription Factor 3	6p22.3 600427	Altered copy number and activity of *E2F3* have been observed in human cancers (Bambury *et al.*, 2015).
E2F4	E2F Transcription Factor 4	16q22.1 600659	Contains a tumour suppressor transactivation domain and plays a role in the suppression of proliferation associated genes (Ginsberg *et al.*, 1994). Component of the E2F complex to which BRCA1 directly binds.
E2F5	E2F Transcription Factor 5	8q21.2 600967	Refer to entry for E2F4
E2F6	E2F Transcription Factor 6	2p25.1 602944	Interacts with chromatin modifying factors and inhibits transcription (Ogawa *et al.*, 2002).

Acronym	Gene Name	Location/ OMIM ID	Potential Role in Breast Cancer Susceptibility
EP300	E1A-Binding Protein, 300-KD	22q13.2 612986	Regulates transcription via chromatin remodelling and plays a role in the stabilisation of TP53 (Gayther *et al.*, 2000, Grossman *et al.*, 2003). Targeted by viral onco-proteins (Arany *et al.*, 1995) Implicated in a variety of cancer types, including breast cancer (Muraoka *et al.*, 1996, Gayther *et al.*, 2000, Le Gallo *et al.*, 2012)
GADD45A	Growth arrest and DNA damage-inducible gene, alpha	1p31.3 126335	Stimulates DNA repair and inhibits damaged cells from entering S phase (Smith *et al.*, 1994). BRCA1 and GADD45A have been shown to play a synergistic role in regulating centrosome duplication and maintaining genomic integrity (Wang *et al.*, 2004).
HLTF	Helicase-like transcription factor	3q24 603257	BRCA1 binds directly to the SWI/SNF complex. Encodes chromatin remodelling factors which have been identified to be disrupted in some cancers (Moinova *et al.*, 2002).
KAT2B	K(Lysine) Acetyltransferase 2B	3p24.3 602303	Associates with EP300 and CBP to play a role in transcriptional regulation through acetyltransferase activity with core histones and nucleosome particles (Yang *et al.*, 1996). Promotes apoptosis (Zheng *et al.*, 2013). Reduced expression is associated with several cancers (Ying *et al.*, 2010, Akil *et al.*, 2012).
PKMYT1	Protein Kinase Membrane Associated Tyrosine/Threonine 1	16p13.3 602474	Negatively regulates the G_2/M cell cycle transition through inhibitory phosphorylation in conjunction with WEE1 (Wells *et al.*, 1999).
PRKDC	Protein Kinase, DNA Activated Catalytic Subunit	8q11.21 600899	Plays a role in cell cycle control, dsDNA break repair and modulation of transcription (Anderson and Lees-Miller, 1992, Hartley *et al.*, 1995). Specific polymorphisms have been associated with an increased risk in cancer susceptibility (Zhou *et al.*, 2012, Zhang *et al.*, 2013b, Hsia *et al.*, 2014, Xiao *et al.*, 2014).
RBL1	Retinoblastoma-like 1	20q11.23 116957	Similar in sequence and possibly function to the RB1 gene, which plays a role in cell cycle regulation (Ewen *et al.*, 1991). BRCA1 acts directly on both RBL1 and RBL2. Forms a complex with HLTF, DP1, E2F4 and E2F5 to mediate transcriptional activation (Chen *et al.*, 2002).
RBL2	Retinoblastoma-like 2	16q12.2 180203	Refer to entry for RBL2
RFC2	Replication Factor C, Subunit 2	7q11.23 600404	Multimetic subunit consisting of 5 subunits (RFC1-5) (Okumura *et al.*, 1995). Component of the BRCA1-associated genome surveillance complex (Wang *et al.*, 2000).
RFC3	Replication Factor C, Subunit 3	13q13.2 600405	Involved in DNA mismatch repair mechanisms (Woerner *et al.*, 2003).
RFC4	Replication Factor C, Subunit 4	3q27.3 102577	Refer to entry for RFC2

Acronym	Gene Name	Location/ OMIM ID	Potential Role in Breast Cancer Susceptibility
RFC5	Replication Factor C, Subunit 5	12q24.23 600407	Refer to entry for RFC2
RPA1	Replication Protein A1	17p13.3 179835	Involved in recruiting DNA repair proteins to sites of DNA damage (Oakley and Patrick, 2010). Missense mutations have been shown to result in defects in dsDNA break repair, leading to tumour development (Wang *et al.*, 2005)
RPRM	Reprimo	2q23.3 612717	Plays a role in p53-induced G_2 cell cycle arrest (Sato *et al.*, 2006). Often aberrantly methylated in several tumour cell lines and multiple cancers (Beasley *et al.*, 2008, Bernal *et al.*, 2008, Ooki *et al.*, 2013).
RPS6KA1	Ribosomal Protein S6 Kinase 1	1p36.11 601684	Involved in control of cell growth and differentiation (Bonni *et al.*, 1999). Polymorphisms are associated with increased risk in some cancers (Lara *et al.*, 2011, Slattery *et al.*, 2011).
***SFN* (14-3-3-σ)**	Stratifin	1p36.11 601209	Expression induced in response to DNA damage, with a loss of expression resulting in impaired G_2/M checkpoint control (Chan *et al.*, 1999). Hypermethylation, resulting in gene silencing, has been shown to result in decreased expression of SFN in breast cancer cells in comparison to normal breast epithelium (Ferguson *et al.*, 2000).
SLC19A1	Solute carrier family 19 (folate transporter) member 1	21q22.3 600424	Plays a role in homologous recombination
SMARCD2	SWI/SNF-Related matrix-associated, actin-dependent regulator of chromatin, Subfamily D, Member 2.	17q23.3 601736	BRCA1 binds directly to the SWI/SNF complex (Bochar *et al.*, 2000). Involved in chromatin remodelling, which is often disrupted in the development of cancers.
UIMC1	Ubiquitin Interaction Motif-Containing Protein 1	5q35.2 609433	Forms a complex with Abraxas to recruit BRCA1 to DNA damage sites (Wang *et al.*, 2007). Directly binds to the BRCT domain of BRCA1 (Sobhian *et al.*, 2007). Missense mutations have been associated with an increased risk of breast cancer susceptibility (Akbari *et al.*, 2009).
WEE1	WEE1 Tyrosine Kinase	11p154. 193525	Coordinates the transition between DNA replication and mitosis, blocking cell division when over expressed (Heald *et al.*, 1993, McGowan and Russell, 1993). High expression levels have been associated with multiple cancers, including breast and ovarian cancers (Porter *et al.*, 2012, Magnussen *et al.*, 2013, Ghiasi *et al.*, 2014).

1.10 Thesis Hypotheses and Objectives

Overarching Hypothesis:

Mutations in proteins involved in *BRCA1/2*-related DNA damage repair and checkpoint control pathways play a role in predisposing individuals to inherited breast cancer.

Overall Aim:

To use targeted gene capture and MPS to sequence known and putative breast cancer susceptibility genes in a cohort of *BRCA1/2* mutation-negative individuals.

The broad aims of this study were:

1. The development of a bioinformatics pipeline for the analysis of *BRCA1/2* mutation-negative individuals with a custom AmpliSeq gene panel (**Chapter 3**)
2. To determine if a pooling approach could be utilised for the identification of rare variants in the BRCA1/2 mutation-negative individuals **(Chapter 4)**
3. To perform an in-depth analysis of sequencing data from mutation-negative individuals for identification of potential susceptibility mutations involved in development of breast cancer **(Chapter 5)**
4. The characterisation of the predicted pathogenic effect of selected variants identified within the patient cohort using CRISPR/Cas9 genome editing (**Chapter 6**)

The experimental outline for the completion of this PhD thesis is illustrated in **Figure 1.5.**

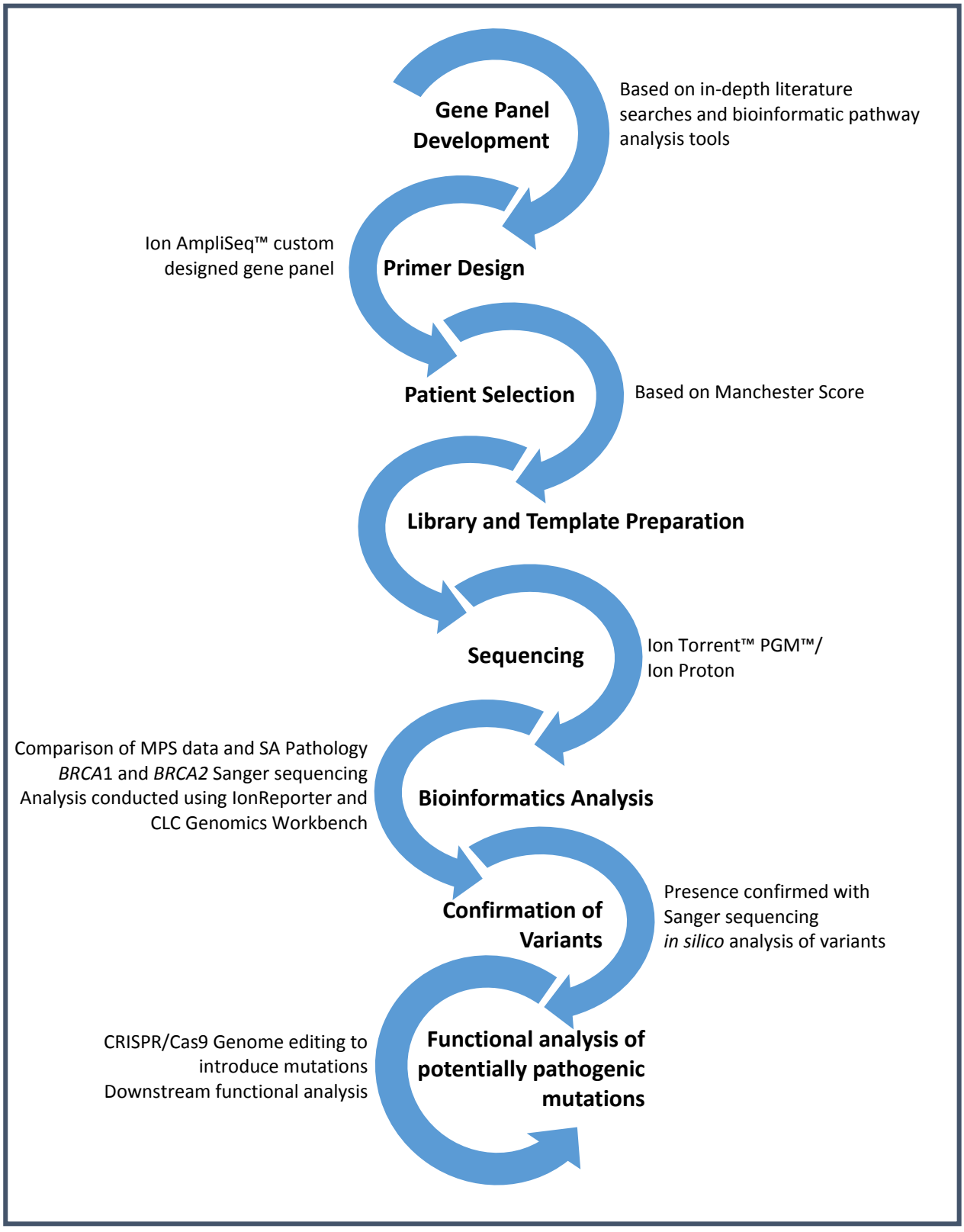


Figure 1.5: Experimental outline for research project carried out within this thesis.

Chapter 2:
Methods and Materials

The following chapter describes common techniques and methods used throughout this this thesis. Specific methods pertaining to only one chapter are presented in their relevant chapters. General buffers and solutions used in this thesis are listed in **Section 2.5**

2.1 Patient Selection

All individuals included in this study had been referred to the Familial Cancer Screening Unit for *BRCA1/2* testing. Patient consent for broad use of genomic material was previously obtained for all patients at the time of venepuncture. Ethics approval was obtained from Southern Adelaide Clinical Human Research Ethics Committee (Application number: 132.13). The patient Manchester scores (refer to **Section 2.1.1** for explanation of Manchester scoring system) of all individuals was used for the selection of patients for the pilot study (13 patients, Refer to **Chapter 3**). The individuals included in the extended study (n=119) were selected based on the date they were referred for genetic testing, with an aim to carry out a longitudinal study over a 12-month period.

2.1.1 Manchester Scores

Patients with a wide range of Manchester scores (5 to 61) were selected for sequence analysis. The Manchester scores of each individual were previously determined by health care professionals at the South Australian Familial Cancer service as outlined in **Table 2.1**. All samples had been previously screened for mutations in *BRCA1* and *BRCA2* and had been found to be mutation-negative. Eleven samples with identified pathogenic *BRCA1/2* mutations were also included as controls.

Table 2.1: Manchester scoring system. Manchester scores of individuals are tallied based on the frequency and types of cancers within familial history, in addition to age of onset. Modified from Evans *et al.* (2005)

Type of cancer	Age at diagnosis	Score
Female breast cancer	<30	11
	30-39	8
	40-49	6
	50-59	4
	>59	2
Male breast cancer	<59	13
	>60	10
Ovarian Cancer	<59	13
	>60	10
Pancreatic cancer	-	1
Prostate cancer	<59	2
	>60	1
		Total = Manchester Score

2.1.2 Genomic DNA isolation

Patient genomic DNA (gDNA) samples were previously extracted from peripheral blood using the Illustra Blood Genomic Prep Mini Spin Kit (GE Healthcare, Bio-Sciences AB, Uppsala, Sweden) as per the manufacturer's instructions. These extractions were carried out in the Department of Molecular Pathology, SA Pathology, which exclusively performs all the inherited breast cancer diagnostics testing for South Australia.

DNA concentrations were measured through spectrophotometry using the dsDNA High Sensitivity Assay and the Qubit 2.0 Flurometer (ThermoFisher Scientific, Massachusetts, USA) as per the manufacturer's protocols (MAN0002326, Life Technologies, California, USA).

2.2 Massively parallel sequencing methods.

All sequences were mapped to the human genome version hg19 (February 2009 build, GrCh37). Ion Reporter (v4.2) and CLC Genomics Workbench (v.6.02) were used for bioinformatics analysis and polymorphisms were analysed using dbSNP (Build 135) unless specified otherwise.

2.2.1 Library preparation and sequencing

Library preparation was carried out following the protocol 'Ion AmpliSeq™ DNA and RNA Library Preparation' Publication number MAN0006735, Revision B.0 (Life Technologies). Reagents were provided in the Ion AmpliSeq Library Kit 2.0 (Life Technologies).

2.2.2 Amplification of targets

In brief, patient gDNA was diluted to a final concentration of 10 ng/µL. Targets were amplified using 1X Ion AmpliSeq Primer Pool (1 or 2), 1X Ion AmpliSeq HiFi Mastermix, 10 ng patient gDNA and nuclease free water to a final volume of 20 µL. Samples were flick mixed and spun down. Reactions were cycled in a Veriti Thermocycler (Applied Biosystems, California, USA) under the cycling conditions outlined in **Table 2.2.**

Table 2.2: PCR Amplification of target regions for library construction with AmpliSeq multiplex primer pools and AmpliSeq Library Preparation Kit

Stage (Repeats)	Temperature (˚C)	Time
Activation (1 x)	99	2 minutes
Denaturation, Annealing and Extension (15 x)	99 60	15 seconds 8 minutes
Hold (1 x)	25	Up to an hour

2.2.3 Amplification of targets

FuPa Reagent (2 µL; concentration not provided) was added to each sample in order to partially digest primer sequences and phosphorylate the amplicons. Samples were flick mixed and spin down. Libraries were incubated in a Veriti Thermocycler as outlined in **Table 2.3.**

Table 2.3: Incubation regime for partial digestion of primer sequences for the generated AmpliSeq libraries

Temperature (˚C)	Time
50	10 minutes
55	10 minutes
60	20 minutes
25	Up to an hour

2.2.4 Partial digestion of primer sequences

Each patient library was assigned a barcode (Ion Xpress™ Barcode Adapters 1-16 and 17-32 Kits, Life Technologies). For each barcode, a mix of Ion P1 Adaptor and Ion Xpress Barcode was prepared at a final dilution of 1:4 for each adaptor. To each library, 4 µL of switch solution, 2 µL DNA Ligase and 2 µL of the Adaptor/Barcode mix was added (concentrations not provided), flick mixed and spun down. Libraries were incubated in a Veriti Thermocycler as outlined in

Table 2.4**.**

Table 2.4: Incubation regime for ligation of adaptors and barcodes to generated AmpliSeq libraries.

Temperature (˚C)	Time
22	30 minutes
72	10 minutes
25	Up to an hour

2.2.5 Purification of the library

In brief, Agencourt AMPure XP Reagent (Beckman Coulter, California, USA) was vortexed and 1.5X sample volume was added to each library. Samples were flick mixed and briefly spun down. Samples were incubated for 5 minutes at room temperature (RT) and placed in a magnetic rack. Supernatant was removed and discarded. Ninety percent ethanol was added to each tube and washed by moving the tube side to side 5 times. Supernatant was removed and repeated for a second wash. All ethanol was removed, and bead pellet was air-dried for 5 minutes. Tubes were removed from the magnetic rack and the DNA was eluted in 50 µL of Platinum PCR Supermix High Fidelity (concentration not provided) along with 2 µL of Library Amplification Primer Mix (concentration not provided). Each sample was flick mixed and briefly spun down. Tubes were placed in a magnetic rack and

supernatant was removed and transferred to PCR tubes. Libraries were cycled in a Veriti Thermocycler as outlined in **Table 2.5.**

Table 2.5: Cycling regime for secondary amplification of purified AmpliSeq libraries

Stage (Repeats)	Temperature (˚C)	Time
Activation (1 x)	99	2 minutes
Denaturation, Annealing and Extension (5 x)	99 60	15 seconds 1 minute
Hold (1 x)	25	Up to an hour

To each amplified library, 0.5X sample volume Agencourt AMPure XP Reagent was added, flick mixed and then briefly spun down. Samples were incubated for 5 minutes at RT, and then placed in a magnetic rack for 5 minutes. The supernatant was removed and aliquoted to new tubes. 1.2X original sample volume of Agencourt AMPure XP Reagent was added to the supernatant, flick mixed and spun down. Samples were incubated for 5 minutes at RT and then placed in a magnetic rack for 3 minutes. The supernatant was removed and discarded, 90 % ethanol was added to each bead pellet and washed by moving the tubes side to side 5 times. The supernatant was removed and repeated for a second wash. Supernatant was removed, and the bead pellets were air dried for 5 minutes. Tubes were removed from the magnetic rack and the libraries were eluted in 50 µL of Low TE Buffer. Tubes were flick mixed and spun down, placed in the magnetic rack and the supernatant was transferred to new tubes. Aliquots of the library were quantified using both the Qubit dsDNA high sensitivity assay (ThermoFisher Scientific, USA) and the BioAnalyser (Agilent, California, USA) or LabChip (PerkinElmer, Massachusetts, USA) as outlined in **Section 2.2.6.**

2.2.6 Library quantification

To determine the size distribution of the libraries, the Agilent High Sensitivity DNA Chips were used on the Agilent 2100 BioAnalyser (Agilent) as per the manufacturer's protocol (G2938-90321 Rev. B, Agilent). Libraries were also quantitated using the LabChip (PerkinElmer) by the Flinders Genomics Facility (Flinders University, South Australia), following the manufacturers protocol.

2.2.6.1 Quantification of libraries via qPCR

Quantification of the library was carried out via qPCR on the ViiA 7 qPCR machine (Applied Biosystems) in a 384 well plate. Analysis was performed with the ViiA7 RUO software. qPCR products

were measured using a TaqMan fluorescent probes, which was supplied as part of the Ion Library Quantitation kit (Applied BioSystems). In brief, a standard curve was generated with 10-fold serial dilutions of an *Escherichia coli* DHB10B control library at 5.8 pM, 0.68 pM and 0.068 pM. Samples were diluted 1:100 in nuclease free water and 9 µL of each diluted sample or standard was combined with 1X Ion Library qPCR mastermix and 1X Ion Library TaqMan Quantitation Assay in a final volume of 20 µL. Samples and standards were analysed in duplicate and run in the ViiA7 as outlined in **Table** 2.6.

Table 2.6: qPCR cycling conditions for quantification of generated AmpliSeq libraries.
UDG; Uracil-DNA glycosylase

Stage (repeats)	Temperature (°C)	Time
Hold (UDG incubation)	50	2 minutes
Hold (Polymerase activation)	95	20 seconds
Cycle (40 x)	95 60	1 second 20 seconds

2.2.7 Amplification of Targets from low concentration libraries

Once quantified, libraries that fell below the specified concentration were reamplified prior to being run on the PGM (1000-5000 pm for BioAnalyser, 300-1500 ng/mL for Qubit). Twenty-five microliters of each library was combined with 75 µL Platinum PCR Supermix High Fidelity and 3 µL Library Amplification Primer Mix. Tubes were flick mixed and cycled in a Veriti Thermocycler as outlined in **Table 2.7.**

Table 2.7: Cycling regime for amplification of low concentration AmpliSeq libraries.

Stage (Repeats)	Temperature (˚C)	Time
Activation (1 X)	98	2 Minutes
Denaturation, Annealing and Extension (10 X)	98 60	Seconds 1 Minute
Hold (1 X)	25	Up to an hour

To purify the reamplified libraries, 150 µL Agencourt AMPure XP Reagent was added to each sample, flick mixed and spun down. Samples were incubated for 5 minutes at room temperature and then placed in a magnetic rack for 3 minutes. The supernatant was removed and discarded. Ninety percent ethanol was added to each bead pellet and washed by moving the tubes side to side 5 times.

The supernatant was removed and repeated for a second wash. Supernatant was removed, and the bead pellets were air dried for 5 minutes. Tubes were removed from the magnetic rack and the Libraries were eluted in 50 µL of Low TE Buffer. Tubes were flick mixed and spun down, placed in the magnetic rack and the supernatant was transferred to new tubes. Aliquots of the re-amplified libraries were re-analysed as outlined in **Section 2.2.6.**

2.2.8 Ion PGM and Ion Proton initialisation and sequencing

Samples were sequenced either on the Ion 318 Chipv2 on the Ion Torrent Personal Genome Machine (PGM; Life Technologies) by the Flinders Genomics Facility (Flinders University, South Australia), or on the Ion P1 chip on the Ion Proton (Life Technologies) by the Lottery West State Biomedical Genomics Facility at the University of Western Australia.

2.2.9 Bioinformatics analysis

For the analysis of Ion Torrent generated sequencing data, a bioinformatics pipeline was developed for both IonReporter (Life Technologies) and CLC Genomics Workbench (QIAGEN, Hilden, Germany) The development and optimisation of the bioinformatics pipeline is outlined in **Chapter 3**.

2.2.9.1 *in silico* analysis

From the variants identified for each individual, those variants considered common within the general population (defined by a minimum allele frequency of MAF >5%) were discarded. The remaining variants were analysed according to their presence in various databases (COSMIC, dbSNP, gnomAD) and their predicted effect on protein function (Polyphen-2, PROVEAN, SIFT, Align-GVGD and Protein domain analysis). Detailed analysis of the selected databases is included in **Chapter 5**. Selected variants of interest detected by MPS were confirmed by Sanger sequencing.

2.3 General molecular biology methods

2.3.1 Genomic DNA isolation

Cells were pelleted via centrifugation for 5 minutes at 500 *x g* at 4 °C, and supernatant was discarded. The pellet was washed twice with ice cold phosphate buffered saline (PBS), with repeated centrifugation and removal of supernatant. Cells were resuspended in 1 volume digestion buffer (0.3 mL for < $3x10^7$ cells, 1 mL for > $3x10^7$ cells). Pellet was flick mixed and incubated at 50 °C for 12-18 hours with shaking. An equal volume of phenol/chloroform/isoamyl alcohol (Sigma-Aldrich,

Missouri, USA) was added to the sample and spun at 1700 *x g* for 10 minutes at 4 °C. The aqueous layer was removed and transferred to a new tube, and half the volume of 7.5 M ammonium acetate and 2 volumes (of original amount of top layer) of 100 % ethanol was added. DNA was recovered by centrifugation at 2100 *x g* for 10 minutes at 4 °C. The pellet was rinsed with 70 % ethanol, flick mixed and centrifuged for 2 minutes at 2100 *x g*, 4 °C. All ethanol was decanted, and the pellet was air dried for at least 5 minutes. DNA was resuspended in Tris-EDTA (TE) buffer and shaken gently at RT for 4-6 hours to facilitate solubilisation. DNA was quantified using the Nanodrop 1000 (ThermoFisher) and stored at 4 °C

2.3.2 Primer Design and Optimisation

PCR primers were designed by eye or through the use of online primer design tools, which included programs such as Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) or Primer3 (http://www.bioinformatics.nl/cgi-bin/primer3plus). Primer pairs were positioned to span regions of approximately 200-800bp, ensuring that the designed amplifiable region incorporated the variant of interest. Sequences were submitted to the Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/Blast.cgi) to predict if primers would result in non-specific products by binding to other regions within the genome. dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) was also utilised to ensure primer binding sites did not include any known polymorphisms. Synthesised primers were provided lyophilised (Integrated DNA Technologies (IDT), Singapore) and were subsequently resuspended in sterile water at a final concentration of 100 µM and stored at -20 °C. Primer sequences are listed in **Appendix F.**

Primers were optimised using genomic DNA extracted from the control cell lines, FH9 and HEK293. All primer sets were optimised individually using a standard or touchdown thermocycling regime. All optimised cycling conditions for each primer set included in **Appendix G.**

2.3.3 Polymerase Chain Reaction

Unless specified otherwise, PCR reactions were set up at a final concentration of 1X PCR Buffer (Applied Biosystems) 1.5 mM $MgCl_2$ (Applied Biosystems), 0.4 mM dNTPs (Life Technologies), 0.8 mM forward primer, 0.8 mM reverse primer (IDT), using 2 units AmpliTaq Gold® Polymerase (Applied Biosystems) or Platinum Taq (Applied Biosystems) plus 2 µL template, in a final volume of 25 µL. Reactions were undertaken in a Veriti Thermocycler (Applied Biosystems) using the cycling conditions described in **Appendix G**.

2.3.4 Agarose gel electrophoresis

PCR, plasmid, gDNA and RNA products were routinely visualised by agarose gel electrophoresis. For making and running gels, a 1X TAE Buffer (**Section 2.5.1**) was used. For analysis of PCR products, gDNA and RNA products a 1.0 % -2.5 % agarose (Scientifix, Victoria, Australia) in TAE buffer was used. For analysis of larger products, predominantly plasmids, a 0.8 % agarose in TAE was used. For the visualisation of products, 1X GelRed (Biotium, California, USA) was added to molten agarose. Products were subsequently viewed using the GeneGenius Imaging System (SynGene, India) unless specified otherwise.

2.3.5 PCR Product Purification

2.3.5.1 Enzymatic purification of PCR Products

PCR products were purified by treatment with Shrimp Alkaline Phosphatase (GE Healthcare, Australia) to degrade residual dNTPs, and exonuclease I (New England Biolabs, NEB, Massachusetts, USA) to degrade single stranded DNA such as residual PCR primers. In brief, 5 Units of exonuclease I (NEB), 1 Unit of shrimp alkaline phosphatase (SAP, NEB) and 1X SAP reaction buffer (GE Healthcare) were added to 5 µL of PCR product and the reaction was incubated at 37°C for 60 minutes. Enzymes were heat inactivated by incubation to 80°C for 20 minutes.

2.3.5.2 Commercial kit for clean-up of PCR products

Additionally, PCR products were purified through various commercial kits that utilised silica membrane-based purification methods. Post visualisation, PCR products were cleaned using the QIAquick PCR purification kit (QIAGEN) as per the manufacturer's protocol. PCR products that were excised from the agarose gels were purified using the QIAquick gel extraction kit (QIAGEN) as per the manufacturer's protocol.

2.3.6 Sanger sequencing

The concentration of the PCR products was determined through agarose gel electrophoresis and comparison to known DNA standards (DMW-100L Ladder (Gene Works, Adelaide), 500 bp DNA ladder (Gene Works, Adelaide), 1 kb DNA ladder (Promega, Wisconsin, USA) 100 bp DNA ladder (Promega), 100 bp DNA ladder (NEB) and 2-log DNA ladder (NEB)). The samples were diluted accordingly to give a concentration of 10 ng/100 bp of product (i.e. 30 ng for 300 bp product). For DNA sequencing, separate aliquots of forward and reverse primers at a concentration of 5 µM, and

the PCR product were provided to the SA Pathology DNA Sequencing Facility for dye terminator sequencing.

2.3.7 RNA Extraction

RNA was extracted using TRI-Reagent. This method involved the addition of 1 mL TRI-Reagent (Sigma-Aldrich) per $1x10^7$ cells and incubated at RT for 5 minutes, Chloroform (100%, Sigma-Aldrich) was added at a ratio of 200 µL per 1 mL TRI-Reagent and samples were mixed vigorously and incubated at RT for 5 minutes. Samples were centrifuged at 1850 x *g* for 15 minutes at 4 ˚C to form a gradient. The top aqueous layer was transferred to a sterile tube, avoiding the interphase. RNA was precipitated through the addition of isopropanol (Sigma-Aldrich) at a ratio of 500 µL per 1 mL TRI-Reagent. Samples were mixed through gentle inversion and incubated at RT for 10 minutes followed by centrifugation at 1850 x *g* for 30 minutes at 4 ˚C. The RNA precipitate formed a pellet on the bottom of the tube, the supernatant was removed, and the RNA pellet was washed with 75% v/v ethanol solution (Sigma-Aldrich). The pellet was mixed vigorously, followed by centrifugation at 1850 x *g* for 10 minutes at 4 ˚C. The supernatant was removed, and the pellet was air dried for approximately 15 minutes. The RNA pellet was resuspended in 50 µL diethyl-pyrocarbinate (DEPC) treated water (Sigma-Aldrich). Samples were stored at -80 ˚C until required. Gloves and pipettes were cleaned with RNase Zap® (Sigma-Aldrich) prior to RNA extraction.

2.3.8 DNA Degradation

All RNA was DNaseI treated to ensure any DNA carried through the extraction process was digested. RNA samples were DNaseI treated as per the manufacturer's instructions for the 'DNA-free kit' (Life Technologies, Australia). In brief, 1X DNaseI Buffer and 2 units of rDNase was added to the RNA sample, mixed gently and incubated at 37 ˚C for 25 minutes. Post incubation, 0.2X volume DNase Inactivation reagent was added, mixed vigorously and incubated at RT for 2 minutes. Samples were centrifuged at 10,000 x *g* for 90 seconds. The supernatant was removed and taking care to avoid the pellet and the supernatant was transferred to a sterile 1.5 mL Eppendorf tube. Samples were stored at -80 ˚C.

2.3.9 Nucleic acid quantification

RNA quantifications were measured spectrophotometrically on a Nanodrop-1000 following the manufacturer's instructions (ThermoFisher Scientific, Australia). Samples (neat and 1:10 diluted in DEPC treated H_2O) were quantified in duplicate and an average was calculated.

2.3.10 Complementary DNA (cDNA) generation

Patient material or RNA extracted from cell lines was used for the generation of cDNA using a maximum of 2 µg total RNA per reaction. For cDNA generation, 0.5 mM dNTPs (Invitrogen, ThermoFisher Scientific, Australia) and 10 ng/µL random primers (Invitrogen) were added to RNA in a total volume of 12 µL and the reaction was incubated at 65 ˚C for 5 minutes, snap frozen on ice and spun down to collect the sample to the base of the tube. To this, 1X first strand buffer (Invitrogen), 0.04 M Dithiothreitol (DTT, Invitrogen) and 40 units RNase Out (Invitrogen) were added to achieve a total volume of 19 µL. Samples were incubated at 25 ˚C for 2 minutes, followed by the addition of 10 Units of SuperScript® II Reverse Transcriptase (Invitrogen). Samples were incubated at 25 ˚C for 10 minutes, 42 ˚C for 50 minutes and 70 ˚C for 15 minutes in Veriti Thermal Cycler (Applied Biosystems). Absence of genomic DNA in RNA preparations was verified by performing replicate reactions with the omission of Reverse Transcriptase enzyme. Samples were diluted with TE buffer and stored at -20 ˚C.

2.3.11 Real-Time PCR (RT-PCR)

Standard reactions were carried out using a ViiA 7 qPCR machine (Applied Biosystems, Australia) in the 384 well format (Applied Biosystems), and performed using ViiA 7 RUO software (Applied Biosystems). The synthesis of dsDNA products during real-time PCR was measured using SYBR-Green (Applied Biosystems, Australia) intercalating dye. A 2X SYBR Green Mastermix (Applied Biosystems) was used for RT-PCR, containing DNA polymerase UP, dNTPs, Uracil-DNA glycosylase (UDG) and ROX reference dye in buffer. Primer and template master mixes were prepared immediately prior to the experiment, with a final concentration of 1X SYBR-Green Master mix, 2 µM of each primer (IDT, Singapore) and various template concentrations ranging from undiluted to 10^{-5} diluted depending on the sample in a total volume of 10 µL. Reactions were prepared in triplicate and each reaction run included Reverse Transcriptase negative, no template and genomic DNA controls to monitor for reagent contamination and primer specificity. All experiments were carried out under the thermal cycling conditions detailed in **Table 2.8** unless stated otherwise.

Table 2.8:Real-time PCR cycling method in ViiA 7. UDG; Uracil-DNA glycosylase

Stage	Repeats	Temperature (°C)	Time
UDG activation	1	50	2 minutes
UDG inactivation; Taq polymerase activation	1	95	2 minutes
Amplification and Extension	40	95	15 seconds
		60	1 minute
Melt Curve	1	95	15 seconds
		60	Heat to 95°C at 0.05°C/second
		95	1 second

2.3.11.1 Real-time PCR analysis

Individual cycle quantification (C_q) values were obtained by setting a threshold manually. Data from ViiA 7 was imported into Microsoft Excel and relative expression levels were calculated using the $2^{-\Delta Ct}$ method. For standard curves, C_q values were used to determine the M-value of the primer pair. The M value is defined as the number of cycles to produce 10 times the amount of template, with the theoretical value being 3.2.

2.4 Cell Culture Methods

All cell culture methods were supplied by Ms Monica Dreimanis (Department of Molecular Medicine and Pathology, Flinders Medical Centre, Flinders University) unless otherwise indicated. All cell culture was performed in a Class I Laminar Flow Hood or a Biosafety Hood as appropriate.

2.4.1 Thawing cells from liquid nitrogen

Cells were removed from liquid nitrogen storage and transferred immediately to 37 °C water bath for rapid thawing of the cells. The cell suspension was transferred to a sterile container and 10 mL of appropriate media was added drop-wise to the cell suspension over 10 minutes. Cells were pelleted by centrifugation 500 x *g* for 5 minutes. Supernatant was removed, and the cell pellet was resuspended in 5 mL of appropriate media for assessment of cell viability.

2.4.2 Subculturing adherent cells.

Cells were removed from the incubator and media was aspirated from the flask. The monolayer was gently rinsed with PBS (5 mL T25, 10 mL T75), rocking the flask back and forwards several times. PBS was aspirated and prewarmed 3 mL 0.05 % Trypsin-EDTA (Sigma-Aldrich) was added to the cells.

Flasks were incubated at RT for 5 minutes (HEK293) or at 37 °C for 15 minutes (MCF10A cells). PBS was added to the flask and cells were transferred to a 15 mL falcon tube and centrifuged at 500 x *g* for 5 minutes (HEK293) or 125 x *g* for 10 minutes (MCF10A). Supernatant was aspirated, and the cell pellet was resuspended in the appropriate media as outlined in **Sections 2.4.5.1** and **2.4.5.2** for HEK293 and MCF10A cells respectively.

2.4.3 Freezing mammalian cell lines

For subsequent retrieval and continuation of culturing, mammalian cell lines were periodically frozen in liquid nitrogen. Cells were counted (see **Section 2.4.4** below) and pelleted by centrifugation. Cells were re-suspended in a final concentration of 15 % DMSO and 25 % FCS in appropriate media at a maximum concentration of 1 x 10^7 cells/mL. Four hundred microlitres of cell suspension was subsequently added to a cryo-vial and stored at -80 °C for a minimum of 24 hours. After 24 hours, samples were transferred to liquid nitrogen storage until required.

2.4.4 Cell counting and viability by Trypan blue exclusion

Cells to be counted were diluted 1:2 or 1:10 as required in in 0.4 % w/v Trypan Blue in PBS (Bio-Rad, California USA). Cells were added to a Neubauer chamber haemocytometer and the number of stained cells (dead cells) and the total number of unstained cells (viable cells) were counted in four 1 mm^2 areas. This value was then divided by four to provide the average number of cells per 1 mm^2. The number of cells per mL was then calculated using $c = n \times d \times 10^4$ where c= concentration of cells/mL, n = average number of cells/mm^2 area and d = dilution.

Alternatively, samples were diluted 1:2 with trypan blue and added to a dual chamber counting slide (Bio-Rad). Cells were counted using the Bio-Rad TC20 automatic cell counter (Bio-Rad). Cells were gated at an appropriate size (4-16 µm for HEK293 and 5-20 µm for MCF10A cells).

2.4.5 Cell Lines

2.4.5.1 Human Embryonic Kidney Cells (HEK293)

Human Embryonic Kidney Cells (HEK293; ATCC® CRL-1573™) were cultured in Dulbecco's Modified Medium (High/Low Glucose DMEM; Sigma Aldrich) supplemented with 10 units/mL Penicillin, 0.1 mg/mL Streptomycin (Sigma Aldrich), 2mM L-Glutamine (Sigma Aldrich) and 10 % Foetal Calf Serum (FCS; Bovogen Biologicals, VIC, Australia). Media was filtered by passing through a 0.22 µm filter (Millipore). Cells were grown in standard conditions (5 % CO_2 at 37 ˚C), in either 10 mL or 20 mL

appropriate media in T25 or T75 flasks respectively. Cells were passaged every 3-4 days when confluent at an approximate ratio of 1:5.

2.4.5.2 Human Breast Epithelial Cells (MCF10A)

MCF10A (ATCC® CRL-10317™) were cultured in Mammary Epithelial Cell Growth Medium (MEBM; Lonza) with Bovine Pituitary Extract (BPE), human epidermal growth factor (hEGF), Insulin, Hydrocortisone, GA-100 (All provided in the MEBM Bullet Kit, concentrations not provided) plus 100 ng/mL Cholera Toxin (Sigma-Aldrich). Cells were grown in standard conditions, (5 % CO_2 at 37 ˚C), in either 5 mL or 12 mL media in T25 or T75 flasks respectively. Cells were passaged every 3 to 4 days, or when >80% confluent, at a ratio of 1:3.

2.4.6 Mycoplasma screening of mammalian cells

Cell lines were screened for Mycoplasma contamination upon establishment in culture, and routinely screened every 3 months whilst in use. Cell medium was screened for the presence of Mycoplasma metabolites thought the Mycoplasma detection kit DigitalTest v2.0 (Biotools.com, Texas, USA). Cell culture media and media alone (negative control) was analysed for the presence for metabolites through a spectrophotometric analysis as per the manufacturers' instructions.

2.5 Buffers

2.5.1 General Buffers and Solutions

Phosphate Buffered Saline	137 mM NaCl
(PBS)	4.3 mM $NaHPO_4$
	1.4 mM KH_2PO_4
	2.7 mM KCl
1X Tris-acetate-EDTA	40 mM Tris-Acetate
(1X TAE)	2 mM EDTA, pH 8.0
	0.001% v/v glacial acetic acid
Tris-EDTA buffer	10 mM Tris, bring to pH 8.0 with HCl
(TE Buffer)	1 mM EDTA
Digestion Buffer	100mM NaCl
For gDNA extraction	10mM TrisCl (pH 8)
	25mM EDTA (pH 8)
	0.5% SDS
	0.1mg/mL Proteinase K

2.5.2 Buffers for CRISPR/Cas9 editing

Sigma Annealing Buffer	10mM Tris-Buffer
	50mM NaCl
	1mM EDTA
IDT Duplex Buffer	100mM potassium Acetate
	30mM HEPES, pH 7.5

LB Media	10g Tryptone
(Lysogeny broth)	5g Yeast Extract
	10g NaCl
	In total volume of 1L MilliQ H_2O
SOC Medium (Super optimal	0.5% Yeast Extract
broth with catabolite	2% Tryptone
Repression)	10 mM NaCl
	2.5 mM KCl
	10 mM MgCl2
	10 mM MgSO4
	20 mM Glucose

2.5.3 Flow Cytometry Buffers

Fixation buffer	100% methanol at -20°C
Permeabilisation Buffer	0.1% Tween-20
	0.1% Sodium Citrate
	1X PBS
Blocking Buffer	1X PBS
	4% BSA

2.5.4 Western Blot Analysis Buffers

10X Running Buffer	25 mM Tris
	192 mM Glycine
	0.1% SDS

2X Laemmli Buffer	125 mM Tris–HCl, pH 6.8
	20% glycerol
	4% SDS
	0.1% bromophenol blue
	5% β-Mercaptoethanol
Blocking Buffer	5% v/v skim milk powder
	0.1% Tween-20
	100 mL PBS
Tris Buffered Saline (TBS)	20 mM Tris-Cl
	150 mM NaCl
	Adjust pH to 7.6 with HCl
Ab Dilutent	5 mL TBS
	5 mL blocking buffer
	Antibody (Primary/Secondary at appropriate volume)
Wash Buffer	TBS
	1% Tween-20

Chapter 3:
Development of a bioinformatics pipeline for analysis of Ion Torrent sequencing data

3.1 Introduction

In a clinical setting, laboratories are transitioning from the gold-standard Sanger sequencing *BRCA* protocol to a more cost- and time-effective MPS analysis. This high throughput approach allows for massively parallel processing of highly multiplexed PCR reactions within a single platform. These platforms often differ in their sequencing chemistries but share the technical paradigm of MPS through clonal amplification of DNA templates. The demand for high throughput, low-cost sequencing has driven the development of multiple platforms that produce thousands of sequences concurrently, with some platforms having the potential to run as many as 500,000 sequencing by synthesis reactions in parallel.

3.1.1 Comparison of *BRCA1* and *BRCA2* sequencing data generated with MPS and Sanger sequencing

The first step in determining the applicability of using the 51 gene panel for detecting sequence variants was to develop an optimised bioinformatics analysis pipeline. This pipeline was developed for the analysis of Ion Torrent data, in order to address the specific sequencing errors known to be associated with Ion Torrent sequencing. This optimisation was needed as the developed best-practices workflows, such as the genome analysis tool kit (GATK, Broad Institute, MIT), are optimised for Illumina sequencing (McKenna *et al.*, 2010, DePristo *et al.*, 2011). Therefore, the development of an in-house pipeline allowed for correction of any potential errors not only for the custom gene panel used within this study, but also those associated with Ion Torrent sequencing. This process was carried out by comparison with the Sanger sequencing data already obtained for *BRCA1* and *BRCA2*.

A major issue associated with MPS analysis is the high rate of false-positive detection (McCall *et al.*, 2014, Mu *et al.*, 2016). In order to optimise the pipeline for minimal false-negatives and false-positives, a pilot study on a relatively small number of patients was carried out. DNA from 13 individuals, whose *BRCA1* and *BRCA2* gene sequences had already been analysed by Sanger sequencing by the SA Pathology diagnostic department, were selected for sequencing with the MPS gene panel. Whilst these individuals were not found to carry any pathogenic *BRCA1/2* mutations, all benign polymorphisms in these genes were annotated, providing a comprehensive panel of sequence variants with which to optimise the Ion Torrent analysis pipeline. Two commercially available sequence analysis programs (IonReporter and CLC Genomics Workbench) were compared to determine which program gave the best sensitivity and specificity.

In addition, this initial analysis was used to determine the optimal number of individual sequencing libraries that were able to be multiplexed on a single Ion Torrent PGM sequencing chip. The aim was to provide both the minimum coverage required for calling variants with a high level of confidence (which at the commencement of this study was recommended to be a minimum of 100X coverage for germline mutations (Chan *et al.*, 2012)), whilst also providing greatest value for money.

3.1.2 Aims

The aims of this chapter were to:

1. Analyse the pilot Ion Torrent sequencing data with 2 commercially available bioinformatics programs, IonReporter and CLC Genomics Workbench.
2. Utilise the *BRCA1* and *BRCA2* variants previously identified by Sanger sequencing to optimise the bioinformatics pipeline and determine the utility of the AmpliSeq gene panel.
3. Determine the maximum number of DNA samples that can be multiplexed in a single sequencing run whilst still achieving optimal coverage for germline mutation analysis.

3.2 Methods

3.2.1 DNA Samples

Thirteen patient samples were selected for this pilot study. DNA extracted from peripheral blood (representing germline DNA) was kindly provided by SA Pathology. Diagnostic mutation analysis for *BRCA1* and *BRCA2* had previously been carried out by Sanger sequencing. All identified sequence variants in these genes had been annotated for each patient sample and this information was also provided by SA Pathology. Further information on these patients, including Manchester scores and sequencing data, can be found in **Chapter 5**, with all individual Manchester scores included in **Appendix C** and all MPS sequence variants included in **Appendix H.**

3.2.2 AmpliSeq library preparation and sequencing

AmpliSeq library preparations were carried out by Dr. Renee Smith in the Flinders Genomics Facility as outlined in the 'Ion AmpliSeq Library Preparation' publication number MAN0006735, Revision 6 (Life Technologies, USA). Reagents were provided in the Ion AmpliSeq Library Kit 2.0 (Ion Torrent, Life Technologies). In order to multiplex samples, libraries were barcoded with the IonXpress Barcode Adapters 1- 32 Kit (Life Technologies, USA).

The size distribution of each library was determined on Agilent High Sensitivity DNA chips on the Agilent 2100 BioAnalyser (Agilent Technologies) as per the manufacturer's protocol (G2938-90321 Rev. B, Agilent Technologies). Libraries were gated from 150-330 bp to quantify only the amplified library. Library concentrations were measured through fluorimetry using the dsDNA High Sensitivity Assay and the Qubit 2.0 Fluorometer as per the manufacturer's protocol (MAN0002326, Life Technologies).

Libraries were then diluted to 10 pM and pooled at equimolar concentrations. Three samples were pooled and sequenced on the first chip and 10 samples were pooled and sequenced on the second chip. All remaining template preparation and sequencing was carried out by Flinders Genomics Facility. In brief, template-positive Ion Sphere Particles (ISPs) were generated via emulsion PCR on the Ion Torrent One Touchv2 (OT2, Life Technologies) as per the manufacturer's instructions. The number of ISPs with template attached was determined with the Qubit Ion Sphere quality control kit (Life Technologies), followed by the selective isolation and enrichment of ISPs with clonally

amplified DNA on the Ion Torrent One Touch (Life Technologies). Sequencing was carried out on Ion318v2 chips on the Ion Torrent PGM according to the manufacturer's instructions.

3.2.3 Data analysis

3.2.3.1 Ion Torrent software analysis

Initial data analysis was carried out using the Ion Torrent Software Suite. Following sequencing, multiplexed data were deconvoluted by grouping sequences based on barcode sequences, which were trimmed and removed (Torrent Browser, v2.2). Reads with a quality score (Phred score) of less than Q20 (representing a mismatch rate of 1 in 100) were then removed. Library sequencing was deemed successful if it returned a minimum of 300,000 reads with a Phred score of Q20 or above. Sequences were then aligned to the human genome reference sequence (hg19/GrCh37). Following alignment, the program performed automated target region coverage analysis and automatically removed regions of poor quality. Run metrics including chip loading efficiency, total read counts and run quality information were also generated. All data were then downloaded from the Ion Torrent server as .fastq files for further analysis with the IonReporter (v4.0, Life Technologies) and CLC Genomics Workbench (v5.0, QIAGEN) programs.

3.2.3.2 IonReporter Analysis

Ion Reporter analysis was carried out using both the Germline High and Low Stringency parameters (**Table 3.1**).

Table 3.1: Germline High-Stringency and Low-Stringency Parameter Settings for Variant Caller: SNP, single nucleotide polymorphisms; Indel, insertion and deletions.

	High Stringency Parameters		Low Stringency Parameters	
Parameter	**SNP**	**Indel**	**SNP**	**Indel**
Minimum coverage each strand	3	3	0	5
Minimum variant score	10	10	10	10
Minimum read proportion	0.15	0.15	0.1	0.1
Minimum total coverage	20	20	6	15
Maximum strand bias	0.95	0.85	0.95	0.85

Variants identified by each of these filtering settings were then analysed through the 'Annotate Variants: Single Sample' workflow, which generated a list of all polymorphisms, insertions and deletions detected within each patient sample. *BRCA1* and *BRCA2* variants from both analysis

pipelines were then compared to those that had previously been identified through Sanger sequencing for the same patient sample. In addition, the Ion Torrent *BRCA1* and *BRCA2* sequencing data for all samples were imported into the Integrative Genomics Viewer (IGV) software (Broad Institute, USA) enabling visualisation of the raw sequencing reads.

3.2.3.3 CLC Genomics Workbench analysis

Ion Torrent sequencing data were imported into the CLC genomics workbench for trimming, mapping and variant calling (**Figure 3.1**). Sequencing reads were trimmed to remove any remaining adaptor sequences. Reads with a length of less than 10 bases were discarded. Trimmed data were then mapped to the human reference genome (hg19/GrCh37) using a minimum length fraction and similarity fraction of 0.95 and 0.9 respectively. All other parameters were left as default. Mapped data were then filtered, to show only the variants that mapped to the regions covered by the AmpliSeq panel. Any variants that did not lie within these regions were masked. Coverage statistics for the regions of interest were generated through the Targeted Regions Coverage report tool using default settings.

Variant calling was carried out using the inbuilt Probabilistic and Qualitative methods. Probabilistic variant analysis identifies changes based on depth of coverage and Qualitative variant analysis identifies changes based on the Phred score of the bases surrounding the potential variant. Filtering parameters for these two pipelines were first run as default and then optimised through comparison to the previously documented *BRCA1* and *BRCA2* variants for each individual. Stringency settings were altered in order to maximise the ability of the MPS to detect the variants identified by Sanger sequencing. The filtering parameters used for Probabilistic and Qualitative variant analysis are outlined in **Table 3.2** and **Table 3.3** respectively.

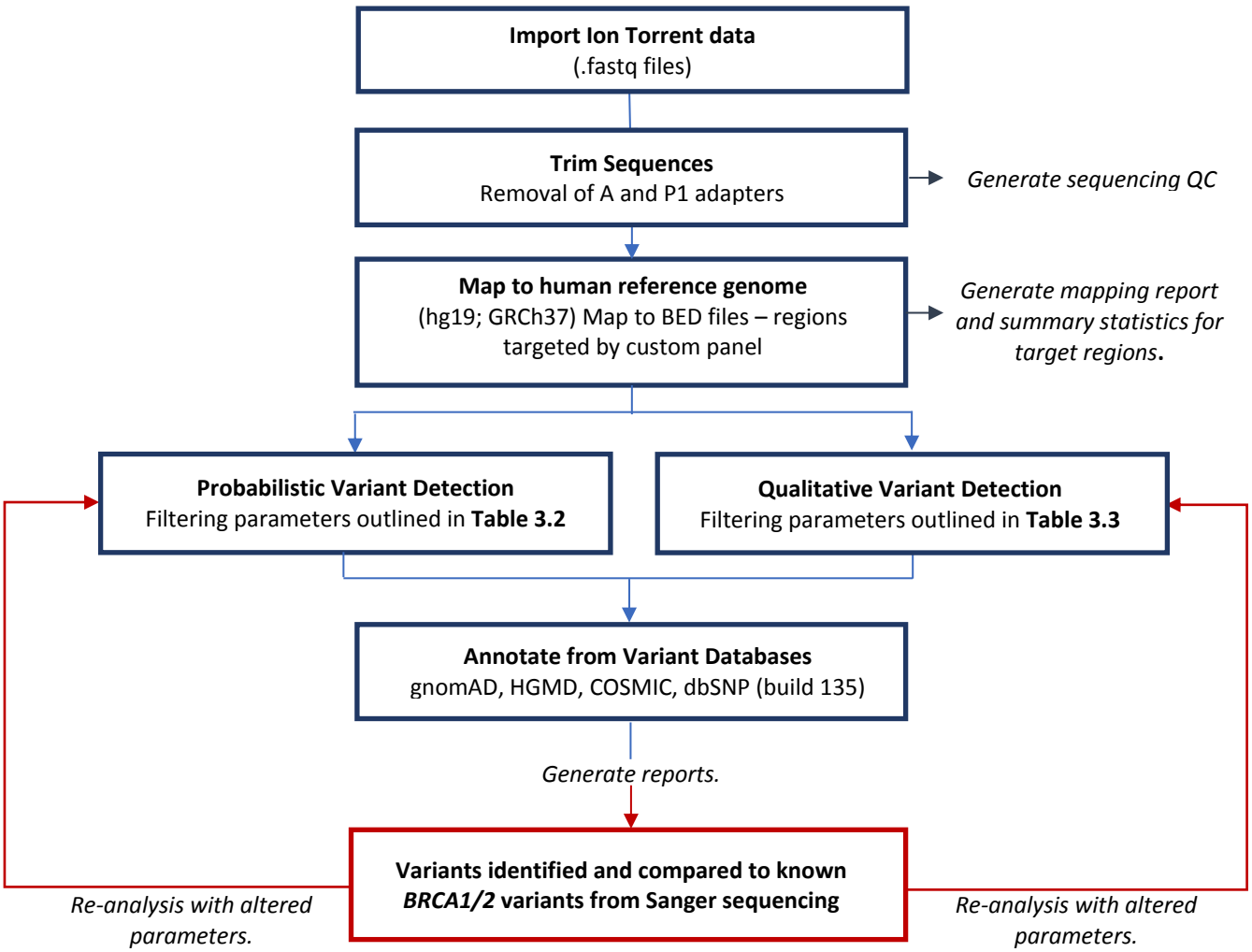


Figure 3.1: CLC Genomics workbench workflow. The loops demonstrate the iterative process undertaken to determine the optimal settings for identifying sequence variants in the Ion Torrent MPS data. gnomAD, Genome Aggregation Database; HGMD, Human Genome Mutation Database; COSMIC, Catalogue of Somatic Mutations in Cancer; dbSNP, SNP database

Table 3.2: Parameters used for optimisation of CLC Probabilistic variant analysis

Variable	Default	Analysis 2	Optimised
Ignore Non-Specific Matches	Yes	Yes	Yes
Minimum Coverage (X)	10	50	50
Variant Probability[1] (%)	90	90	85
Require Presence in Forward and Reverse Reads	Yes	Yes	Yes
Filter 454/Ion Homopolymer Errors	No	Yes	Yes

[1]**Variant Probability**: Minimum value of the variant probability required for the variant to be called.

Table 3.3: Parameters used for optimisation of CLC Qualitative variant analysis.

Variable	Default	Analysis 2	Analysis 3	Optimised
Neighbourhood Radius (bp)	5	5	10	10
Maximum gap and mismatch count[1]	2	2	2	2
Minimum neighbourhood quality [2] (Phred)	15	20	20	20
Minimum central quality[3] (Phred)	20	25	25	25
Ignore Non-specific matches	Yes	Yes	Yes	Yes
Ignore Broken Pairs	Yes	Yes	Yes	Yes
Minimum Coverage	10	10	50	50
Minimum Variant Frequency (%)	35	35	45	35
Maximum Expected Alleles	2	2	2	35
Require Presence in Forward and Reverse Reads	No	Yes	Yes	Yes
Filter 454/Ion Homopolymer Indels	No	Yes	Yes	Yes

[1]**Maximum gap and mismatch count**: This is the number of gaps and mismatches allowed within the length of the read.
[2]**Minimum neighbourhood quality:** The average quality score of the nucleotides in a read within the specified radius has to exceed this threshold for the base to be included in the calculation for this position.
[3]**Minimum central quality:** This allows for reads whose central base quality falls below the specified value being ignored (Qiagen, 2017)

3.3 Results

3.3.1 Concentration analysis of the AmpliSeq Libraries

Thirteen patient sample libraries were generated, and each library DNA yield was quantified with both the Qubit Fluorometer and the Agilent BioAnalyser (**Table 3.4**). According to the manufacturer, AmpliSeq libraries are expected to yield concentrations between 300-1500 ng/mL as determined by the Qubit DNA assay and 2000-10000 pM as determined by the BioAnalyser High Sensitivity DNA Kit.

Table 3.4: Quantification of amplified patient libraries. Libraries with low concentrations (<300 ng/mL and/or <2000 pM) are indicated by blue shaded boxes. Libraries with high concentrations (>1500 ng/mL and/or 10,000 pM are indicated by green shaded boxes. For comparison purposes, libraries which failed to sequence are indicated by a red shaded box in the pool column.

Patient ID	Pool	Qubit concentration (ng/mL)	BioAnalyser concentration (pM)
SABC001	1	874	2691.0
	2	1390	3041.9
SABC002	1	887	1073.6
	2	661	452.6
SABC003	1	1070	2519.0
	2	744	3150.9
SABC004	1	590	3199.8
	2	230	1004.3
SABC005	1	103	351.6
	2	374	2021.9
SABC006	1	1630	2171.8
	2	1490	3838.1
SABC007	1	1620	7284.5
	2	688	4009.6
SABC009	1	464	589.3
	2	184	1136.3
SABC022	1	779	3350.7
	2	1320	8272.1
SABC027	1	950	3296.1
	2	953	3514.2
SABC042	1	1670	6970.0
	2	734	883.2
SABC115	1	724	3613.1
	2	1050	5355.8
SABC124	1	1170	3981.2
	2	1180	1151.6

Three libraries were found to have low concentrations by both Qubit and BioAnalyser analysis (SABC004 pool 2, SABC005 pool 1, SABC009 pool 2), however only 1 of these libraries failed to sequence (SABC004). Six additional libraries were found to have low concentrations only by BioAnalyser analysis, of which 2 did not successfully sequence in this pilot experiment (SABC124 pools 1 and 2). Three libraries were found to have high concentrations only by Qubit analysis (SABC006 pool 1, SABC007 pool 1, SABC042 pool 1), which did not affect their sequencing. An example of BioAnalyser analysis can be seen in **Figure 3.2**.

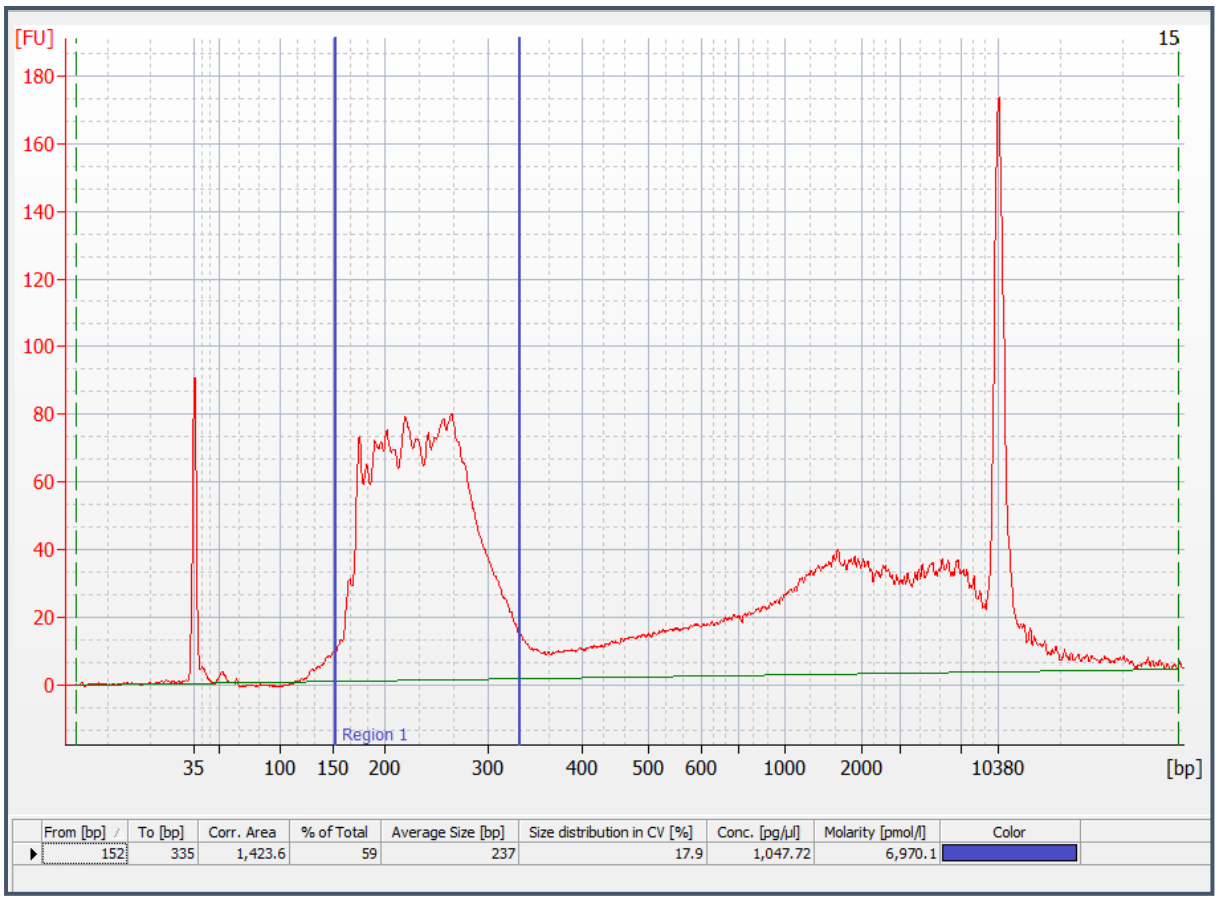


Figure 3.2: Example of a BioAnalyser electrogram an AmpliSeq library. The blue lines flank the expected library amplification and indicate the gated region from which concentration was determined. Peaks at 35 bp and 10380 bp are due to the low and high molecular weight markers which are present in each run, in order to align the sample with the ladder for quantitation. Sample SABC042, Pool 1. X-axis, base pairs; Y-axis, arbitrary fluorescence.

3.3.2 Library sequencing

Following quantification, barcoded libraries were pooled at equimolar concentrations (based on BioAnalyser analysis) and amplified via emulsion PCR on the Ion OneTouch 2 system. A quality control check was carried out on the samples both pre- and post-enrichment to determine the number of ISPs which contained amplified templates (

Table 3.5).

Table 3.5: Pre- and post-enrichment Ion Sphere Particle templating.

	Chip 1		Chip 2	
	Pre-enriched	**Post Enriched**	**Pre-enriched**	**Post Enriched**
Templated ISP	8%	69%	19%	72%

It is recommended that the library samples show approximately 10-25% templated ISPs prior to enrichment, however there is no recommended value for post-enriched samples. Despite pre-enrichment values of templated ISPs falling below recommended guidelines for chip 1, the chip was still sequenced as the PGM sequencer was a recent addition to the Flinders Genomics Facility and the sequencing capacities of this machine were not well understood at the commencement of this study.

3.3.3 Raw sequencing data

Chip 1 contained 3 barcoded patient libraries (SABC002, SABC005, SABC009). This relatively low number of patients was selected as this was the first time this AmpliSeq library had been used and it was unclear if equal coverage could be obtained across all target regions. The first sequencing run was highly successful and resulted in an average coverage of 850X for each of the 3 patients (**Table 3.6**). As the aim was to obtain a minimum of 100X coverage, the remaining 10 libraries were multiplexed on Chip 2, resulting in an average of 320X coverage for each library.

Table 3.6: Sequencing run summary. Q20, one misaligned base per 100. Uniformity refers to the level of equal representation of the generated sequencing reads across the targeted regions.

Run	Patient ID	Bases	≥Q20	Reads	Mean read length (bp)	Mapped Reads	On Target (%)	Coverage (X)	Uniformity (%)
1	SABC002	220,762,925	183,104,648	1,737,315	127	1,729,584	96.54	992.8	94.17
	SABC005	153,146,388	126,887,134	1,160,384	132	1,156,617	95.31	677.6	90.61
	SABC009	201,632,278	167,600,423	1,546,314	130	1,543,211	96.43	904.7	94.02
2	SABC001	77,716,558	67,227,950	629,566	123	625,698	95.22	343.2	93.85
	SABC003	65,433,079	56,621,189	512,027	128	509,163	97.43	291.1	95.39
	SABC004	32,521,689	27,772,584	241,785	135	240,236	95.10	141.5	54.42
	SABC006	104,285,793	89,556,592	854,774	122	849,884	96.52	464.5	94.17
	SABC007	74,085,014	64,137,183	579,746	128	576,100	97.47	330.4	96.59
	SABC022	68,661,720	59,586,996	555,760	124	552,865	97.02	308.7	93.53
	SABC027	67,098,080	58,087,069	544,311	123	540,561	96.56	299.7	95.41
	SABC042	91,847,290	79,250,719	720,843	127	716,429	95.95	407.8	94.05
	SABC115	66,288,826	57,090,127	534,562	124	531,232	95.82	289.9	95.54
	SABC124	61,153,519	52,677,911	470,506	130	467,849	97.75	275.0	51.99

The majority of patient samples showed high quality reads with a high level of even coverage and uniformity. However, samples SABC004 and SABC124 had unusually low levels of uniformity (<60%). This can be explained for sample SABC004 due to the fact that pool 2 was identified as having a low concentration from the BioAnalyser analysis (**Table 3.4**), and this pool did not successfully sequence on the PGM. More surprising was the fact that SABC124 pool 2 failed to sequence despite having an acceptable size distribution and concentration (**Table 3.4**). Therefore, for these two samples only one pool was successfully sequenced, resulting in approximately 50% coverage of the target region. Despite this, these samples were still analysed for the regions that were sequenced once the bioinformatics pipeline was established.

3.3.4 Analysis of *BRCA1* and *BRCA2* variants to optimise bioinformatics pipeline.

Analysis of the Ion Torrent sequencing data was carried out with two bioinformatics programs, CLC Genomics Workbench and IonReporter. To identify the optimal pipeline for analysis of all genes included on the custom AmpliSeq panel, a bioinformatics pipeline was first developed using the *BRCA1* and *BRCA2* genes, which had already been sequenced by SA Pathology in all individuals. Due to design limitations of the AmpliSeq algorithm, the multiplex primer pool only covered 95.3% of *BRCA1* and 91.8% of *BRCA2* coding sequences, and therefore only regions covered by both Sanger sequencing and the Ion Torrent panel were compared in this initial pilot analysis.

3.3.4.1 IonReporter Analysis

After mapping the data to the human reference sequence, variant calling was carried out under germ-line High stringency and Low stringency parameters as outlined in **Table 3.1**. The IonReporter variant analysis pipelines identified a number of SNPs in addition to various indels in each of the samples analysed. The number of variants identified by both the high and low stringency pipelines within *BRCA1* and *BRCA2* are shown in comparison to the Sanger-identified variants in **Table 3.7**.

Table 3.7 illustrates that the majority of polymorphisms within *BRCA1* and *BRCA2* can be detected in both Sanger sequencing and MPS. While there are several variants that were false positives and negatives in the MPS data, these are due to limitations associated with the Ion Torrent sequencing chemistry and the design of the AmpliSeq panel (discussed in **Section 3.4.6**).

Table 3.7: *BRCA1* and *BRCA2* sequence variants detected through IonReporter analysis pipelines in comparison with Sanger sequencing data. Patient ID is indicated in the top row. SS, Sanger sequencing; L, Low Stringency variant analysis; H, High Stringency variant analysis. Blue, true variants detected by Sanger sequencing and MPS analysis method; Red, false positives in Ion Torrent sequencing; Green; false-negatives in Ion Torrent sequencing.

	SABC005			SABC001			SABC027			SABC006			SABC124			SABC115			SABC009			SABC002			SABC042			SABC004			SABC007			SABC022			SABC003		
Variant	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**	**SS**	**L**	**H**
	BRCA1																																					
c.-19-115T>C																																						
c.442-34C>T																																						
c.2077G>A																																						
c.2082C>T																																						
c.2170C>T																																						
c.2311T>C																																						
c.2612C>T																																						
c.2792A>G																																						
c.3113A>G																																						
c.3119G>A																																						
c.3548A>G																																						
c.4308T>C																																						
c.4837A>G																																						
	BRCA2																																					
c.-26G>A																																						
c.425+67A>C																																						
c.426-89T>C																																						
c.681+56C>T																																						
c.865A>C																																						
c.1365A>G																																						
c.1504A>C																																						
c.2229T>C																																						
c.2971A>G																																						
c.3396A>G																																						
c.3624G>A																																						
c.3807T>C																																						
c.5744C>T																																						
c.6841+78delAAT																																						
c.7242A>G																																						
c.8149G>T																																						
c.8851G>A																																						

There were 6 differences identified between Sanger sequencing and the 2 different MPS filtering parameters used in the IonReporter software. Interestingly, 4 of the 5 differences were detected with both High and Low Stringency parameters, suggesting these variants were not able to be filtered out by the analysis software regardless of the parameters used (**Table 3.7**). Importantly, all MPS variant calls that were discordant with Sanger sequencing were found to be false-positives or –negatives of the MPS analysis, and not due to variants being missed by the diagnostic Sanger sequencing analysis.

The single false-positive variant from the MPS analysis of *BRCA1* (c.4837A>G) was most likely due to its location within a homopolymer region of 5 G nucleotides. There was not a clear reason for the false-positive identified in *BRCA2* (c.8149G>T). For further confirmation, repeated Sanger sequencing of this DNA sample confirmed that this variant was not present in this individual and was a true false-positive of the MPS data. These two false positives were consistent across both filtering parameters. A third false-positive was detected in the MPS analysis of *BRCA2* (c.1504A>C) but only in the Low Stringency analysis. Visual inspection of this sequence found that it was present at the end of a sequencing read (2 bp before 3' termination) and as such was almost certainly an Ion Torrent sequencing artefact; repeated Sanger sequencing of the DNA sample confirmed this (results not shown).

There were 3 variants identified by Sanger sequencing that were not detected by the MPS analysis in *BRCA1* and *BRCA2*. The BRCA1:c.3548A>G variant was successfully identified in 7 individuals, however was not identified in SABC124. Analysis of data illustrated that this variant was not detected in this individual as one pool failed to sequence successfully. Additionally, there were two false-negative variants identified in *BRCA2* from the Ion Torrent sequencing. Visual inspection of the sequencing reads covering the first variant (865A>C) indicated that the variant was present within a homopolymer stretch of 4 A nucleotides, and as such was most likely filtered out as it was thought to be a homopolymer error rather than a true sequence variant. A second false-negative variant was not identified in *BRCA2* (c.3642G>A). Analysis of sequencing data identified that this variant was found within a stretch of 5 G nucleotides, as was present in approximately 40% of reads, and was also likely filtered out as it too was deemed a homopolymer error.

3.3.4.2 CLC genomics workbench analysis

CLC Genomics workbench variant identification was carried out through the in-built Probabilistic and Qualitative variant analysis functions. These functions allow numerous parameters associated with stringency to be varied.

3.3.4.2.1 CLC Probabilistic variant analysis

Initial variant analysis with default Probabilistic parameters identified 30 false-positives in *BRCA1* (including 4 false-positives being identified in more than 5 patient samples, **Table 3.8**) and 87 false-positives in *BRCA2* (including 8 false-positives being identified in more than 5 patient samples, **Table 3.9**). The majority of these variants were indels, a common sequencing error found in Ion Torrent data in and around homopolymer regions (Bragg *et al.*, 2013). In Analysis 2, homopolymer indels were filtered out, resulting in the number of false-positive variants significantly decreasing (1 false-positive in *BRCA1* and 3 false-positives in *BRCA2*). Whilst Analysis 2 greatly decreased the effect of indels on false-positive detection, this analysis resulted in an increase in false-negative calls, as it failed to detect 7 variants which were known to be present by Sanger sequencing (3 variants in *BRCA1* and 4 variants in *BRCA2*). By reducing variant probability to 85% in the final analysis pipeline, 5/7 false-negatives were removed. The final optimised Probabilistic variant analysis resulted in 1 false-positive in *BRCA1* (c.4387A>G); which was identified in all variant analyses and was due to the location of the variant within a homopolymer stretch. In *BRCA2*, the optimised Probabilistic analysis identified 1 false-positive (c.8149G>T), the reason for which is unclear, and 1 false-negative (c.3624G>A), which was incorrectly filtered out as a sequencing error within a homopolymer stretch.

Table 3.8: *BRCA1* sequence variants detected through the optimisation of CLC Probabilistic variant pipelines in comparison with Sanger sequencing data. Patient ID is indicated in the top row. SS, Sanger sequencing; D, Default parameters; A2, Analysis 2 parameters; O, Optimised Parameters (as outlined in **Table 3.2**) Blue, true variants detected by Sanger sequencing and MPS analysis method; Red, false positives in Ion Torrent sequencing; Green; false-negatives in Ion Torrent sequencing.

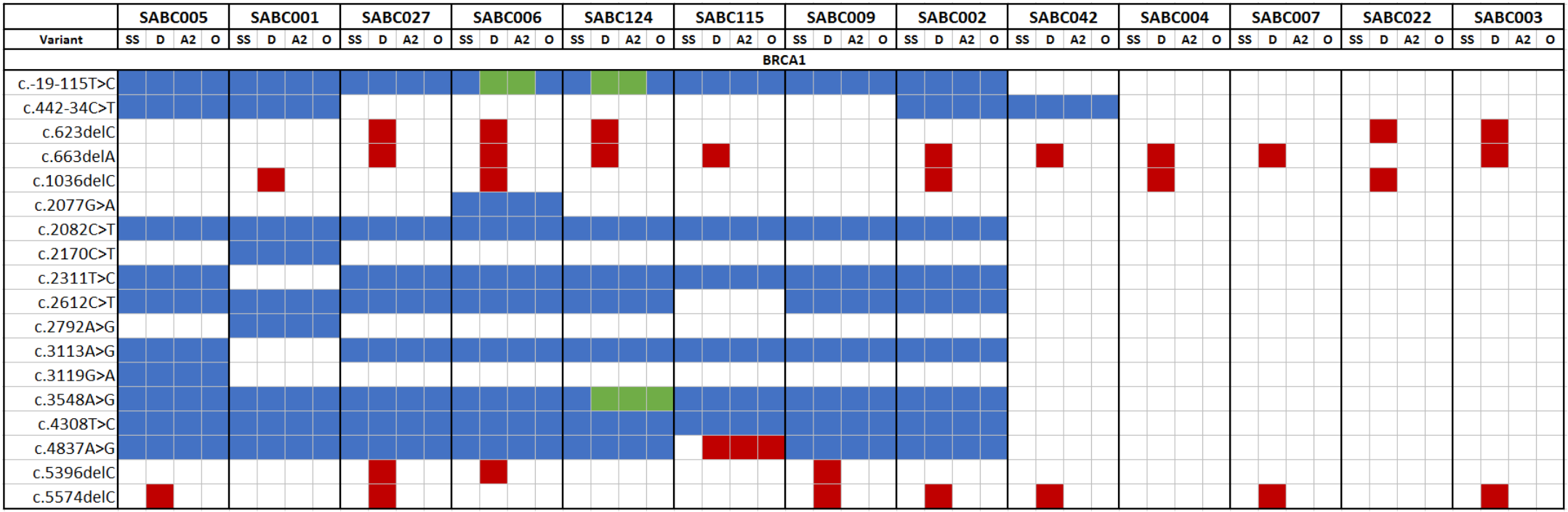
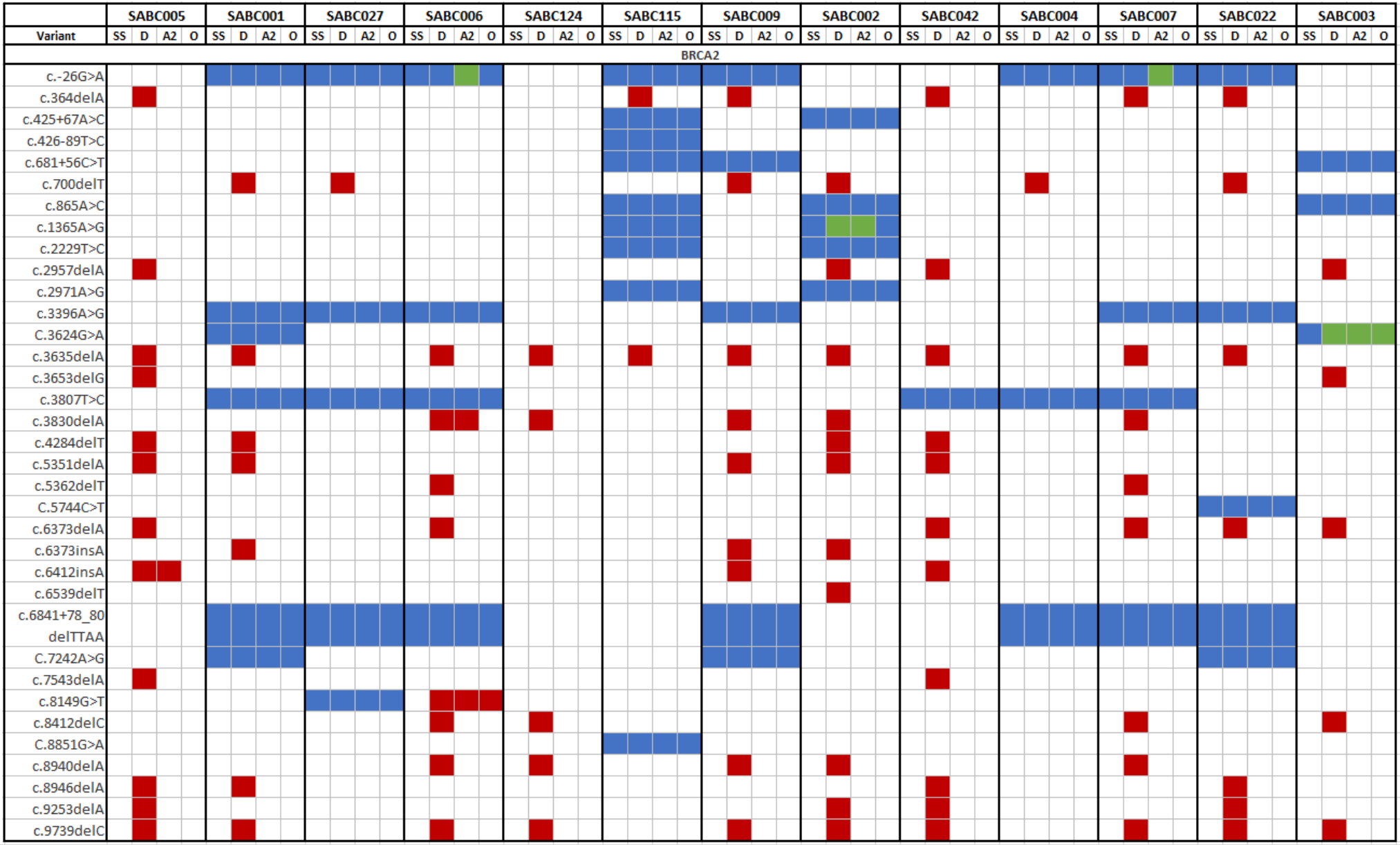


Table 3.9: *BRCA2* sequence variants detected through the optimisation of CLC Probabilistic variant pipelines in comparison to Sanger sequencing data. Patient ID is indicated in the top row. SS, Sanger sequencing; D, Default parameters; A2, Analysis 2 parameters; O, Optimised Parameters (as outlined in **Table 3.2**) Blue, true variants detected by Sanger sequencing and MPS analysis method; Red, false positives in Ion Torrent sequencing; Green; false-negatives in Ion Torrent sequencing.



3.3.4.2.2 CLC Qualitative variant analysis

The default parameters for Qualitative variant analysis (used in Analysis 1) have relatively low stringency; for example, the default settings do not require variants to be present in both the forward and reverse direction, nor are variants within homopolymer regions filtered out. This resulted in a significantly high number of false-positive variants being identified, consisting of 48 false-positives in *BRCA1* (including 5 false-positives being identified in more than 5 patient samples) and 145 false-positives in *BRCA2* (including 17 false-positives being identified in more than 5 patient samples). Again, the majority of these were indels, presumably due to intrinsic errors in the Ion Torrent sequencing chemistry.

In order to increase the stringency in Analysis 2, parameters were altered such that variants were:

1. required to be present in both sequencing directions;
2. removed if they were present in homopolymer regions; and
3. were required to meet an increase in multiple quality scores (maximum gap and mismatch count, minimum neighbourhood quality, minimum central quality).

By altering these sequence quality parameters, the increased stringency of variant detection significantly decreased the number of false positives, resulting in 2 false positives in *BRCA1* and 4 false positives in *BRCA2*. Visual inspection of the raw reads and coverage frequency of these variants found that these regions had low coverage, of approximately 10-15 reads only. Therefore, in Analysis 3, variants were required to have:

1. a minimum coverage of 50X; and
2. a minimum read proportion of 45%.

Whilst these parameters removed 1/2 false-positives in *BRCA1* and 3/4 false-positives in *BRCA2*, this analysis also had an increased number of false-negatives, with 6 in *BRCA1* and 7 in *BRCA2*. It was apparent that these variants were not detected due to the higher stringency for the read proportion of 45%. Upon altering this to the default setting of 35%, only 2 false-negatives remained in each of *BRCA1* and *BRCA2*.

Therefore, from the optimised parameters, there remained 6 variants that differed in comparison to the available Sanger sequencing. The single false-positive from the MPS analysis of *BRCA1* (c.4837A>G) was also detected in this individual in both high and low stringency IonReporter analysis

and was most likely due to its location within a homopolymer region of G_5 nucleotides. Additionally, the single false-positive in *BRCA2* (c.8149G>T) was also detected by both Ion Reporter analyses, and there was not a clear reason for the identification of this variant. Repeated Sanger sequencing of this sample was carried out to verify that the variant was not present, confirming it was a true false-positive within this sample. Two false-negatives were identified in *BRCA1*, with the c.-19-115T>C variant present in 30% and 33% of sequencing reads in these two individuals. This resulted in this variant being filtered out in 2 individuals as it fell below the specified read proportion metrics (35%) and was thought to be a sequencing error. The second false-negative (c.3548A>G) was not detected in one individual, as one pool of the library preparation failed to amplify. The two false-negatives within *BRCA2* (c.865A>C and c.3624G>A) were both due to the location of the variants within homopolymer stretches.

Table 3.10: *BRCA1* sequence variants detected through the optimisation of CLC Qualitative variant pipelines in comparison to Sanger sequencing data. Patient ID is indicated in the top row. SS, Sanger sequencing; D, Default parameters; A2, Analysis 2 parameters; A3, Analysis 3 parameters; O, Optimised Parameters (as outlined in **Table 3.3**) Blue, true variants detected by Sanger sequencing and MPS analysis method; Red, false positives in Ion Torrent sequencing; Green; false-negatives in Ion Torrent sequencing.

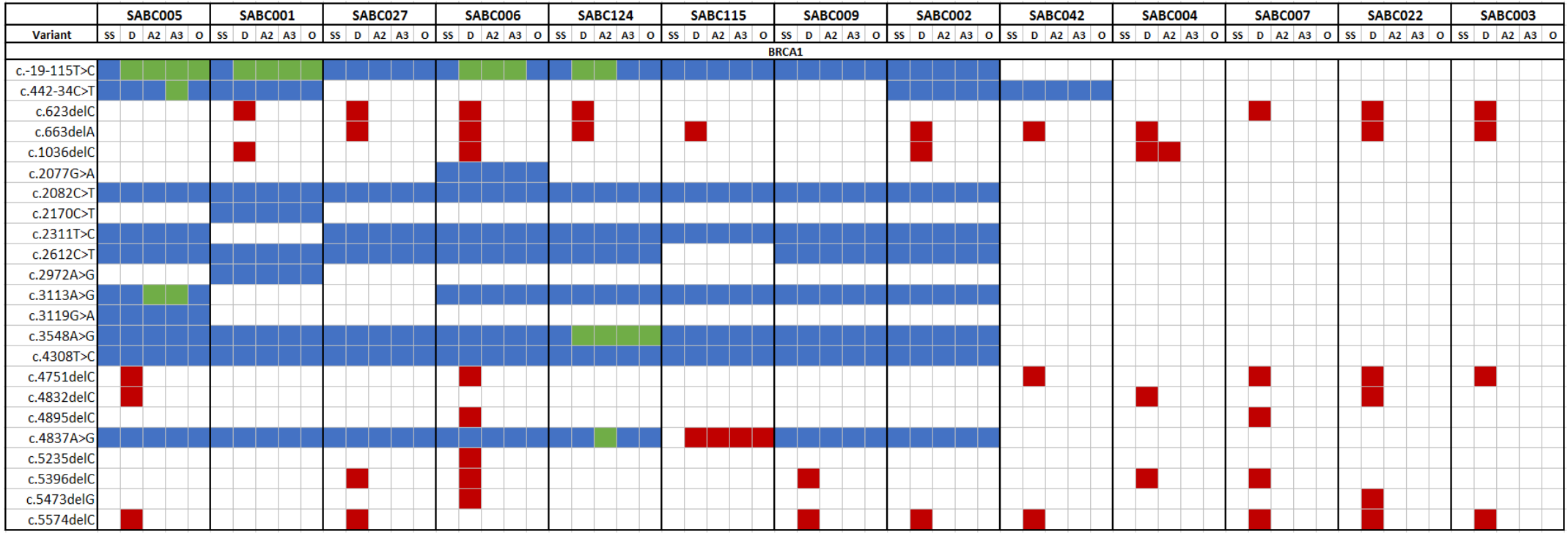
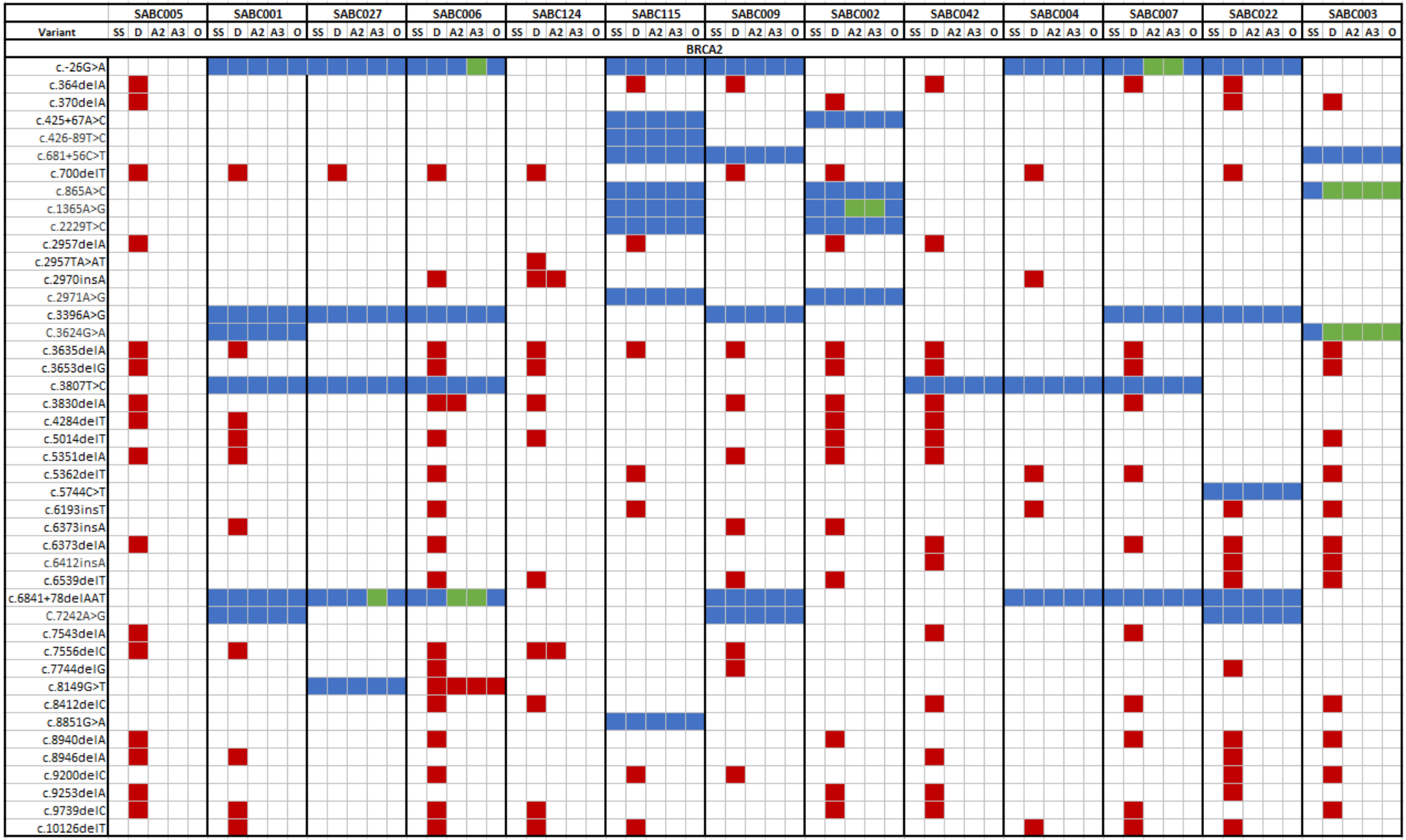


Table 3.11: *BRCA2* sequence variants detected through the optimisation of CLC Qualitative variant pipelines in comparison to Sanger sequencing data. Patient ID is indicated in the top row. SS, Sanger sequencing; D, Default parameters; A2, Analysis 2 parameters; A3, Analysis 3 parameters; O, Optimised Parameters (as outlined in **Table 3.3**) Blue, true variants detected by Sanger sequencing and MPS analysis method; Red, false positives in Ion Torrent sequencing; Green; false-negatives in Ion Torrent sequencing.



From the overall analysis using each of the optimised pipelines, it was determined that CLC Probabilistic and IonReporter High Stringency variant analysis would be used for all further analyses as these provided the most consistent results (summarised in **Table 3.12**). Despite taking longer for the initial optimisation, the probabilistic variant analysis feature of CLC genomics workbench was utilised in conjunction with IonReporter High Stringency Analysis. In support of this approach, previous studies have demonstrated that the Torrent Suite Variant Caller has been shown to result in a high level of false positives, with a reduced sensitivity (Quail *et al.*, 2012, Bragg *et al.*, 2013, Yeo *et al.*, 2014, Buzolin *et al.*, 2017).

Table 3.12: Summary of number of variants in *BRCA1* and *BRCA2* identified through each of the optimised MPS pipelines in comparison to Sanger sequencing data for 13 patients. Numbers shaded in red indicate a difference when compared to the number and/or location of variants identified though Sanger sequencing.

Patient ID	Sanger Sequencing	IonReporter		CLC Genomics Workbench	
		High Stringency	Low Stringency	Qualitative Variant Analysis	Probabilistic Variant Analysis
SABC001	15	15	15	14	15
SABC002	14	14	14	14	14
SABC003	3	2	2	1	2
SABC004	3	3	3	3	3
SABC005	10	10	10	10	10
SABC006	13	14	14	14	14
SABC007	4	4	4	4	4
SABC009	13	13	13	13	13
SABC022	5	5	5	5	5
SABC027	13	13	13	13	13
SABC042	2	2	2	2	2
SABC115	15	16	16	16	16
SABC124	8	8	7	7	7

3.4 Discussion

3.4.1 Library quantification

The Ion AmpliSeq protocol recommends quantifying libraries by one of three different methods: Qubit fluorimetry, BioAnalyser capillary electrophoresis and the Ion Library Quantitation qPCR Kit. The Qubit Fluorometer utilizes fluorescent dyes that specifically bind to DNA, and fluorescence is only emitted when bound to these target molecules. This makes this approach more sensitive than standard UV absorbance, which can be skewed by the presence of protein, free nucleotides or excess salts. In addition, this method is very fast and cost effective. The BioAnalyser is a chip-based capillary electrophoresis system, with the output consisting of a virtual gel image which provides information not only on concentration, but also the size distribution of the library (for example **Figure 3.2**). Unfortunately, both the Qubit and BioAnalyser methods have relatively low sensitivity, and therefore further PCR amplification of the library needed to be carried out prior to quantification. This requirement for PCR amplification of the library is not optimal because it may result in a preferential amplification of certain amplicons within each multiplexed pool. This can occur due to GC bias, which can be detrimental in the generation of MPS data, as some regions may have greater sequence coverage than others (Robin *et al.*, 2016). This could also skew allelic balance and result in false identification of variants.

The final and most sensitive method for determining library concentration is the Ion Library Quantitation qPCR Kit (Life Technologies). This approach uses qPCR analysis to determine the concentration of each library with reference to an *E. coli* standard. One main advantage of this method is that it does not require additional amplification of the libraries prior to quantitation. This approach is also more sensitive as it specifically detects only fragments which will be able to be sequenced in the library. This is because during the library preparation, two different adapters are added (A and P1), resulting in library fragments which contain either A-P1/P1-A, A-A or P1-P1. Fragments with A-P1/P1-A are the only library fragments which can be effectively sequenced on the Ion Torrent, and therefore represent the relevant part of the library to be quantified. The advantage of the qPCR quantification system is that it specifically quantifies only the amount of amplifiable/useable library fragments (i.e. correctly adapted A-P1/P1-A library fragments). As the BioAnalyser and Qubit instruments quantify all library molecules regardless of adaptors, these approaches are unable to discriminate incorrectly adapted fragments. This qPCR method was carried out for the first 3 libraries sequenced on the initial chip; however, this process is prohibitively

expensive, and for the large number of individuals selected for this study, unfortunately it was not financially viable. At the commencement of this study, the Ion Library TaqMan Quantitation kit was approximately $1600 for 250 reactions. Each sample is run in duplicate or triplicate, meaning at most, it is possible to quantify 125 samples from one kit. However, for each run, it is also necessary to include the 3 *E. coli* standards in duplicate, further reducing the number of samples that can be analyzed per kit. A cost comparison between the three quantification platforms found that the qPCR method was the most expensive, costing approximately $13 per sample, whereas the BioAnalyser/LabChip cost $5 per sample, and the Qubit DNA assay cost $1 per sample. Therefore, it was considerably more cost-effective to carry out both the BioAnalyser and Qubit assays in conjunction, rather than the qPCR-based approach.

Libraries were therefore quantitated with both Qubit and BioAnalyser approaches. It is clear from **Table 3.4** that concentrations determined using the BioAnalyser and Qubit were noticeably different, despite being the same sample. One reason for this is that the BioAnalyser estimates the average size of each library based on specific DNA intensities at certain sites. Therefore, complex multiplexed libraries with an excess of primers or any contaminants tend to distort the true average size and will generate inaccurate values (Robin *et al.*, 2016). In order to minimize the effects of these discrepancies, the size distribution of each library was visually inspected using the BioAnalyser analysis (**Figure 3.2**) to determine if effective library amplification had occurred, however libraries were pooled based on the Qubit derived concentrations. Furthermore, the molarity values determined by the BioAnalyser were required for downstream calculations before combining samples for sequencing, and therefore both quantification methods were used for each library preparation.

3.4.2 Multiplexing patient libraries across multiple sequencing chips.

The number of libraries combined within a single run is dependent on the size of the sequencing chip as well as the level of coverage required for the analysis. At the time of this investigation, it was reported that a minimum coverage of 100X was required for accurate germline variant detection (Chan *et al.*, 2012). Given that this targeted sequencing panel hadn't been previously used, it was initially decided to combine a relatively small number of libraries to determine the capabilities of the sequencing technology empirically. An average of 850X coverage was achieved from this initial sequencing (Table 3.6)This is a significantly higher level of coverage than required, therefore 10 patients were combined for the second run to determine the limits of the subsequent multiplexed

sequencing runs. This resulted in an average coverage of 320X for these 10 patients. This is still sufficient depth to accurately identify germ-line sequence variants within the selected patients. As this is still a greater than required level of coverage for these individuals, it is possible to combine a greater number of individuals on each sequencing run. From the results obtained, it is possible to combine three times as many AmpliSeq libraries (e.g. 30 libraries) on each sequencing run and still achieve the desired approximate 100X coverage for each individual sample. The limiting factor in terms of multiplexing libraries is the cost associated with the barcodes required for multiplexing. At the time of this study, it cost approximately $2500 for 16 barcodes. This cost associated with multiplexing large numbers of samples becomes limiting in terms of the number of individuals that can be included on one sequencing run. Intial attempts were made to multiplex 16 samples on the second MPS run, however, several libraries failed to amplify sucessfully and meet the required DNA yield metrics, and as a result, only 10 libraries were run on the subsequent chip.

3.4.3 MPS sequencing summary

As indicated in **Table 3.6**, sequencing of the generated libraries on the PGM was successful. Uniform sequencing coverage ensures that reads are distributed evenly across a targeted region and greatly helps with variation detection (Bodi *et al.*, 2013). This is an important variable as there are many biases associated with MPS sequencing, including, but not limited to, issues with preparation of AT-rich libraries, and sequencing of both GC-rich and homopolymer regions (Quail *et al.*, 2008, Quail *et al.*, 2012, Bragg *et al.*, 2013). For all individuals analysed in this pilot study, there was a high level of uniformity (> 90%, as recommended). This indicates that the two primer pools for each patient sample were multiplexed at equimolar concentrations. Only two samples SABC004 and SABC124 failed to meet these criteria (54.42 % and 51.99 % respectively), however this is attributed to the fact that one pool from each individual failed to produce sequence. In theory, this could have been predicted from the quantification data of these pools, however there were two additional pools which also flagged with low concentrations by both methods (SABC005 pool 1 and SABC009 pool 2) which went onto generate good sequencing data on the Ion Torrent. Therefore, concentration is not an accurate predictor of success of the library in downstream sequencing. If possible, it would have been useful to also obtain qPCR quantification for these libraries, to determine if this highly sensitive approach to quantification is able to accurately predict libraries which will not generate high quality sequence data.

3.4.4 IonReporter Analysis

Using IonReporter and both the high stringency and low stringency variant caller pipelines, many variants were identified for each of the 13 patients sequenced. Initially, variant calling parameters were compared between the inbuilt Germline High and Low Stringency pipelines within the IonReporter Software Suite. One advantage of IonReporter is that it has already been optimised for the plethora of small insertions and deletions present within Ion Torrent generated data (Rusmini *et al.*, 2016), resulting in a reduction in the false positive and negative error call rate compared to other analysis programs (as discussed in **Section 3.4.6**).

Several patterns of sequencing errors are recognized based on the sequencing chemistry employed by this MPS system. Ion Torrent data have a high ratio of false positives in the identification of small insertion and deletion mutations (Boland *et al.*, 2013, Zhang *et al.*, 2015a, Damiati *et al.*, 2016). However, this platform also demonstrates high accuracy in the identification of SNPs (Fujita *et al.*, 2017). The issue of false-positive mutations associated with Ion Torrent Panel sequencing raises a couple of areas of concern, regarding data analysis and the specificity of massively multiplex PCR reactions (McCall *et al.*, 2014). Given the large number of amplicons in this targeted gene panel, it is time consuming and ultimately impractical to manually curate each sequence. As such, automated software such as Ion Torrent's Variant Caller is essential to the future of this technology in a laboratory setting (McCall *et al.*, 2014). As a result, analysis was carried out between both pipelines to retain as many 'true' polymorphisms as possible for further downstream analysis.

The High Stringency settings are optimised for the identification of minimum false-positives using PGM chips present in a higher proportion of reads, whilst the Low Stringency settings are optimised for variants present in a high frequency of sequencing reads and minimal false-negative calls (Life Technologies, 2017). However, the High Stringency pipeline does have the potential to filter out variants of clinical significance due to its stringency.

Both stringency parameters were concordant with the *BRCA1* and *BRCA2* sequencing data obtained from SA Pathology, apart from 3 false-positive and 3 false-negative sequence variants. Two of 3 false-positive variants were identified by both the High and Low Stringency Parameters, while only 1 was identified by the Low Stringency Analysis but was filtered out by the High Stringency analysis. All 3 false-negative variants were missed in analyses with both the High Stringency and Low Stringency filtering parameters.

As previously mentioned, the false-positive BRCA2:c.1540A>C variant was located 2bp from the end of a sequencing read. The identification of this variant can be attributed to its location, as the accuracy of base calling is known to decrease near the end of reads. Additionally, this A is the last nucleotide within a 5 homopolymer stretch, which would have also contributed to the erroneous variant identification (Quail *et al.*, 2012, Bragg *et al.*, 2013) This variant was detected through the Low Stringency Analysis as it was present within enough of the reads (17/385 reads) to meet the pre-determined cut off of 5% of reads. However, this variant was excluded when this individual was re-analysed with the High Stringency Analysis, which required variants to be present in 20% of reads. It is important to note that this variant was only detected by the Ion Torrent Low Stringency Pipeline and was not found by CLC Genomics Workbench.

3.4.5 CLC genomics workbench analysis

Several studies have used programs other than IonReporter to analyse Ion Torrent data (Chan *et al.*, 2012, Vogel *et al.*, 2012, Yeo *et al.*, 2012, Rusmini *et al.*, 2016). Therefore, in addition to IonReporter, CLC Genomics Workbench was also used to identify variants. CLC Genomics Workbench is a program for the analysis of MPS data generated via all sequencing platforms, and as such it is not optimised for the sequencing errors associated with each type of sequencing chemistry. As a result, it took significantly longer to optimise the analysis parameters for these data (**Figure 3.1**).

The variant calling parameters needed to be optimised to maximise 'true' variant detection when compared to *BRCA1* and *BRCA2* Sanger sequencing data. CLC Genomics Workbench allowed the import of raw data generated from Ion PGM Sequencer. Only SNPs and multiple nucleotide variants (MNVs) that mapped to the target region were considered and compared to the pathology Sanger sequencing results. Initially, many deletions were detected for both the qualitative and probabilistic pipelines. This is consistent with the sequencing errors associated with Ion Torrent chemistry (Strom *et al.*, 2015). Additionally, the default parameters for both of these pipelines did not require variants to be present in both directions, nor were errors in homopolymer regions ignored. As these errors are known to be the main contributor of false-positives to MPS data generated with the Ion Torrent, it was necessary to filter these out (Loman *et al.*, 2012, Yeo *et al.*, 2012, Bragg *et al.*, 2013). As a result, the number of incorrectly called variants reduced significantly. From visual analysis of these deletions, it was evident that the majority were associated with the two most common sequencing errors known to affect Ion Torrent Sequencing chemistry:

1. Presence within a homopolymer run
2. Location at the 5' or 3' end of short sequencing reads.

These sequencing errors are known to severely affect the rate of false positives identified from PGM data (Loman *et al.*, 2012, Yeo *et al.*, 2012), and as such, need to be stringently filtered out.

When optimising the parameters for the Qualitative variant analysis, several variants were removed from analysis when altering the minimum read proportion from 35 % to 45 % (**Table 3.3**, Analysis 3). This could be attributed to potential skewing as a result of PCR amplification. From detailed analysis of variants that were excluded through this filtering analysis, it was determined that the minimum read proportion was too high, resulting in the elimination of a high number of true variants, and as a result this was reset back to the default (35 %) for further analysis. It is known that PCR amplification bias is a prevalent issue, particularly in GC rich regions and in repetitive regions. Furthermore, polymerase slippage occurs during amplification of polyA runs and AT dinucleotide repeats, often resulting in poor read quality and has the potential to result in an allelic imbalance (Aird *et al.*, 2011). Therefore, this slight skew in PCR amplification bias may be affecting the analysis of MPS data thorough the qualitative variant analysis method

3.4.6 Analysis of *BRCA1* and *BRCA2* sequences

Sequence alignments for variants with discordant results were manually inspected with the Integrative Genomics Viewer (IGV). Three putative variants (in 4 individuals) were observed from analysis of the raw MPS data, however subsequent Sanger sequencing failed to confirm the presence of these polymorphisms (**Table 3.7** – **3.11**). This highlights the potential inaccuracies of the variant caller software in determining the presence or absence of variants within MPS data.

3.4.6.1 False-negative variants

The *BRCA1* and *BRCA2* variants were compared to those documented for each individual by SA Pathology. In the regions which were covered by the AmpliSeq gene panel, all variants previously documented by *BRCA* screening were detected as outlined in **Table 3.7** to **Table 3.12**. Following the analysis of the validation cohort using the optimised Ion Reporter and CLC genomics workbench pipelines, 4 false-negatives were identified. Two of these were identified in all 4 bioinformatics pipelines used, and 2 were analysis specific.

The first false-negative variant was BRCA2:c.3548A>G, which was not detected in individual SABC124. This variant was missed by all analyses, due to issues with library amplification. This region was not successfully sequenced in SABC124, as multiplexed pool 2, which contained the primers for the amplification of this region produced a library which did not meet the library quantification metrics. Despite this, this library was still run on the sequencing chip, which failed to generate sequence. Additionally, variant BRCA2:c.3642G>A was not detected in individual SABC003. This variant had previously been identified and called by all analyses with IonReporter and CLC Genomics Workbench in SABC001. Further investigation of this variant identified that this polymorphism was present within a homopolymer stretch of 5 G nucleotides. Visual curation of the data indicated that the variant was present in 40% of the reads, however, was filtered out by IonReporter software as it was thought to be a homopolymer error. However, this variant was accurately called in individual SABC001, in which the variant was present in 53% of the reads. This skew in PCR amplification, coupled with the presence of the variant within a homopolymer stretch explains why this variant was missed within this sample.

Additionally, the polymorphism BRCA1:c.-19-115C>T was missed by the Qualitative variant analysis in SABC001, despite accurate detection in multiple other samples. Further analysis of this variant indicated its presence in only 33% and 30% of sequencing reads in the two individuals, indicating a skew in initial amplification of the DNA library in this individual's sample. Due to the filtering parameters associated with Qualitative variant analysis, this variant was considered to be a sequencing artefact and was filtered out. Furthermore, the BRCA2:c.865A>C variant was not identified in SABC003 through the Qualitative variant analysis. As previously mentioned, this variant was present within a stretch of 4 A nucleotides and was thought to be a homopolymer error.

3.4.6.2 False-positive variants

Overall analysis of variants detected resulted in the identification of two false-positive variants called by all programs, and one false-positive identified only by IonReporter Low Stringency analysis.

In individual SABC006, BRCA2:c.8149G>T was detected by all four variant calling pipelines. This variant was present in 50.8% of reads, and had a read depth of approximately 800X, however it was not identified on the initial Sanger sequencing records. As a result, this region was re-sequenced by Sanger sequencing, however this variant was still not detected in the patient sample. This variant was most likely associated with the incorporation of an incorrect base in the initial stages of the

generation of the library, as it was located within the middle of a sequencing read and was not surrounded by a repetitive run of nucleotides. The fact that it was present in 50.8% of reads implies that the error must have occurred very early in library generation.

BRCA1:c.4837A>G was identified in SABC115, again by all filtering parameters used. Further analysis of this variant identified that it was located within a stretch of G nucleotides and was a sequencing artefact. It was surprising that this error was not picked up by IonReporter as this software is manufactured specifically for the analysis of Ion Torrent generated data and had previously filtered out a true variant for the same reason.

The final false-positive variant, BRCA2:c.1504A>C in SABC124, was detected by Low Stringency analysis only. Visual inspection of this sequence found that it was present at the end of a sequencing read (2 bp before 3' termination) and as such was almost certainly an Ion Torrent sequencing artefact, as these are common issues associated with Ion Torrent sequencing and the variant was filtered out with increased stringency. The change in the required read proportion resulted in this variant being filtered out, with this variant only being present in in 17/385 of reads (5%), whilst High stringency analysis required the variant to be present in 20% of reads. Whilst this indicates that the high stringency analysis outperforms the low stringency analysis in this situation, the low stringency analysis may be beneficial for the analysis of somatic mutations or for the analysis of clonal mutations, which are present in a smaller proportion of sequencing reads.

3.4.7 Variants excluded from analysis

A number of additional intronic and exonic variants documented by SA Pathology were not confirmed using the AmpliSeq gene panel sequencing method. This is because these regions were not covered by the AmpliSeq primers, and thus were excluded from this analysis. The coverage of *BRCA1* and *BRCA2* achieved with this panel was 95% and 92% respectively. This lack of complete exon coverage is not acceptable for diagnostic purposes and is a limitation of this targeted MPS panel (**Appendix B**). Since the initial design and conception of this panel, Life Technologies has released the Ion AmpliSeq *BRCA1* and *BRCA2* panel, which contains analysis of the entire coding regions of these genes, plus 10-20 bp of flanking intronic sequences for analysis of splice sites.

With rapid improvements in the field over time, complete coverage of all desired genes will be easily attainable. However, another limitation of this custom AmpliSeq panel is the inability for

modifications (discussed further in **Chapter 7**). Since being designed in 2013, there have been numerous other genes associated with the development of breast cancer (as discussed in **Chapter 7**), however, the precise nature of the multiplexed primer pools unfortunately makes it impossible to expand the regions of interest.

In summary, a bioinformatics pipeline was generated for the analysis of Ion Torrent generated data. Through the comparison of MPS sequencing data of *BRCA1* and *BRCA2* to Sanger sequencing data, it was possible to establish an optimised, stringent pipeline in order to detect sequence variants in the remaining sequencing data generated for all patients.

Chapter 4:
Tri-Pool-Seq analysis

4.1 Introduction

Despite the reduction in the cost of massively parallel sequencing in recent years, sequencing many individual samples is still economically challenging. The most expensive aspect of massively parallel sequencing is the generation of the libraries themselves, with Ion Torrent libraries costing approximately $400 per individual sequenced. Recently, several studies have highlighted the utility of a pooling approach (Pool-seq), in order to maximise the number of individuals that can be sequenced from one library by significantly reducing costs (Anand *et al.*, 2016, Jin *et al.*, 2016, Ryu *et al.*, 2018). In addition to this, the Pool-seq approach has further benefits, including a reduction of DNA required from each individual and a reduction in overall workload.

Multiple studies have demonstrated that pooled-DNA sequencing is an efficient and cost-effective technique to identify rare variants in target regions (Calvo *et al.*, 2010, Diogo *et al.*, 2013, Jin *et al.*, 2016). Diogo *et al.* (2013) utilised the pooling approach to sequence 25 genes of interest in 500 individuals with rheumatoid arthritis and 650 case controls; this study identified 281 rare protein coding variants associated with this condition. A study carried out by Ryu *et al.* (2018) analysed a large cohort of Ashkenazi Jewish individuals (n=1000) through sequencing a 56 gene-panel on only 40 generated libraries, allowing a population-based analysis. Additionally, Anand *et al.* (2016) implemented the Pool-Seq approach to sequence 996 individuals in 83 pools and demonstrated that the pool-seq allele frequencies were robust and reliable through comparisons to public variant databases. These pooling approaches also work with varying levels of complexity, with sample numbers ranging from 12-50 patients per pool, albeit resulting in varying levels of coverage being achieved. These studies demonstrate the general utility of the pooling approach to increase the number of individuals that can be screened within a minimal number of sequencing reactions.

However, an unavoidable limitation of the predominant pooling methodology used thus far (Pool-Seq) is the loss of individual sample information. This experimental design results in the pools being multiplexed in a manner such that it is not possible to determine which individual the rare variant is present in. Therefore, this approach requires the downstream analysis of all individuals included within the pool to identify which specific individual the rare or causative variant is present in. This is a significant flaw in the experimental design and represents the main issue which is addressed within the Tri-Pool-Seq approach utilised within this study. This pooling strategy is designed in such a way that cross referencing of pools can allow for the identification of the specific individual with the variant. This approach was first detailed by Chi *et al.* (2014), which utilised this strategy for the

identification of rare mutations within sodium azide induced mutant rice populations. This study was successfully able to identify 16 mutations within specific mutagenized rice plants through the complex pooling methodology and subsequent deconvolution of sequencing data (Chi *et al.*, 2014).

Another disadvantage of all pooling strategies is that they generally result in lower sequencing coverage than sequencing samples individually. However, it has been previously demonstrated that germline mutations do not require deep coverage for identification, as they should be present in 50% of sequencing reads (assuming heterozygosity). Therefore, it may be a useful strategy to employ this approach for the analysis of an increased population size for the identification of rare and potentially pathogenic mutations in inherited breast cancer. This study represents the first application of the Tri-Pool-Seq methodology to human samples.

Twenty-five patients were included within each pool, resulting in the generation of a 5 x 5 x 5 cube. This approach allowed the sequencing of 125 individual samples through the generation of only 15 pooled libraries. As a proof-of-principle study, 8 pools were initially selected for sequencing, allowing for the analysis of 18 individuals across the multiplexed patient pools. In order to determine the specificity and sensitivity of this Tri-Pool-Seq method, each of the patient samples included in each pool were also sequenced individually.

4.1.1 Aims

The aims of this chapter are to carry out a proof of principle analysis to confirm the utility of the three-dimensional pooling method, including

1. Identification of rare variants within pooled patient samples and allocating the identified variants back to their respective patient samples
2. Comparison of rare variants identified through the pooling approach and individual sequencing approach for the identification of true variants and false negatives.

4.2 Methods

4.2.1 Patient Selection

Ethics approval was obtained from Southern Adelaide Clinical Human Research Ethics Committee (132.13). Patient consent for broad use of genomic material was obtained from all patients at the time of venepuncture. All individuals had been referred to the Familial Cancer Screening Unit in South Australia for *BRCA1* and *BRCA2* genetic testing (ranging from November 2009 – January 2013).

4.2.2 DNA integrity analysis

In order to minimise the potential for sample dropout within the pools, DNA integrity was determined via a quantitative real-time PCR assay (Brisco *et al.*, 2010). DNA integrity of the selected patient samples was analysed via a previously established quantitative real-time PCR (qPCR) assay within the department. The quality of DNA was analysed through the amplification of two different sized amplicons of the *GalT* (galactose-1-phosphate uridylyltransferase) gene (100 bp and 300 bp). Each of the samples selected were run in duplicate for each amplicon size. The two different fragment sizes were selected as it allows the analysis of the average and maximum expected sizes of products produced from the AmpliSeq panel.

In brief, qPCR reactions were set up at a final concentration of 1X PCR Buffer (Applied Biosystems, Australia) 5 mM MgCl (Applied Biosystems), 0.3 mM dNTPs (Fisher Biosciences) 0.4 mM forward primer (IDT), 0.4 mM reverse primers (short or long; IDT), FAM labelled *GalT* locked nucleic acid (LNA) Probe (IDT), using 1 unit Platinum Taq Polymerase (Life Technologies) plus 10 ng template in a final volume of 25 µL. Samples were analysed using a Bio-Rad iQ5 real-time PCR detection system (California, USA), with iQ5 optical system software and were determined to be of optimal integrity if the short and long fragments amplified within 0.5 C_qs of each other.

4.2.3 Pooling of patient samples

DNA from 125 individuals was pooled in equimolar concentrations (125 ng, as determined through Qubit fluorometry quantification) according to a three-dimensional cube strategy (**Figure 4.1**). Pooling was designed in a manner that if a variant was present in three pools, it would be possible to identify the individual sample carrying the variant.

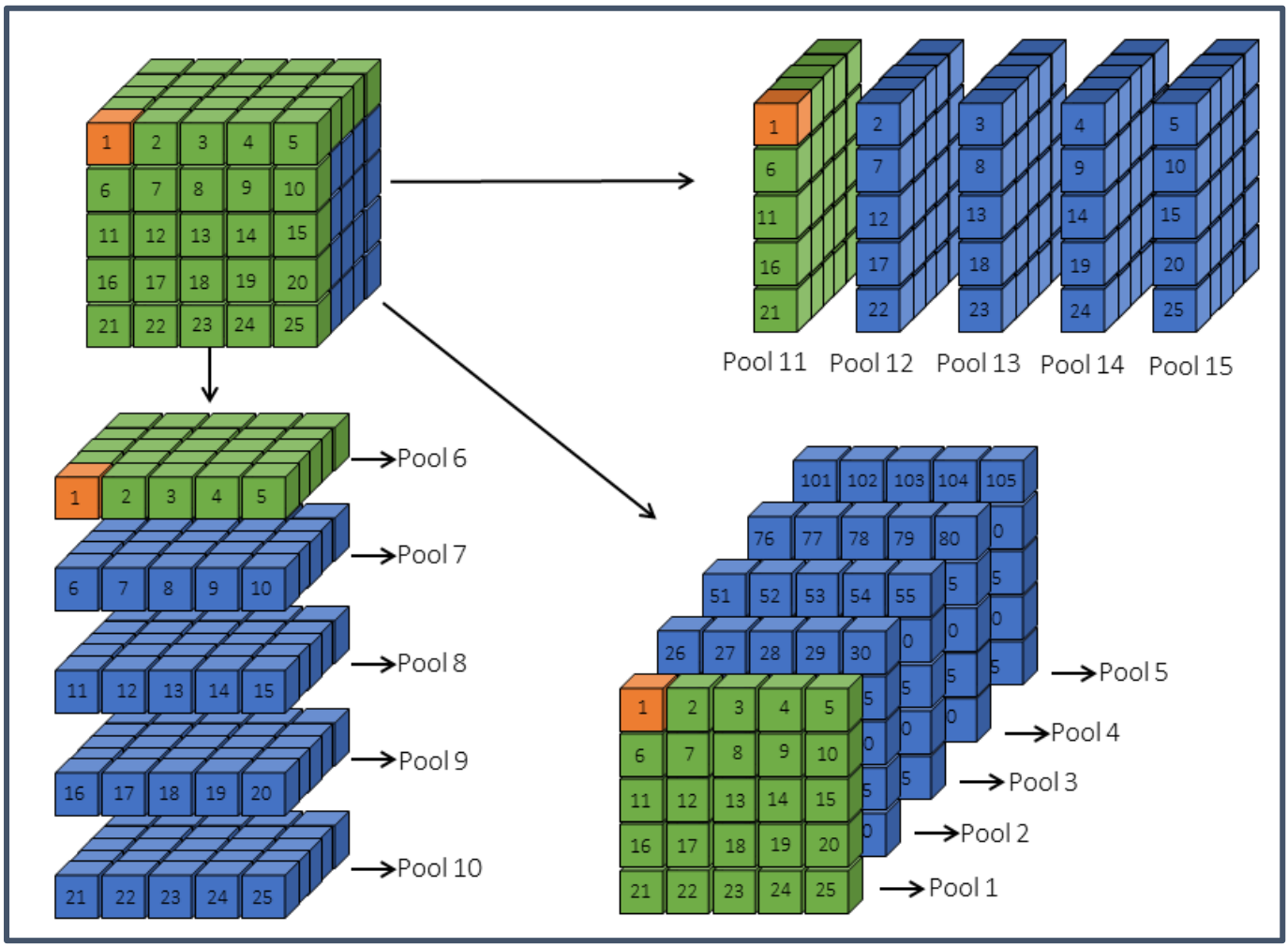


Figure 4.1: Schematic illustration of tri-pool-seq strategy. 125 patient samples were arranged in the form of 5 x 5 x 5 arrays, from which 15 DNA libraries would be created through the pooling of equimolar patient DNA. An example of the analysis is that a variant common to pools 1, 6 and 11 (green pools) would indicate the source as sample 1 (orange square).

4.2.4 Library preparations

Library preparation was carried out following the protocol 'Ion AmpliSeq™ DNA and RNA Library Preparation' Publication number MAN0006735, Revision B.0 (Life Technologies). Reagents used for the library preparation were provided in the Ion AmpliSeq Library Kit 2.0 (Life Technologies). The standard protocol was followed, with the following modifications under the guidance of the Life Technologies field application specialists:

- Forty nanograms of patient DNA was used as input;
- Initial amplification was cycled 15 times (protocol recommended 14 cycles) with 8-minute annealing/extension times (protocol recommended 4 minutes/cycle);

- Libraries of a low concentration (<300 ng/mL) were re-amplified (consisting of an additional 10 cycles of amplification, with subsequent 2X clean up with AMPure Beads); and
- 90% ethanol used for all clean up steps (protocol recommended 70% ethanol).

4.2.5 Library sequence analysis

Libraries generated from individual patient samples were sequenced on a 318v2 chip on an IonTorrent PGM at the Flinders University Genomic Facility. Libraries generated from pooled samples were sequenced either on a 318v2 chip on an IonTorrent PGM at the Flinders University Genomic Facility (pools 1, 6 and 11) or on an Ion P1 chip on the Ion Proton (Life Technologies) by the LotteryWest State Biomedical Facility Genomics at the University of Western Australia (pools 2, 3, 7, 8 and 15).

4.2.6 MPS data analysis

Data analysis was carried out by Bioinformatician Dr. Lesley-Ann Gray at the Australian Genome Research Facility (AGRF; Melbourne). Data analysis involved conversion of BAM files to FASTQ, with QC carried out on all FASTQ files, with variants of a Phred score less than 25 removed. Reads were trimmed, with removal of all barcode, adapter and amplicon sequences. Additionally, all homopolymers and sequencing artefacts were removed. Reads were aligned to the hg19 (February 2009 build, GrCh37). Variants were called using the GATK Haplotype Caller. Variants that were present in at least 1/50 reads were called and filtered through comparison to the provided BED files. Minimum allele frequencies of identified variants were annotated from gnomAD, with variants of low frequency (MAF<0.05) utilised for further analysis. Variants identified in the individual patient samples and the corresponding 3 pooled libraries, were compared.

4.3 Results

4.3.1 Generation of patient pools

In order to prevent skewing of patient representation in the pooling analysis, DNA integrity of the patient samples was analysed via quantitative real time PCR analysis. This analysis identified that all samples selected for the pooling approach were of sufficient quality as the short and long fragments amplified within 0.5 C_q of each other (**Table 4.1**). Additionally, samples appeared to be of similar quality as they all amplified within 2 C_q of each other. Following this, patient samples were pooled according to **Table 4.2** and DNA libraries were generated.

Table 4.1: DNA integrity analysis of samples included in all three pools for Tri-Pool-Seq method. DNA integrity analysed through qPCR amplification of a "short" (100 bp) and "long" (300 bp) product of *GalT*. Mean C_q for each amplicon of each individual included

Patient ID	Mean Cq	
	Short	Long
SABC007	29.88	30.32
SABC013	29.98	30.04
SABC025	29.98	30.23
SABC031	29.65	29.72
SABC042	30.06	30.07
SABC050	30.08	30.40
SABC059	29.93	30.12
SABC064	29.98	30.23
SABC065	29.94	30.05
SABC070	30.10	30.29
SABC071	30.19	30.23
SABC077	30.33	30.48
SABC085	29.44	29.87
SABC098	30.45	30.93
SABC102	29.16	30.03
SABC114	30.60	30.67
SABC127	29.99	30.03
SABC131	30.18	30.47

Table 4.2: Individual patient samples and the corresponding pools they are present in.

Individual Sample	Contained in Pools
SABC007	2, 8, 15
SABC013	3, 8, 15
SABC025	2, 7, 15
SABC031	1, 7, 11
SABC042	1, 5, 11
SABC050	1, 8, 11
SABC059	3, 6, 11
SABC064	2, 8, 11
SABC065	2, 7, 11
SABC070	3, 7, 15
SABC071	2, 6, 15
SABC077	3, 6, 15
SABC085	1, 7, 15
SABC098	1, 6, 15
SABC102	3, 7, 11
SABC114	3, 8, 11
SABC127	1, 8, 15
SABC131	2, 6, 11

4.3.2 MPS Run Summaries

Barcoded adapters were used to combine 3 multiplexed patient pools on one Ion318v2 chip (Run 1) and 5 multiplexed pools on an Ion P1 chip (Run 2) with overall run metrics shown in **Appendix D.**

The sequencing run summaries for each patient pool library are shown in **Table 4.3**. The majority of patient pools showed high quality reads, however patient pool 1 showed lower levels of uniformity and % on target in comparison to all other sequencing pools. This was particularly evident when the libraries were mapped back to the designed bed files, which identified that the off-target regions were located in pseudogenes or repeat regions. Interestingly, this was not observed when samples were individually sequenced (refer to **Appendix H** for all sequencing run data.)

A total of 21,605,556 unique reads were aligned to the reference sequence and achieved an average on-target percentage of 93.4 % in 8 pools (**Table 4.3**). Mean target coverage for pooled samples ranged from 647X to 2460X with an average of 1676X. As 25 individuals are included in each patient pool, rare heterozygous variants would be expected to be observed in 1/50, or 2 % of sequencing reads. Relative to the mean depth of 1676 reads, the expected depth of a variant present in one

individual in the heterozygous state would be approximately 30 reads, which should be sufficient depth for accurate variant detection.

Table 4.3: Sequencing run summary. Q20, one misaligned base per 100. Uniformity refers to even distribution of sequencing reads across the targeted regions. Run 1 consisted of 3 pooled samples sequenced on an Ion318v2 chip on the Ion Torrent PGM. Run 2 consisted of 5 pooled libraries sequenced on an Ion P1 chip on the Ion Proton.

Run	Pool ID	Bases	≥Q20	Reads	Mean read length (bp)	Mapped Reads	On Target (%)	Coverage (X)	Uniformity (%)
1	001	173,145,736	150,948,300	1,384,031	125	1,372,986	78.63	634.7	89.05
	006	188,094,811	165,937,149	1,524,757	123	1,513,269	91.36	802.6	93.36
	011	149,140,790	130,895,162	1,232,034	121	1,219,807	93.34	647.9	90.78
2	002	507,137,357	452,970,662	3,531,959	143	3,522,535	97.70	2,308	95.62
	003	483,298,041	430,975,578	3,359,785	143	3,349,851	98.0	2,207	95.59
	007	443,101,499	393,139,096	3,152,896	140	3,140,668	92.17	1,920	96.34
	008	530,028,391	472,117,289	3,701,781	143	3,691,485	98.16	2,428	95.15
	015	540,380,924	482,283,091	3,718,313	145	3,709,575	97.86	2,460	95.38

4.3.3 Bioinformatics analysis of pooled samples

Analysis of the tri-pool-seq method involved comparison of the variants identified within each pool to those identified within the patient samples contained with the corresponding pools. This was carried out for all 18 patients that were covered through the various combinations of the pooled patient libraries (**Table 4.4**). Comparison of these variants demonstrated that there were considerably more variants identified in the pools than was expected from the individual sequencing data (pooling false positives). Further analysis of these variants identified that the majority of these false positives were common within the general population (MAF >0.05) and would most likely be present within multiple individuals included within each pool (Refer to **Appendix E** for full analysis of pooling variants). Therefore, analysis of variants was limited to the rare variants, defined as an allele frequency of <0.05 as determined by the gnomAD database (including absence from the database). Rare variants detected in all pools (true variants) or missed completely or partially by the pooling approach (false negatives) are detailed in **Table 4.4.** Overall, these results indicated that a higher proportion of variants were identified through the individual sequencing analysis than were identified through the pooling approach.

Table 4.4: Variant analysis from three-dimensional pooling for analysis of 18 individuals. Total number of variants identified by pooling analysis and in individually sequenced patient included. Analysis then focussed on rare variants (MAF<0.05 as determined by gnomAD). The total number of rare variants within each sample indicated, with true variants (green) and false negatives (orange).

Patient ID	Pools	Total number			Rare variants			
		Variants identified by pooling method	Variants identified in all 3 pools	Variants identified in patient	Total number of rare variants in patient sample	Variants in all pools and patient	Variants in patient and 1 or 2 pools	Variant in patient only
SABC007	2, 8, 15	299	65	153	93	13	7	73
SABC013	3, 8, 15	253	86	117	52	26	6	20
SABC025	2, 7, 15	262	82	128	57	18	7	32
SABC031	1, 7, 11	315	82	137	74	25	16	33
SABC042	1, 5, 11	333	107	187	103	30	18	55
SABC050	1, 8, 11	283	82	129	52	14	18	20
SABC059	3, 6, 11	307	88	126	54	22	8	24
SABC064	2, 8, 11	277	95	140	49	20	11	18
SABC065	2, 7, 11	289	112	161	70	30	16	24
SABC070	3, 7, 15	260	83	120	50	19	8	23
SABC071	2, 6, 15	301	81	145	73	22	14	38
SABC077	3, 6, 15	300	97	130	55	25	11	19
SABC085	1, 7, 15	294	69	116	58	20	10	28
SABC098	1, 6, 15	301	90	140	71	25	22	24
SABC102	3, 7, 11	299	101	150	64	23	14	27
SABC114	3, 8, 11	280	84	124	51	19	13	19
SABC127	1, 8, 15	278	98	137	61	29	11	21
SABC131	2, 6, 11	300	92	140	60	16	22	22

Combined analysis of the pools indicated that many rare variants were missed through this pooling approach. In most samples, 30-45% of rare variants were detected in all three pools. Alarmingly, there were several individuals where less than 30% of the rare variants were detected through the pooling methodology (SABC007; 14%, SABC050; 27%, SABC131; 26%). The discrepancy between the expected and observed variants was further analysed. In-depth analysis of the missed variants identified that functionally relevant mutations were missed in some individuals. A pathogenic *PALB2* mutation (PALB2:c.3116delA) was missed in both SABC025 and SABC042, and a pathogenic *BRCA1* mutation (BRCA1:c.4869delT) was missed in individual SABC070.

4.4 Discussion

Multiple studies have utilised the Pool-Seq methodology for the identification of rare variants (Calvo *et al.*, 2010, Harakalova *et al.*, 2011, Diogo *et al.*, 2013, Anand *et al.*, 2016, Ryu *et al.*, 2018). This approach allows for a significant increase in the number of patients sequenced with a decrease in the associated costs and labour time (Ryu *et al.*, 2018). These studies have been invaluable in enabling the cost-effective identification of disease-associated SNPs in complex diseases, where very large numbers of affected individuals and control samples are required. However, the main limitation is that the multiplexed nature of these pools means that it is often difficult, if not impossible, to determine which individual the variant of interest was identified in without further downstream sequencing analysis. The three-dimensional pooling strategy utilised in this study was designed in order to mitigate this issue.

4.4.1 DNA integrity analysis

Prior to pooling, the integrity of the individual DNA samples selected was analysed through qPCR. DNA analysis was carried out to ensure all samples included in each pool were of optimal quality, in order to prevent over- or under-representation of each patient sample in the generated pool, prior to library preparation. Studies have shown that multiplexed PCRs can result in uneven and unspecific amplification of specific targets, which may be due to the different efficiencies in primer binding and extension (Chi *et al.*, 2014). Therefore, integrity analysis was carried out to determine if the DNA utilised for the patient pools was suitable. This would help prevent additional PCR bias, as sub-optimal samples may not be amplified with similar efficiencies, creating random biases (Marroni *et al.*, 2012). Additionally, it has been demonstrated that samples that are degraded or are of unverified quality require longer to amplify than those of optimal integrity (Brisco *et al.*, 2010). Therefore, through this analysis, it was possible to determine that the samples selected were of similar quality. This assay verified that all samples analysed through this method were of similar quality, as they reached the cycle threshold within two cycles of each other (**Table 4.1**).

This analysis was carried out to obtain equimolar representation of all 25 patient samples within each patient pool, preventing drop out of individual samples. Whilst this analysis indicated that samples were of similar quality, comparison between the variants identified through the pooling approach and the individually sequenced samples showed that there was a large proportion of rare variants within each sample that were missed (variants in patient only, **Table 4.4**). This may illustrate that there was drop out of some of the samples included within the patient pools. However, the

deconvoluted analysis identified rare variants specific to each sample, indicating that sequencing of all samples in the pilot study was successful (to some extent), with rare variants being detected for all 18 patients included in this initial analysis. It is possible that these variants may have been present within more than one individual within the pool and were incorrectly utilised to verify the successful amplification of patient DNA. Unfortunately, it was not possible to measure if equal representation of each sample within the pool was achieved through this analysis. However, analysis of sequencing results was utilised to determine if there was an over or under-representation of variants within samples (as discussed in **Section 4.4.3**).

Whilst qPCR was utilised for DNA integrity analysis in this instance, there are other methods that could be utilised to QC the DNA prior to pooling. Analysis could be carried out spectrophotometrically with the Nanodrop. The absorbance profile (260/280 nm ratio) generated through this method allows for the detection of contaminants such as salts, phenols, proteins and polysaccharides which are known to interfere with DNA sequencing (Abdel-Latif and Osman, 2017). This method allows for the quantification of DNA concentration, in addition to analysis of DNA purity. Additionally, other methods include analysis via electrophoresis, with DNA samples visually analysed for signs of degradation and contamination. However, QC through both the Nanodrop and electrophoresis does not illustrate how amplifiable the DNA samples obtained are. The extracted DNA may contain a high level of damage, from a variety of sources including extraction methods, repeated freeze thawing, contaminants within the sample etc. Therefore, a combination of both of these methods in addition to analysis through qPCR may be beneficial in future to determine both the purity of the extracted sample and the utility of the DNA for amplification in the future.

4.4.2 Sequencing of the generated patient pools

As indicated by **Table 4.3**, sequencing of the 8 patient pools on the Ion PGM and Ion Proton was successful. Pool 1 showed a lower level of uniformity and % on target reads in comparison to all other sequencing pools. Pool 1 was only 78% on target, whilst all other generated pools were >90% on target. When the sequenced pools were mapped back to the BED files, it was evident that the regions that were off target within this pool were located within pseudogenes and repeat regions. A somewhat underappreciated problem with MPS is the misalignment of short reads to a reference genome. When reads are mapped to the incorrect genomic location, or discarded if they are too divergent, this results in the generation of biased allele frequencies. Whilst these issues can easily be identified when sequencing individual DNA samples, this is harder to identify in pooled samples,

particularly for low frequency alleles (i.e. rare variants). This is due to the fact that the variation in coverage is often small, and therefore is difficult to detect (Schlotterer *et al.*, 2014).

In addition to this, the misalignment of sequences may be attributed to the background noise that is generated through the sequencing of 25 patients within a single pool. Every read generated represents an independent sequencing event from a large pool of chromosomes. Due to the high error rates of MPS sequencing (0.48 % – 1.78 % for Ion Torrent Sequencing (Quail *et al.*, 2012, Song *et al.*, 2017)), it is difficult to distinguish between sequencing errors and low frequency variants within the pooled libraries. Unlike the analysis of individually sequenced samples, this issue cannot be overcome through the analysis of multiple reads within the same region, as rare variants are only expected to be observed in a small number of reads (2-4% of reads). As such, genuine rare variants are often mistaken for background sequencing noise and filtered out. Particularly, this sequencing noise is commonly reported in Pool-Seq studies (Harakalova *et al.*, 2011, Schlotterer *et al.*, 2014, Anand *et al.*, 2016). Due to the slight sequence variation and background noise observed, it is possible that these 'noisy' sequences may have been incorrectly mapped back to the pseudogenes rather than the genes included on the panel.

Moreover, although DNA integrity was analysed prior to pool generation and library preparation, there may have been regions within pool 1 that were particularly difficult to amplify, resulting in a signal drop out and underrepresentation of these regions in the amplified and sequenced libraries, resulting in a lower proportion of on target reads in comparison to all other generated libraries These issues could have arisen at any step during the patient pooling, library preparation, and sequencing and as such cannot be corrected. However, this non-uniform representation of the desired regions within this pool may have resulted in a bias in the generated data and therefore affected the respective deconvolution of variants based on presence in pools. As such, it would be beneficial to re-generate the sequencing library for this particular pool and re-sequence it in an effort to achieve a more uniform coverage of the desired regions.

4.4.3 Variant identification through the pooling methodology

For the analysis of polymorphisms within the pooled libraries, it was anticipated that rare variants would be observed within 4% of sequencing reads (1/25 alleles) for a homozygous variant in one individual, and 2% of reads (1/50 alleles) for a heterozygous variant in one individual. Analysis of the proportion of reads that rare variants were detected in indicated that a majority (55%) of those that

were picked up by the pooling method were overrepresented within their populations, which could have either been attributed to a preferential amplification or over-representation of the specific patient sample within the pool. Alternatively, it may have been that the rare variant was present in more than one individual included within the generated patient pool. This is more likely to be the case, as analysis of the genuine variants that were known to be present in only one individual were not often detected through the pooling methodology at all (as discussed in **section 4.4.3.1**).

4.4.3.1 A large proportion of rare variants were missed within each patient sample

Previous pooling studies have demonstrated that a significant proportion of the identified variants are often rare and novel, with approximately 50-85% of variants identified through these studies falling into these categories (Harakalova *et al.*, 2011, Anand *et al.*, 2016, Ryu *et al.*, 2018). This study is in line with these findings, with 45-60% of variants identified from our Tri-Pool-Seq analysis being rare or novel. However, further analysis found that the pooling data identified a mean of 22 rare variants per individual, whilst the individual sequencing identified an average of 64 rare variants per individual. This represents a high number of false-negatives (>50%) from the Tri-Pool-Seq method. As previously mentioned, 3/8 sequencing pools generated had a significantly lower level of coverage (mean 693X coverage) than the other 5 sequenced pools (mean 2264X coverage). This decrease in coverage may have resulted in variants being missed as they were only present within a small number of reads and may have been incorrectly filtered out as sequencing errors. These pools with a lower sequencing coverage may have then affected the success of the deconvolution process, as it may have resulted in variants being missed, resulting in variants only being correctly identified within the pools with a higher read depth. As a result, this would have resulted in variants being identified within one or two pools only, in addition to the individual patient sequencing and therefore were filtered out through the deconvolution process of the specific individual. In order to combat this, it would be beneficial to ensure all pools sequenced have a similar level of coverage in future.

As mentioned, the major challenge in detection of rare mutations from large populations is to correctly distinguish genuine mutations from sequencing errors, as the latter confound with low frequency alleles within the sequenced pools (Harakalova *et al.*, 2011, Schlotterer *et al.*, 2014, Anand *et al.*, 2016). Although 5 of the sequencing pools generated in this study had a mean read depth of 2264X coverage, this was still insufficient for the detection of a significant proportion of rare variants within all sequencing pools. This is surprising, as there were only 25 samples included

within each pool, with each individual having a mean coverage of 90X within these 5 sequencing pools. At this level of coverage, it was anticipated that all variants, even those only present within one sample should have been detected. However, it was observed that the low frequency alleles in the generated pool (i.e. a heterozygous variant in 1 individual in a pool of 25 samples would be present in 1/50 alleles) could not reliably be distinguished from background noise, and as such may have been filtered out, or the background noise was incorrectly annotated as a sequence variant through this process. This may have resulted in the failure of this approach, as it is difficult to distinguish between sequencing noise and true sequence variants.

It may be possible to rectify this through an increase in sequencing coverage, allowing for a greater read depth in an attempt to eliminate the sequencing artefacts. However, it has been recommended by Schlotterer *et al.* (2014) that a minimum of 50X coverage of each individual within the pool, which was achieved for 5/8 pools in this study. Additionally, Ryu *et al.* (2018), achieved a mean coverage of 1068X for pools of 25 individuals and was able to identify rare variants through a pooled approach. According to the literature, the read depth achieved in this study was sufficient, however in this study it did not result in the identification of a majority of rare variants. This could be associated with the difficulty in differentiating between background noise and true sequencing changes and different programs designed specifically for the analysis of pooled sequencing data should be used. Additionally, replicated sequencing of the pools may be a useful way to differentiate between sequencing errors and genuine sequence changes. This may be a useful way to not only reduce the error rate associated with the Pool-Seq approach used, but also to determine if any drop out of individuals within each pool had occurred. Whilst this would have been a beneficial analysis to carry out, the cost and time associated with this was beyond the scope of this study.

As Pool-Seq approaches increase in popularity, so too does the generation of programs for the analysis of pooled sequencing data. These programs use known errors from the sequencing of genuine polymorphisms to analyse pooled sequencing data to improve SNP calling within data sets generated from pooled DNA samples (Li *et al.*, 2008, DePristo *et al.*, 2011, Schlotterer *et al.*, 2014). Such programs include CRISP (Bansal, 2010), Syzygy, (Calvo *et al.*, 2010) and VipR (Altmann *et al.*, 2011) one or a combination could be utilised in future. These programs are designed and utilised for the accurate calling of variants within pooled samples, based on different metrics. CRISP is designed to detect SNPs and short indels from DNA pools that have been subjected to high throughput sequencing. CRISP leverages sequence data from multiple generated pools to detect

rare and common sequence variants, but requires multiple pools for comparison in order to accurately determine genuine sequence changes (Bansal, 2010, Bansal *et al.*, 2011). Whilst both Syzygy and VipR are algorithms designed to detect SNPs and indels from pooled sequencing data based on quality scores and distribution scores, determined by the coverage achieved at any one position (Calvo *et al.*, 2010, Altmann *et al.*, 2011, Rivas *et al.*, 2011). Through the analysis of these various factors, including coverage, quality scores, and through comparisons to the other generated sequencing pools, these programs are more accurately able to distinguish between false positives, arising from sequencing errors compared to real variant alleles. The utilisation of these programs would have been beneficial as it may have been possible to pick up more rare variants within the pools, which would then be more representative of the rare variants identified in the individually sequenced patients.

4.4.3.2 False positive variants identified through Tri-Pool-Seq

Of the rare and novel variants detected within each pool, it was observed that approximately only 15% were rare and the remaining 85% were novel. This has also been reported in previous pooling studies (Harakalova *et al.*, 2011, Chi *et al.*, 2014) and has been speculated to be associated with the low filtering threshold that is required for the detection of single variant allele in a pool of alleles (50 alleles for the pools generated in this study), complicating the distinction of true sequencing variants from noise. This increase in false positives was observed in all samples, with approximately 30% of rare variants detected through the pooling approach being present in all 3 pools, but not the individual sequencing data. Ryu *et al.* (2018) reported a false positive rate of 6.3% within their Pool-Seq study and Anand *et al.* (2016) demonstrated a false positive rate of 53.9% and 6.7% both pre- and post- quality filtering. Both of these studies have demonstrated significantly lower false positive rates than what has been observed within this study.

It has been suggested that there is an overall high level of accuracy in the detection of known variants, which is mainly attributed to publicly available reference datasets and SNP-array data (Anand *et al.*, 2016). However, the same cannot be said for novel rare variants, which have not only been reported in multiple Pool-Seq approaches (Harakalova *et al.*, 2011, Anand *et al.*, 2016), but also observed within this study. False positive rare variants are one of the most challenging aspects of the pooling methodologies used, due to the high number of alleles present within the sample. The detection of low frequency alleles within these samples is often confounded with the detection of sequencing errors, generating many false positives. MPS technologies are not completely error

free, with Ion Torrent sequencers being prone to higher error rates than other sequencing platforms. When analysing individual diploid sequencing data, it is easier to detect and correct small sequencing errors, as the allele frequency, particularly in inherited conditions, can only be one of a few discrete values (those being not present, 0%; heterozygous, 50%; homozygous, 100%). However, with multiple diploid organisms pooled for sequencing, the possible allele frequencies can become many possible values, making it extremely difficult to detect or even correct for these deviations in observed allele frequencies that may be due to sequencing errors. As such, this makes it difficult, if not almost impossible to discriminate genuine rare variants from background noise generated from sequencing errors based on allele frequencies alone. As such, a more stringent, filtering approach is required to combat these high levels of false positives.

4.4.3.3 Pool size has been shown to affect the success of the pooling methodology

As Pool-Seq approaches have been designed to maximise the number of individuals that can be screened in a smaller number of sequencing reactions, it can be used to determine allele frequency estimates from the cohort being analysed. Importantly, determination of allele frequencies from small sized pools (containing <50 individuals) has been shown to yield suboptimal results for allele frequency estimates (Schlotterer *et al.*, 2014). However, a population-based approach was not the main goal of this analysis, and as such, a smaller pool size was utilised. Previous other Pool-Seq approaches have illustrated success, with smaller pool sizes which supported the pool size of 25 used in this study (Jin *et al.*, 2016, Ryu *et al.*, 2018). Additionally, as the primary objective of this pilot study was to determine if the Tri-Pool-Seq methodology was able to accurately detect rare variants and to determine if they could be traced back to the individual of origin, a smaller pool size results in less alleles within a pool to screen and was anticipated to result in more variants being detected than observed.

Whilst it has been recommended by Schlotterer *et al.* (2014) that pool sizes should be >40 individuals, Harakalova *et al.* (2011) reported that with a pool of 20 individuals the sequencing noise was too high, which resulted in a detrimental number of false positives. They recommended a decrease in pool size for future studies, hypothesising that a decrease in sample size would allow for the expected allele frequency to be increased, allowing for a more obvious distinction between the sequencing noise and a single heterozygote allele call. Multiple studies have since been carried out utilising pools of 10-25 individuals, which have successfully identified rare variants (Calvo *et al.*, 2010, Diogo *et al.*, 2013, Anand *et al.*, 2016, Ryu *et al.*, 2018). These studies have all utilised the

developed pooling analysis programs and have shown minimal detection of false positives and negatives. In future, it may be beneficial to decrease the size of the pools generated as it may increase the coverage per chromosome present within the pool. This increase in sequencing coverage could be utilised to more accurately detect genuine sequence changes from sequencing errors within the low frequency alleles.

4.4.3.4 Known pathogenic mutations not identified through Tri-Pool-Seq methodology

Due to the pooling strategy and the data analysis approach utilised within this study, a significant proportion of rare variants, some of which were functionally significant, were missed. A pathogenic *PALB2* mutation (PALB2:c.3116delA) was missed in both SABC025 and SABC042, and a pathogenic *BRCA1* mutation (BRCA1:c.4869delT) was missed in individual SABC070. These variants are known to be pathogenic mutations associated with the development of cancer, which have functional implications for the individuals and the families they are present in and were not detected through the Tri-Pool-Seq methodology. That these important variants were not identified provides strong evidence that the Tri-Pool-Seq approach is not a valid approach for the identification of rare and potentially causative variants within this cohort. The fact that these variants are single nucleotide deletions rather than single nucleotide base changes may have been a reason as to why they were missed in this approach. These variants were most likely incorrectly filtered out as sequencing errors, rather than being identified as genuine sequencing changes. As these variants were not identified within any of the pools they were contained within, it is probable that these variants were incorrectly filtered out in all sequencing pools, despite the high level of sequencing coverage achieved. Whilst this study is the first of its kind, there are several tweaks that need to be made in order to re-attempt the Tri-Pool-Seq approach for the identification of rare, potentially significant variants which can then be traced back to a specific individual.

Overall, the results of this preliminary analysis indicate that the pooling method was not an effective approach for the identification of rare variants within this patient cohort, as a significant proportion of rare variants were missed within each of the patients included within this analysis. The inability of this methodology to identify potentially disease causative mutations within these individuals highlights that this approach cannot be utilised for the accurate analysis of the individuals included within this patient cohort. Further work is required for the troubleshooting of this pooling methodology before proceeding to utilise this approach.

Chapter 5:

Identification of variants in *BRCA1/2* mutation-negative individuals with a familial history of breast cancer.

5.1 Introduction

Advances in sequencing technologies have resulted in genetic testing becoming common practise, particularly for inherited cancer risk evaluation (Easton *et al.*, 2015). Analysis of these individuals has played a vital role in the identification of disease associated loci and have been used to generate catalogues of genetic variation within both the diseased and general population. With an increase in the number of individuals being sequenced, the need for a clear interpretation of the pathogenicity of identified sequence variants has become apparent. This categorisation is carried out through a variety of mechanisms. Initial analyses are usually carried out through various *in silico* programs, including databases containing population and clinical variance information or prediction programs, focussing on protein structure and conservation.

For these resources to have the most impact, it is imperative that these databases and programs contain the right data to accurately identify disease-associated and causative variants from the broader spectrum of variants present within all human genomes (MacArthur *et al.*, 2014). The majority of variants that have been associated with genetic illnesses and phenotypes have not only been determined through *in silico* work, but also functional assays to determine the pathogenicity of the identified variants (Tavtigian *et al.*, 2008b). Issues in variant interpretation arise due to conflicting interpretations of pathogenicity, or when there have been changes to the assessment of variant pathogenicity (Lincoln *et al.*, 2015). This is particularly significant when variants have been included in public repositories without detailed functional analysis (MacArthur *et al.*, 2014). This is exceptionally evident in the ClinVar database, with 85% of entries being reported only once, with a minimum of 40% of these entries lacking functional evidence (Kobayashi *et al.*, 2017).

Although the vast majority of variants that are reported to be causative are indeed causative mutations, false assignment of causality is a significant issue. Detailed analyses of numerous studies have found that often, false-positive causative variants have been shown to be common polymorphisms within the general population, lacking direct evidence for pathogenicity and are often even identified in control populations used within the studies (Bell *et al.*, 2011, Norton *et al.*, 2012, Xue *et al.*, 2012). While it is recognised that functional analysis is a time-consuming process, this is a vital step in variant interpretation that needs to be carried out once variants have been subjected to rigorous *in silico* analyses. As the volume of patient sequencing data increases, it is critical that candidate variants are subject to thorough evaluation through the readily available

databases and through prediction software to prevent mis-annotation of the suspected pathogenicity. These analyses help to eliminate an array of benign and functionally insignificant variants, resulting in a more comprehensive list of variants for functional analysis. While the potential pathogenicity of identified sequence variants can be established through *in silico* analysis, pathogenicity should not be established through these programs alone. Functional analysis, particularly for novel variants, is imperative as false assignments of pathogenicity can have severe consequences for patients and families. This can result in incorrect prognostic, therapeutic or reproductive advice. This has the potential to lead to unnecessary treatment, familial cascade testing and prophylactic surgery. It is therefore imperative to ensure thorough analyses of predicted causative variants through both *in silico* and functional analyses are carried out before any clinically actionable variants are reported back to affected individuals and their respective families (MacArthur *et al.*, 2014).

For many genes, the most commonly identified sequence variants include known pathogenic mutations (often nonsense or frameshift mutations resulting in premature protein truncation), common neutral SNPs and VUS with an uncertain clinical risk. VUS are particularly problematic with regard to genetic counselling and treatment decisions. Their ambiguity requires further evidence (and therefore investigation) that the identified missense variants are actually pathogenic before they can be acted upon. Many of these VUS are either of uncertain clinical significance or have conflicting interpretations of pathogenicity, and whilst variants may be segregated in to high- or low-risk categories, there is a need to prioritise causative variants from the many candidates identified (Goldgar *et al.*, 2004, Higasa *et al.*, 2016). Classification of variants is established through several means. This includes epidemiological observations, comprising of family history and segregation of disease-associated alleles. It also includes direct variant frequency analyses and indirect measures such as amino acid conservation and quantification of the severity of amino acid change when present. This information is used in addition to evidence obtained through functional analyses (Goldgar *et al.*, 2004).

The overall evidence for the gene and/or variant in disease pathogenesis needs to be considered. This involves detailed analysis of all available data from both population frequency and clinical significance in addition to in-depth literature analysis (Doss *et al.*, 2014). While there are recommended strategies for variant analysis, universal guidelines for the prioritising of sequence

variants do not exist. Population frequency is a crucial criterion used for the clinical interpretation of sequence variants, with rarity being a prerequisite for pathogenicity (with the exception of founder mutations). Defining the threshold at which a variant is deemed too common is difficult, with laboratories often setting conservative allele frequency thresholds (Higasa *et al.*, 2016). An additional issue associated with allele frequency is the use of an ethnicity-specific reference genome for the population specific identification of rare sequence variants (Higasa *et al.*, 2016). Databases used to determine population frequencies within this study were comprised of summary data compiled from gnomAD, however it has been demonstrated that significant variation is observed in specific populations, and an Australian reference genome may be of benefit. Furthermore, the analysis of sequence conservation and effect of amino acid change is a useful tool in variant analysis as it can be applied to all missense variants and does not require extensive patient or familial history (Goldgar *et al.*, 2004). However, this may only be indirectly related to disease risk, and is a predictive tool, with further functional validation required for promising variants.

Classification of rare, non-truncating sequence variants is often problematic, as it is unclear if subtle changes within the screened genes alter function enough to play a role in cancer predisposition (Easton *et al.*, 2007). Interpreting the clinical significance of rare missense variants, particularly ones missing from public repositories, poses a challenge. These rare sequence variants are often non-pathogenic but are infrequent in the general population. Due to this there is added complexity in determining between the potentially pathogenic and private variants in these individuals. Despite this difficulty, there are a number of bioinformatics tools available to predict the effect of sequence variants on gene and protein function. According to The American College of Medical Genetics, these prediction tools can be utilised to prioritise missense variants for functional analysis, providing a supporting level of evidence for or against pathogenicity (Richards *et al.*, 2015). There is a myriad of *in silico* analysis programs, with each published study using a different combination of programs in parallel to determine the potential pathogenicity of identified sequence variants. These analysis programs utilise protein sequence, structural information or a combination of both, in addition to the biochemical properties of the amino acids to classify sequence variants as either pathogenic or neutral. Additionally, some programs consider the 3D structure of the protein, providing valuable information about sequence conservation, environmental changes upon mutation, stability and flexibility of the protein. As there is no “one size fits all” model for evaluation of potentially pathogenic sequence variants, it is necessary to ensure thorough and rigorous analysis is carried out

before proceeding with functional analysis of variants. This can be completed through the use of multiple *in silico* analysis programs and public data bases, but also emphasises the need to functionally validate the identified variants either *in vitro* or *in vivo.*

5.1.1 Identification of inherited breast cancer mutations

In recent years, diagnostic laboratories have transitioned from Sanger sequencing to MPS based panel approaches for identification of mutations within breast cancer susceptibility genes. This process allows for the simultaneous testing for mutations in the highly penetrant *BRCA1* and *BRCA2* genes, in addition to other moderate-risk genes. Genes included on these panels have varying levels of evidentiary support for their role in inherited cancer susceptibility. There are a range of multigene panels commercially available – most commonly ranging from 10 to 25 genes (Easton *et al.*, 2015, Judkins *et al.*, 2015, Tung *et al.*, 2015, O'Leary *et al.*, 2017, Rosenthal *et al.*, 2017) , which are increasingly being used in inherited breast cancer risk assessment. However, it has been demonstrated that an increase in the number of genes sequenced does not correlate to a significantly increased mutation detection rate (Lincoln *et al.*, 2015, Tung *et al.*, 2015, Maxwell *et al.*, 2016, Prapa *et al.*, 2017). Additionally, it is noted that panel sequencing should only be offered when indicated, once uninformative *BRCA1/2* testing has been carried out, and the individuals either have a familial history of breast cancer or a personal history of early onset breast cancer (Easton *et al.*, 2015, Prapa *et al.*, 2017). This approach aligns with the work that has been carried out within this study.

From the current commercially available inherited breast cancer gene panels, the most commonly interrogated genes are *BRIP1, CHEK2, ATM, PALB2, TP53* and *BARD1* (Winship and Southey, 2016, Prapa *et al.*, 2017), all of which are included on the diagnostic portion of this gene panel. It has been demonstrated that mutations within susceptibility genes other than *BRCA1/2* only accounts for an additional 10% of inherited breast cancer cases (Tung *et al.*, 2015). This illustrates that further investigation is required to determine the cause of the remaining 70% of inherited breast cancer cases.

In addition to sequencing a range of known breast cancer susceptibility genes, this study has also included 31 genes within the custom AmpliSeq panel that have the potential to be involved in breast cancer development. These genes function within 3 key cellular pathways and loss of function is

predicted to result in a similar phenotype to cancer attributed to a BRCA mutation. These pathways include DNA damage repair, G_2/M Cell cycle checkpoint control and homologous recombination. It is hypothesised that genes coding for proteins which directly interact with BRCA1/2 and or other proteins involved in DNA damage repair and checkpoint control may play a role in predisposing families to inherited breast cancer. This study illustrates a complementary approach to screen multiple genetic loci which have already been implicated in breast cancer predisposition, but also the potential to identify novel cancer susceptibility genes within this patient cohort

5.1.2 Aims and hypotheses

The aims of this chapter are to:

1. Use the custom AmpliSeq gene panel and the previously established bioinformatics pipeline on a cohort of individuals with inherited breast cancer for the interrogation of selected genes involved in DNA damage repair and cell cycle control.
2. Identify potentially pathogenic mutations within 19 genes known to be involved in breast cancer predisposition.
3. Identify potentially pathogenic mutations within 32 genes hypothesised to be involved in breast cancer predisposition.

5.2 Methods

5.2.1 Patient selection

Patient consent for broad use of genomic material was previously obtained for all patients at the time of venepuncture. Ethics approval was obtained from Southern Adelaide Clinical Human Research Ethics Committee (Application number: 132.13). Patients were selected based on two approaches. Initially, the Manchester scoring system was used to select patients for analysis (refer to **Chapter 2** for explanation of the Manchester Scoring System). Samples were selected from a cohort of individuals referred for genetic testing from June 2005 – June 2014. The majority of samples sequenced were from August 2011 – October 2012, ensuring a wide spread of Manchester scores was achieved, including 11 BRCA1/2 mutation-positive individuals.

5.2.2 AmpliSeq library preparation and sequencing

Library preparation was carried out following the protocol 'Ion AmpliSeq™ DNA and RNA Library Preparation' Publication number MAN0006735, Revision B.0 (Life Technologies). Reagents used for the library preparation were provided in the Ion AmpliSeq Library Kit 2.0 (Life Technologies). Library preparations were carried out following the flow diagram in **Figure 5.1**, with modifications from method indicated below the figure.

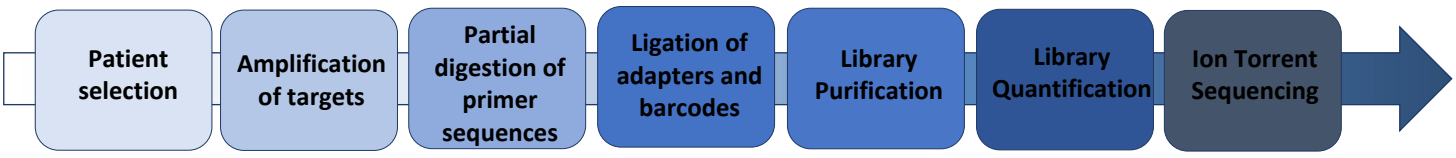


Figure 5.1: Flow diagram for MPS library preparation

The standard protocol was followed, with the following modifications under the guidance of the Life Technologies field application specialists:

- Half reaction volumes were carried out to maximise the number of patients that could be analysed from each kit;
- Ten nanograms of patient DNA was used as input;
- Initial amplification was cycled 15 times (protocol recommended 14 cycles) with 8-minute annealing/extension times/cycle (protocol recommended 4 minutes/cycle);

- Libraries of a low concentration (<300 ng/mL) were re-amplified (consisting of an additional 10 cycles of amplification, with subsequent 2X clean up with AMPure Beads); and
- 90% ethanol was used for all clean up steps (protocol recommended 70% ethanol).

Samples were sequenced either on the Ion 318 Chipv2 on the Ion Torrent Personal Genome Machine (PGM; Life Technologies) by the Flinders Genomics Facility (Flinders University, South Australia), or on the Ion P1 chip on the Ion Proton (Life Technologies) by the LotteryWest State Biomedical Facility Genomics at the University of Western Australia.

5.2.3 Bioinformatics analysis

Bioinformatics analysis was carried out using the optimised pipelines discussed in Chapter 3. Subsequently, all variants were filtered based on their prevalence within the general population. Variants that were of low frequency (<5% MAF) were retained and subjected to further analysis (**Figure 5.2**). Several established databases were used for the analysis of variants, to determine the potential functional significance of the identified variants (missense, nonsense, splice site and variants within the 3' and 5'UTR) within each individual sample.

5.2.3.1 Variant database analysis

As illustrated in **Figure 5.2**, five databases were used to determine the functional significance of all variants identified within the patient cohort. Databases included dbSNP and gnomAD, which provide information on variant prevalence within the general population, and COSMIC, HGMD and ClinVar, which provide information on clinical significance.

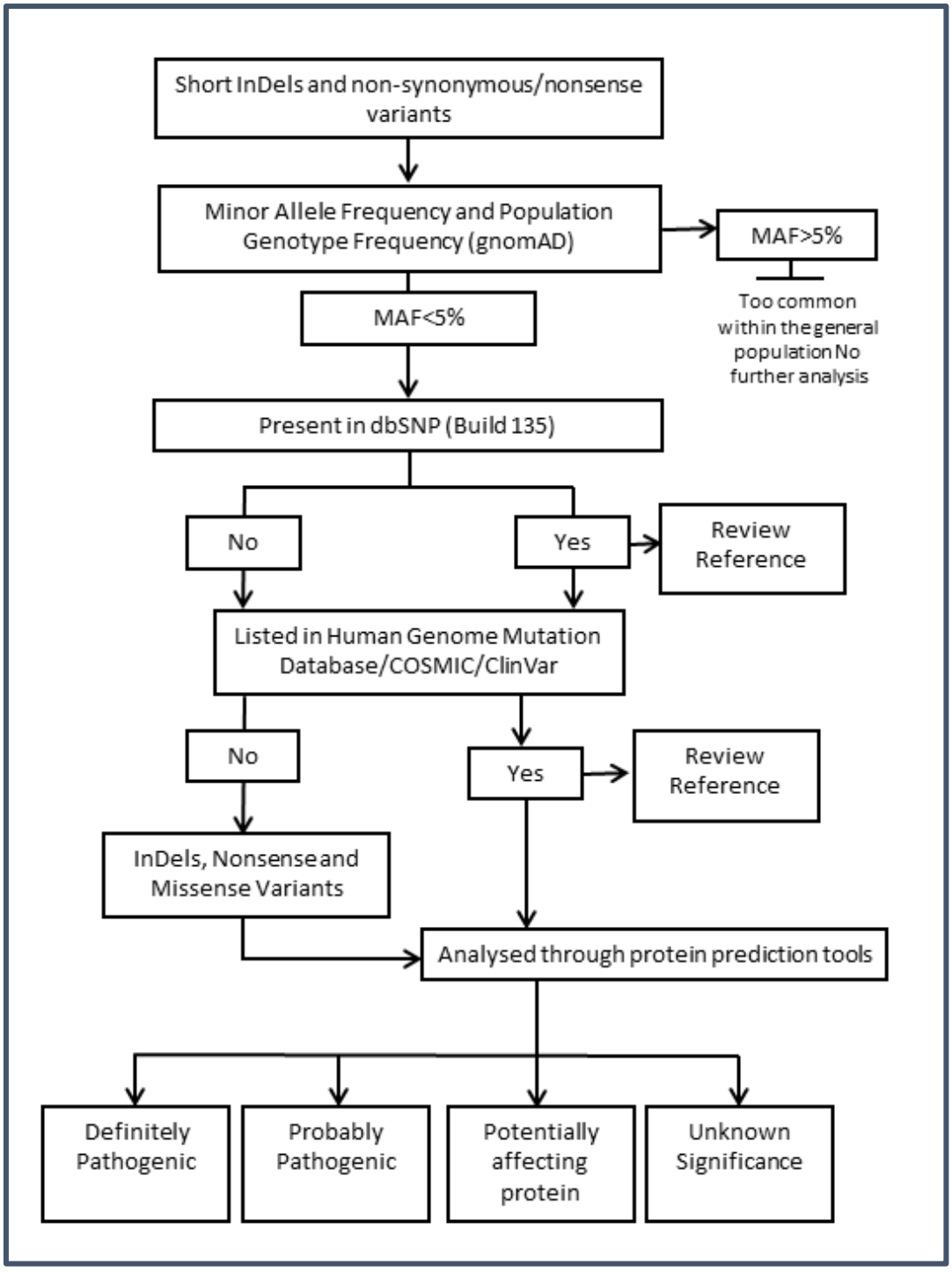


Figure 5.2: Workflow showing process of filtering variants to identify those of potential pathogenicity (Adapted from McCarthy *et al.* (2013). Indels, Insertions or Deletions; gnomAD Genome Aggregation Database; MAF, Minimum Allele Frequency. #Prediction tools include *in silico* analysis of variants using various protein prediction programs SIFT, PROVEAN, PolyPhen-2 and Align-GVGD in addition to analysis of protein functional domains.

5.2.3.2 Population frequency databases

1. dbSNP database (build135)

This database contains frequency information for variants identified from a variety of genomic studies, including the 1000 genomes project and exome sequencing projects (ESP; http://www.ncbi.nlm.nih.gov/SNP/).

Variants with MAF <5% were considered for further analysis.

2. gnomAD database

The genome aggregation database (gnomAD) contains the aggregation and analysis of both exome and whole genome sequencing data. It consists of whole exome datasets from 125,748 individuals and 15,708 whole genome data sets (Karczewski *et al.*, 2019) (https://gnomad.broadinstitute.org/).

Total MAF provided in gnomAD is used for the analysis of individuals in this patient cohort.

Variants with MAF <5% were considered for further analysis.

5.2.3.3 Clinical significance databases

3. Catalogue of somatic mutations in cancer database (COSMIC)

This database is a catalogue of somatically acquired mutations identified in human cancers. It is comprised of mutations identified from the analysis of approximately 4800 genes and 250000 tumours, resulting in the identification of 50000 mutations associated with somatic cancer (Forbes *et al.*, 2008); http://cancer.sanger.ac.uk/cosmic).

If variants were found in database, associated references were reviewed before its inclusion for further analysis.

4. Human gene mutation database (HGMD)

This database is a comprehensive collection of missense and nonsense mutations, regulatory and splicing variants, insertions and deletions, repeat expansions and gross gene lesions within 3600

different nuclear genes associated with human inherited disease. There are over 96000 different germline mutations and disease associated polymorphisms within this database (Stenson *et al.*, 2009); http://www.hgmd.cf.ac.uk/ac/index.php).

If variants were found in database, associated references were reviewed before its inclusion for further analysis.

5. Clinical variance database (ClinVar)

This database is a public archive of relationships between medically important sequence variation and phenotypes. Each submission includes the reported variation, interpretations of the variant to human health and evidence supporting each submission (Landrum *et al.*, 2014); https://www.ncbi.nlm.nih.gov/clinvar/).

Review any references associated with variant – variants listed as having uncertain significance, conflicting interpretations of pathogenicity or variants not present in database were analysed further.

5.2.3.4 *in silico* analysis to determine pathogenicity of variants

Variants retained following population and clinical significance database analysis were further analysed by *in silico* analysis to predict the effect of the change on the resultant protein. To be considered for further analysis, variants had to be predicted to effect protein function from 3 of the 4 analysis programs used. These effects on protein function included predictions of the variant being pathogenic, damaging, possibly damaging, deleterious, or lying within a functional domain or key region of protein function.

5.2.3.4.1 PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping) predicts the impact of an amino acid substitution on structure and function of proteins based on Bayes posterior probability. The *Homo sapiens* sequence alignment was imported into the PolyPhen-2 server (v2.2.2 release 2011_2012; http://genetics.bwh.harvard.edu/pph2/) and the functional consequences of these selected variants were analysed. This prediction tool generates a number of scores for each variant, including an overall score, and a sensitivity and specificity (Adzhubei *et al.*, 2013). The PolyPhen-2 program

also generates multiple sequence alignment of 75 amino acids from approximately 100 different species to examine sequence conservation.

Variants with scores >0.8 were classed as possibly or probably damaging.

5.2.3.4.2 SIFT and PROVEAN

SIFT (Sorting Intolerant from Tolerant; http://sift.jcvi.org/) analyses sequence homology and the physical properties associated with amino acid substitutions to determine if the amino acid changes are tolerated. This analysis determines the median sequence conservation which measures the diversity of the sequences selected for prediction and generates a score (Sim *et al.*, 2012, Hu and Ng, 2013). PROVEAN (Protein Variation Effect Analyser; http://provean.jcvi.org/index.php) predicts whether a protein sequence variant affects protein function (Choi *et al.*, 2012). This analysis generates a score based on the analysis of the top 30 closely related sequences and measures the change in sequence similarity of the query sequence before and after the introduction of an amino acid variation.

SIFT analysis of SNPs: Variants with scores <0.05 were predicted to be damaging.
SIFT analysis of indels: Variants with scores ≥0.5 were predicted to be damaging
PROVEAN: Variants with scores <-2.5 were predicted to be pathogenic.

5.2.3.4.3 Align-GVGD analysis

Align GV-GD analysis is based on multiple sequence alignments of 9 species from the Homologene feature of NCBI (http://www.ncbi.nlm.nih.gov/homologene/). Align-GVGD combines the biophysical characteristics of amino acids and multiple sequence alignments of proteins and analyses the level of sequence conservation between species. These values are used to determine a score for both Grantham Variation (GV), which is the biochemical variation at each position, and a Grantham Differentiation (GD) score, which identifies the difference between the biochemical properties of the variant position and the amino acid being assessed. These two scores are used to determine a class, ranging from c0 to c65, corresponding to the amino acid substitution being neutral to likely deleterious respectively (Tavtigian *et al.*, 2006).

Variants ranked as Class C15 – C65 were predicted to be likely to highly likely pathogenic.

5.2.3.4.4 Splice site analysis

The predicted effect of the identified variants on splicing were analysed using the Human Splicing Finder (http://www.umd.be/HSF/index.html). This tool was utilised to determine if any of the identified variants affected existing splice sites, in addition to the identification of any splicing silencers and splicing enhancers that may be affected by the variants (Desmet *et al.*, 2009).

5.2.3.4.5 Protein domain analysis

The protein feature of NCBI database (http://www.ncbi.nlm.nih.gov/protein) was utilised to assess the protein domains within each gene to determine whether any of the variants lay within any important functional domains. Additionally, the protein database, UniProtKB (http://www.uniprot.org/) was used to determine if any of the variants lay within any structural or functional domains of significance or if the protein variant had been shown to affect protein-protein interactions in previous studies.

Determined if variant was present within functional domain or key region of protein function.

5.2.4 Statistical analyses

The statistical significance of the identified variants was carried out with the assistance of Mrs. Mary Barnes (Flinders Centre for Epidemiology and Biostatistics, Flinders University). A one-sided z-test was used to compare the frequency of the identified variants with the observed frequency in GnomAD. For this statistical analysis, the null hypothesis was that the rate of mutation was the same as the reference population, and the alternative hypothesis was that the rate of mutation is greater in those with breast cancer than the general population. A p-value of less than 0.05 was considered significant for all statistical calculations

5.2.5 Confirmation of variants of interest

All potentially pathogenic variants were confirmed though Sanger sequencing. Refer to **Chapter 2** for detailed methods, **Appendix F** for primer sequences and **Appendix G** for cycling conditions.

5.3 Results

5.3.1 Manchester scores of individuals included in study

For this study, 133 individuals with a wide range of Manchester scores were selected. As illustrated in **Figure 5.3**, the Manchester scores ranged from 5 – 53, but clustered around the range of 10-25. Eleven individuals with known *BRCA1/2* mutations were also included in this analysis as part of the longitudinal study over a 12-month period.

5.3.2 Variant identification in patient cohort

Each patient sample was individually sequenced using the custom AmpliSeq gene panel on the Ion Torrent PGM or IonProton (Refer to **Appendix D** for MPS run summaries and coverage metrics). Of the 133 samples initially selected, sequencing data was successfully generated for 131 samples.

A cumulative total of 16,347 sequence variants were identified (consisting of 1,041 different variants) with a mean of 123 variants identified in each individual (range 65 – 168; **Appendix H**). An average of 28 low frequency variants (<5% MAF) were detected in each individual (range 14 -47; **Appendix H**)

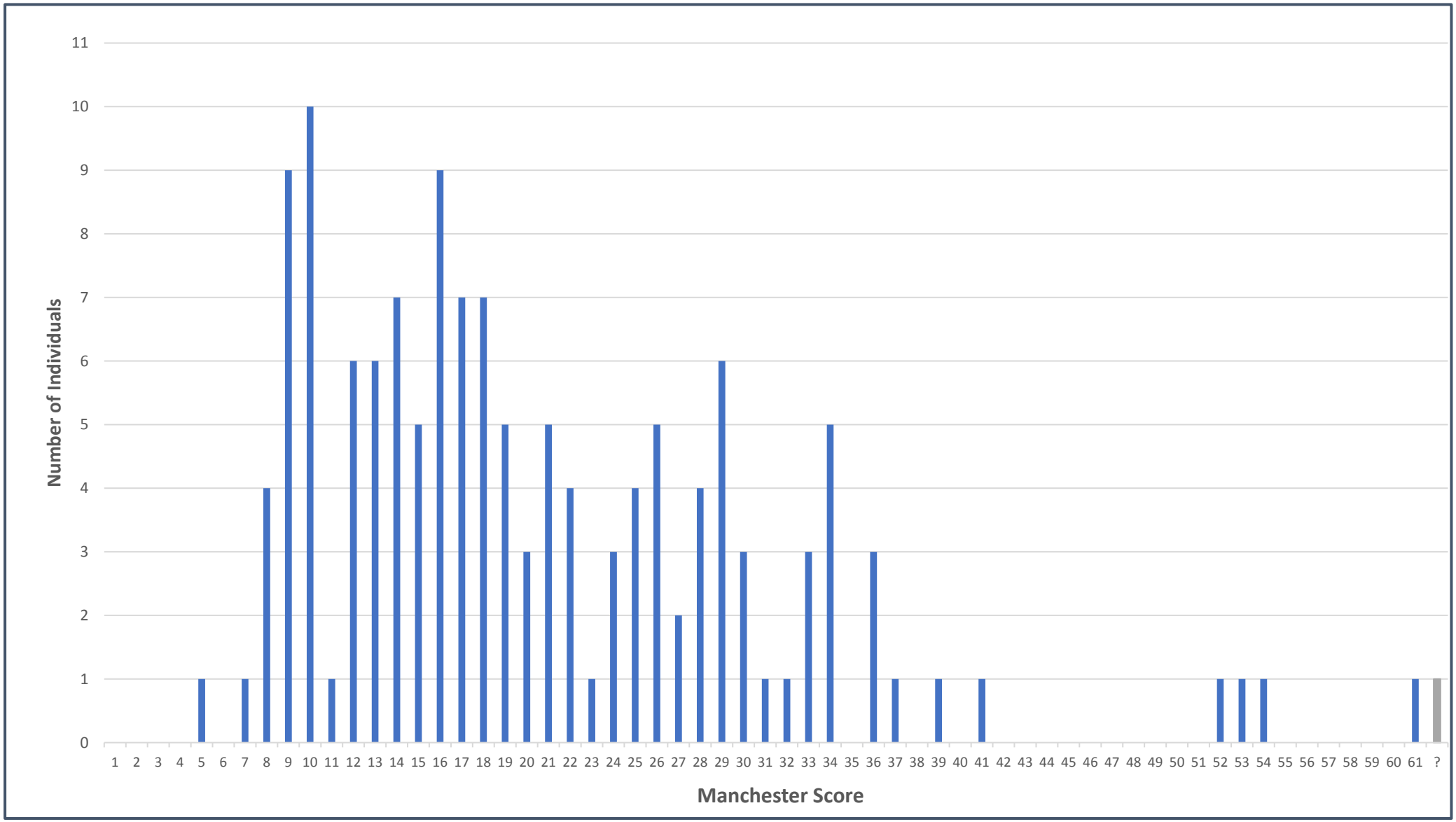


Figure 5.3: Manchester scores of individuals with hereditary breast or ovarian cancer included in this study (n=133). The Manchester score for one individual is unknown, as indicated by the final grey shaded column.

5.3.3 Analysis of sequence variants

From the initial 1041 variants identified in the patient cohort, 403 were eliminated based on their relatively high frequency in the general population (MAF > 5%), leaving a total of 638 low frequency variants. Of these variants, 119 variants were identified as intronic variants that were not predicted to affect splicing and were not analysed further. Furthermore, 25 low frequency variants were identified as synonymous mutations. These variants were subjected to splice site analysis and were eliminated as they were not predicted to have any effect on splicing (as depicted in **Figure 5.4**).

The remaining 494 variants were analysed through COSMIC, HGMD, UniProt and ClinVar databases to analyse the potential significance of each detected variant. This resulted in the elimination of 328 variants, which were either identified as benign or were eliminated based on data from the clinical significance databases and in-depth literature analysis as they had been shown to not play a role in cancer development. The remaining 166 variants were then subjected to *in silico* analysis with various protein prediction programs (including PolyPhen-2, SIFT Align-GVGD and PROVEAN). This resulted in a reduced list of 82 potentially pathogenic variants identified within 84 individuals. The database analysis for the 82 variants of interest is summarised in **Table 5.1**, with the *in silico* analysis results summarised in **Table 5.2**. The statistical significance of the identified variants is summarised in **Table 5.3.** The frequency of the potentially pathogenic variants found within this study was compared to the allele frequencies on gnomAD. All statistically significant variants are indicated in red.

Of these 82 potentially pathogenic variants, one was previously reported pathogenic in the literature, 44 variants were identified in genes contained within the diagnostic portion of the panel, whilst 38 variants were identified in genes contained within the discovery portion (**Figure 5.4**). These potentially pathogenic variants of interest were identified in 84 of the 120 *BRCA1/2* mutation negative individuals, with 36 individuals having no potentially pathogenic mutations identified in the 51 genes analysed. Several individuals had multiple potentially pathogenic variants identified within multiple genes. The number of identified variants and their frequency within the patient cohort is depicted in **Figure 5.5**.

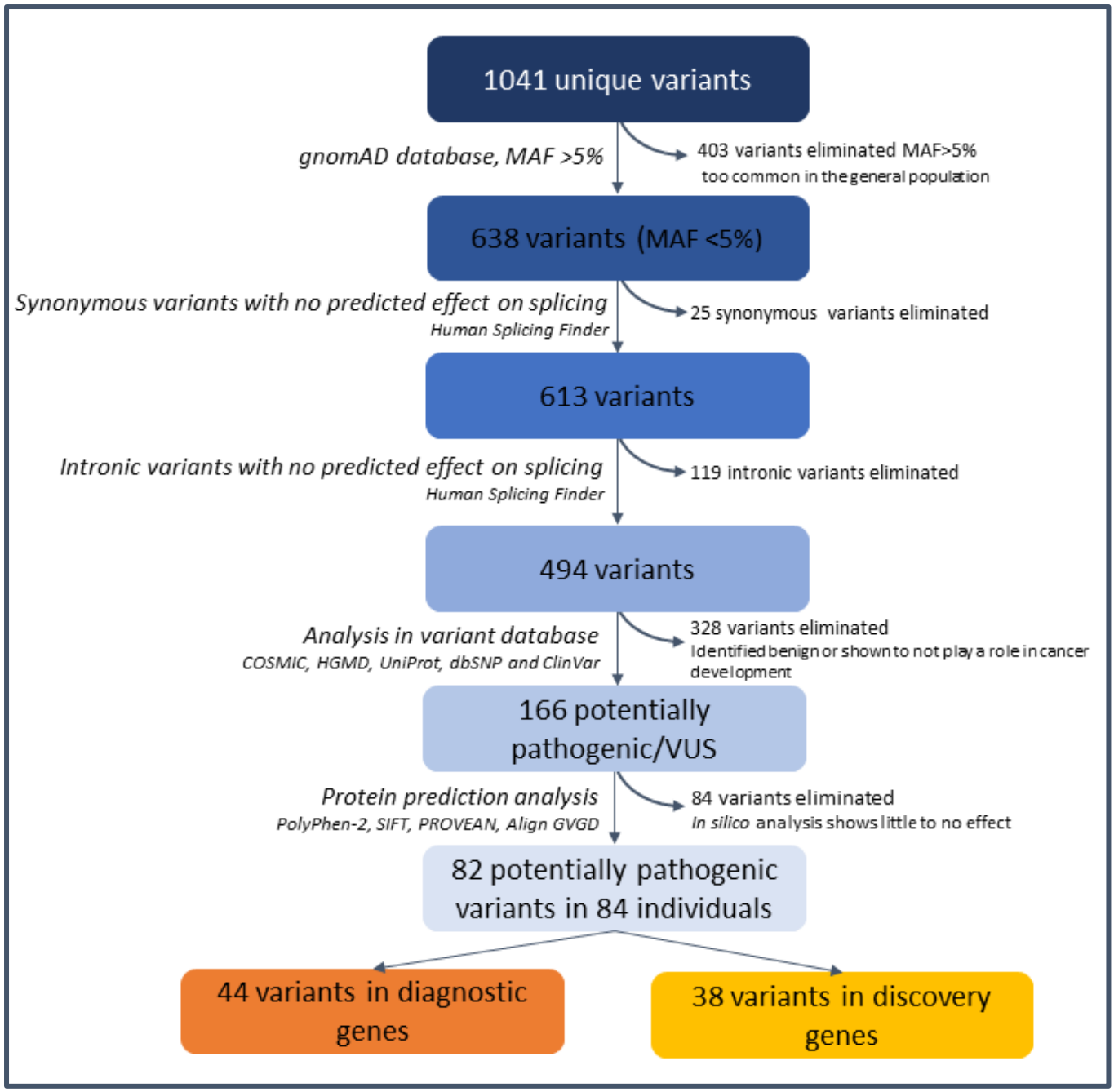


Figure 5.4: Filtering and analysis of variants found within the patient cohort for the identification of potentially pathogenic variants for further analysis. Results of filtering analysis carried out as illustrated in **Figure 5.2**. VUS, Variant of uncertain significance.

Table 5.1: Database analysis of predicted pathogenic variants. Transcript variants listed in HGVS nomenclature. MAF; Minimum allele frequency as determined by gnomAD(%). Presence in COSMIC, HGMD and gnomAD annotated. Variant location within any domain or region of the protein annotated from UniProt. Significance of variant in human disease as annotated in ClinVar and any references or other phenotypes associated with selected variant annotated from any databases used. HGMD, Human Gene Mutation Database, gnomAD; Genome Aggregation Database [#]Indicates the number of *in silico* analysis programs that predict the identified variant as pathogenic (As outlined in **Table 5.2**) Literature cited pertains to specific identified variant.

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by *in Silico*[#]	References and additional notes
ATF1	c.571C>G	P191A	9	2.376	N	CM067641	No domain	Y	Not Present	3/4	Susceptibility to lung cancer (Rudd *et al.*, 2006)
ATM	c.2T>C	M1T	1	<0.001	N	CM960095	No domain	Y	Path/Likely Pathogenic	3/4	Start lost, associated with Ataxia Telangiectasia (Gilad *et al.*, 1996)
ATM	c.998C>T	S333F	1	0.134	COSM5020963	N	No domain	Y	Benign/Likely Benign	3/4	Somatic mutation resulting in Haemangioblastoma (Shankar *et al.*, 2014)
ATM	c.1010G>A	R337H	2	0.006	COSM21301	CM0910483	No domain	Y	Conflicting interpretations of pathogenicity (likely benign/ uncertain significance)	2/4	Somatic mutation in breast cancer, intestinal adenocarcinoma (Zehir *et al.*, 2017) and oesophageal squamous cell carcinoma (Hao *et al.*, 2016), Susceptibility to inherited breast cancer (Tavtigian *et al.*, 2009)
ATM	c.1892C>T	P631L	1	0.000	N	N	No domain	Y	Uncertain Significance	1/4	
ATM	c.2119T>C	S707P	3	0.774	COSM41595	CM013692	No domain	Y	Likely Benign	0/4	Predicted predisposition mutation in somatic/inherited breast cancer (Dork *et al.*, 2001, Bozhanov *et al.*, 2010, Fletcher *et al.*, 2010) and Acute Myeloid Leukaemia (Hirsch *et al.*, 2016)
ATM	c.2572T>C	F858L	3	0.845	COSM21826	CM061641	No domain	Y	Conflicting interpretations of pathogenicity (likely benign/uncertain significance)	2/4	Rare polymorphism, predicted pathogenic in multiple haematopoietic malignancies, none confirmed somatic (Fang *et al.*, 2003, Gumy-Pause *et al.*, 2006, Kanagal-Shamanna *et al.*, 2014). Associated with increased radiosensitivity and development of inherited breast cancer, often linked with ATMc.3161C>G (Gutierrez-Enriquez *et al.*, 2004, Fletcher *et al.*, 2010)

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by *in Silico*[#]	References and additional notes
ATM	c.3161C>G	P1054R	7	1.658	COSM5932709	CM973365	No domain	Y	Benign	4/4	Confirmed somatic basal cell carcinoma (Sharpe *et al.*, 2015). Haplotype with ATM:c.3161C>G, increased radiosensitivity and development of inherited breast cancer (Larson *et al.*, 1997, Gutierrez-Enriquez *et al.*, 2004, Fletcher *et al.*, 2010) .
ATM	c.4258C>T	L1420F	1	1.114	COSM6495411	CM000653	No domain	Y	Conflicting interpretations of pathogenicity (likely benign/ uncertain significance)	1/4	Associated with somatic AML (Hirsch *et al.*, 2016) and Melanoma (Zehir *et al.*, 2017). Associated with increased breast cancer susceptibility (Fletcher *et al.*, 2010)
ATM	c.5558A>T	D1853V	1	0.443	COSM21628	CM083593	No domain	Y	Benign/Likely Benign	4/4	Association with bilateral breast cancer in conjunction with ATM:c.38-8T>C (Heikkinen *et al.*, 2005)
ATM	c.7390T>C	C2464R	1	0.037	COSM758329	CM016183	FAT Domain	Y	Conflicting interpretations of pathogenicity (likely benign/ uncertain significance)	2/4	Somatic mutation associated with B-CLL and intestinal adenocarcinoma (Kovaleva *et al.*, 2016). Potential breast cancer susceptibility (Dork *et al.*, 2001)
ATM	c.8305_8306insC	W2769fs	1	0.000	N	N	PI3K/PI4K Domain	N	Not Present	2/3	8307G>A (p.W2769X) is a null mutation associated with Ataxia Telangiectasia development (Gilad *et al.*, 1996). Predicted to cause nonsense mediated decay (NMD) of transcript.
BARD1	c.1670G>C	C557S	5	1.435	N	CM021950	Flexible linker domain, required for initiation of apoptosis	Y	Benign/Risk Factor	1/4	Increased prevalence observed in hereditary breast cancer. Associated with breast cancer predisposion in *BRCA1/2* mutation-negative families (Karppinen *et al.*, 2004).
BARD1	c.1972C>T	R658C	2	0.731	N	CM067650	No domain	Y	Benign	3/4	Variant confers increased susceptibility to lung cancer (Rudd *et al.*, 2006)
BRCA1	c.*1086A>C	3'UTR	1	0.000	N	N	3'UTR	N	Not Present	N/A	
BRCA1	c.*1288A>T	3'UTR	1	0.000	N	N	3'UTR	N	Not Present	N/A	
BRCA1	c.*1438G>A	3'UTR	1	0.000	N	N	3'UTR	N	Not Present	N/A	
BRIP1	c.517C>T	R173C	2	0.357	N	CM035889	Helicase ATP-Binding domain & Nuclear localisation signal	Y	Benign/Likely Benign	3/4	Potential breast cancer susceptibility (Wong *et al.*, 2011b)

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by *in Silico*[#]	References and additional notes
BRIP1	c.2108A>T	K703I	1	0.001	N	CM142739	No domain	Y	Not Present	3/4	Ovarian cancer susceptibility (Kanchi *et al.*, 2014)
CDH1	c.1004G>A	R335Q	1	0.002	COSM6023913	N	Cadherin 2 Tandem Repeat Domain	Y	Uncertain Significance	3/4	Confirmed somatic mutation in breast cancer (Zehir *et al.*, 2017) and basal cell adenoma (Jo *et al.*, 2016)
CDH1	c.1493A>C	D498A	1	<0.001	N	N	Cadherin 4 Domain	Y	Uncertain Significance	3/4	
CDH1	c.1774G>A	A592T	1	0.313	COSM19758	CM994192	Cadherin 4 Domain	Y	Conflicting interpretations of pathogenicity (likely benign/ uncertain significance)	3/4	Somatic mutation associated with HER2+ and ER/PR.HER2+ breast carcinomas (Boyault *et al.*, 2012), intestinal adenocarcinoma (Ascano *et al.*, 2001) and thyroid cancer (Soares *et al.*, 1997)
CDKN2A	c.442G>A	A148T	5	1.981	COSM3736958	CM004869	No domain	Y	Benign	2/4	Somatic mutation associated with AML (Hirsch *et al.*, 2016) and pancreatic carcinoma (Dal Molin *et al.*, 2015). Associated with increased risk of melanoma (Debniak *et al.*, 2005b). Low-penetrance predisposition to inherited breast cancer (Debniak *et al.*, 2005a)
CHEK1	c.601A>G	M201V	1	0.002	COSM5662670	N	Protein kinase domain, Interaction with CLSPN	Y	Not Present	3/4	Somatic mutation associated with small cell lung carcinoma (George *et al.*, 2015)
CHEK2	c.254C>T	P85L	1	0.240	N	CM077521	No domain	Y	Benign/Likely Benign	1/4	Pathogenic in somatic osteosarcomas, associated with development of Li-Fraumeni (Miller *et al.*, 2002). Not associated with increased breast cancer susceptibility, however results in 50% reduced activity of CHEK2 protein (Bell *et al.*, 2007)
CHEK2	c.599T>C	I157T	1	0.425	COSM3693990	CM993368	Forkhead associated (FHA) domain – phosphopeptide recognition domain.	Y	Conflicting interpretations of pathogenicity (uncertain significance/likely pathogenic)	2/4	Frequently identified in normal European populations (Allinen *et al.*, 2001). Increased prevalence in breast cancer individuals and identified in both familial and unselected breast cancer cases. Results in a low risk increase in inherited breast cancer susceptibility (Nevanlinna and Bartek, 2006). Somatic mutations associated with cancers in lung, intestine, kidney and ovary (Beltrame *et al.*, 2015, Gadd *et al.*, 2017)

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by *in Silico*[#]	References and additional notes
CHEK2	c.1304C>T	A392V	1	0.000	N	N	Protein Kinase domain	N	Uncertain Significance	3/3	
E2F2	c.794C>T	T265I	1	0.210	N	N	Dimerization domain	Y	Not Present	2/4	
E2F3	c.838T>A	C280S	3	<0.001	N	N	Dimerization domain of E2F transcription factors	Y	Not Present	1/4	
E2F3	c.1315G>A	G439R	1	0.001	N	N	Transactivation domain, Retinoblastoma protein binding domain.	Y	Not Present	1/4	
E2F4	c.917_918insCAG	S307dup	1	0.684	COSM435515	N	No domain	Y	Not present	2/3	Confirmed somatic mutation in breast cancer
E2F4	c.918_920delCAG	S307del	9	0.426	COSM435516	N	No domain	Y	Not present	2/3	Confirmed somatic mutations associated with adenocarcinoma, colon cancer (Giannakis *et al.*, 2014) and mouth carcinoma (Al-Hebshi *et al.*, 2016). Predicted to result in NMD of transcript
EP300	c.2207A>G	H736R	1	0.001	N	N	No domain	Y	Not Present	1/4	
EP300	c.6627_6638delCCAGTTCCAGCA	N2209_Q2213delinsK	1	0.174	COSM6853547	N	Interaction domains with HTLV-1 Tax and NCOA2	Y	Not Present	2/3	
EP300	c.6668A>C	Q2223P	6	2.429	COSM4387478	N	Interaction with NCOA2 domain	Y	Benign/Likely Benign	1/4	Somatic mutation in Haemangioblastoma (Shankar *et al.*, 2014) May be associated with Rubineten-Taybi syndrome.
EP300	c.6964_6964delC	H2324fs	1	0.000	COSM1566439	N	No domain	N	Not present	2/3	Confirmed somatic mutations in intestinal adenocarcinoma (Wang *et al.*, 2014), colon carcinoma (Mouradov *et al.*, 2014, Giannakis *et al.*, 2016) and endometroid carcinoma (Zehir *et al.*, 2017)
EP300	c.6983C>T	S2328F	6	0.001	N	N	No domain	N	Not Present	3/4	
HLTF	c.932A>G	N311S	3	2.617	N	N	No domain	Y	Not Present	1/4	
HLTF	c.2440C>T	P814S	1	0.333	N	N	No domain	Y	Not Present	1/4	
HMMR	c.383C>G	S129C	4	1.026	N	N	Chromosome segregation ATPase	Y	Not present	4/4	
HMMR	c.2163A>C	Q705H	1	0.000	N	N	No domain	N	Not present	3/4	
KAT2B	c.1957C>T	R653W	1	0.312	N	N	No domain	Y	Not Present	2/4	

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by in Silico#	References and additional notes
KAT2B	c.2137C>A	P713T	1	0.869	N	N	No domain	Y	Not Present	3/4	
MRE11A	c.274G>A	E92K	1	0.001	N	N	No domain	Y	Not Present	1/4	
NBN	c.1651delA	R551fs	1	0.000	COSM1458549	N	No domain	N	Not Present	1/3	Confirmed somatic mutation associated with Intestinal adenocarcinoma (Giannakis *et al.*, 2016) and oesophageal squamous cell carcinoma (Lin *et al.*, 2014a) Predicted to result in NMD of transcript.
NBN	c.2165G>C	W722S	1	0.000	N	N	No domain	N	Not Present	3/4	
NQO2	c.86A>G	E29G	6	2.194	N	N	No domain	Y	Not Present	3/4	
NQO2	c.173G>A	G58D	4	2.757	N	N	No domain	Y	Not Present	2/4	
PALB2	c.2816T>G	L939W	2	0.094	N	CM105609	WD1 repeat, required for POLH DNA synthesis stimulation, Interaction with RAD51, BRCA2 and POLH	Y	Conflicting interpretations of pathogenicity (Benign/ likely benign/uncertain significance)	3/4	
PALB2	c.2993G>A	G998E	9	1.615	N	CM098533	As above	Y	Benign/Likely Benign	3/4	Increased risk of inherited breast cancer (Sluiter *et al.*, 2009)
PALB2	c.3116delA	N1039X (Ter)	2	0.001	N	CM070242	As above	Y	Pathogenic/Likely pathogenic	3/3	Known pathogenic in literature (Rahman *et al.*, 2007)
PKMYT1	c.434T>C	F145S	1	0.000	N	N	Protein kinase domain	N	Not present	3/4	
PKMYTI	c.451C>G	R151G	1	0.016	N	N	Protein kinase domain	Y	Not present	3/4	
PRKDC	c.3730insG	L1244P	1	2.609	N	N	No domain	Y	Not present	2/4	
PRKDC	c.5119T>A	L1707Q	4	0.218	N	N	No domain	Y	Uncertain Significance	3/4	
PRKDC	c.8694C>T	R2898C	14	3.780	N	N	FAT Domain/KIP Binding Domain	Y	Benign	3/4	
PRKDC	c.11805G>A	G3935S	1	0.051	N	N	PI3K/PI4K Domain	Y	Not Present	2/4	
PRKDC	c.11989T>C	L3996P	1	0.022	N	N	PI3K/PI4K Domain	Y	Not Present	2/3	
RAD50	c.379G>A	V127I	1	0.161	N	N	No domain	Y	Conflicting interpretations of pathogenicity (benign/uncertain significance)	1/4	

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by in Silico[#]	References and additional notes
RAD50	c.980G>A	R327H	1	0.330	N	CM068746	Coiled-coil domain	Y	Conflicting interpretations of pathogenicity (benign/uncertain Significance)	2/4	Results in reduced RAD50 protein and associated with low-penetrance predisposition to inherited breast cancer (Tommiska *et al.*, 2006).
RAD50	c.2793_2794delCAinsAC	931_932NKdel/insKQ	2	0.000	N	N	Coiled-coil domain	N	Uncertain Significance	1/3	Predicted to result in NMD of transcript.
RAD51	c.824A>G	D275G	1	0.000	N	N	No domain	N	Not Present	2/4	
RAD51D	c.26G>C	C9S	1	0.041	N	CM128416	Preferentially binds ssDNA	Y	Not Present	3/4	Susceptibility to inherited breast and ovarian cancer (Gutierrez-Enriquez *et al.*, 2014)
RAD51D	c.383G>A	G128D	1	0.000	COSM6003014	N	No domain	N	Not Present	2/4	Somatic mutation in prostate carcinoma (Kumar *et al.*, 2016)
RAD51D	c.497T>C	L184P	1	0.001	N	N	No domain	Y	Uncertain Significance	3/4	
RAD51D	c.698A>G	E233G	5	1.144	N	CM045804	No domain	Y	Benign/Likely Benign	2/4	Associated with low-penetrance predisposition to inherited breast cancer (Rodriguez-Lopez *et al.*, 2004)
RBL1	c.940G>A	G314S	1	0.000	N	N	No domain	N	Not Present	3/4	
RBL2	c.2487A>T	R829S	2	0.719	N	N	Domain B, Spacer region	Y	Not Present	2/4	
RFC3	c.246T>G	I82M	1	0.013	N	N	No domain	Y	Not Present	2/4	
RFC4	c.1034A>G	H345R	1	0.001	N	N	No domain	Y	Not Present	2/4	
RPA1	c.2T>C	M1T	1	0.000	N	N	No domain	N	Not present	3/4	
RPS6KA1	c.1125delC	S375fs	1	0.000	N	N	ACG Kinase C Terminal	N	Not present	2/3	
RPS6KA1	c.1141delG	S378Afster18	2	0.000	N	N	ACG Kinase C Terminal	N	Not present	2/3	Predicted to result in NMD of transcript
SLC19A1	c.395C>T	A132V	1	0.024	N	N	No domain	Y	Not Present	3/4	
TP53	c.869G>A	R290H	1	0.015	COSM6023506	CM065493	Binding domain of HIPK1, ZNF385A, AXIN1 and E4F1	Y	Conflicting interpretations of pathogenicity (likely benign/uncertain significance)	0/3	Not confirmed somatic mutation in Myelodysplastic syndrome; Possibly involved in Li-Fraumeni Syndrome (Anensen *et al.*, 2006) and hereditary cancer predisposing syndrome.
UIMC1	c.43C>T	R15W	3	1.057	N	N	Necessary for transcriptional repression	Y	Not Present	3/4	

Gene	Transcript Variant	Protein Variant	Number of patients	MAF (%)	COSMIC	HGMD	Protein Domain	gnomAD	ClinVar	Damaging by in Silico[#]	References and additional notes
UIMC1	c.999G>T	Q333H	1	0.004	N	N	AIR region	Y	Not Present	3/4	
UIMC1	c.1690T>C	Y564H	2	0.207	N	N	Zinc finger like region	Y	Not Present	3/4	
UIMC1	c.1756G>T	A586S	1	0.047	N	N	No domain	Y	Not Present	2/4	
UIMC1	c.2045_2046delTT	F682C fsTer14	1	0.003	N	N	No domain	Y	Not Present	1/3	
WEE1	c.628G>T	G210C	1	0.704	N	N	No domain	Y	Not Present	1/4	
XRCC2	c.509A>G	E170G	1	0.003	N	N	No domain	Y	Not Present	2/4	

Table 5.2: *In silico* analysis of predicted pathogenic variants. MAF; Minimum allele frequency as determined by gnomAD (%). Effect on splicing predicted by Human Splicing Finder. ESE; exonic splicing enhancer, ESS; exonic splicing silencer. SIFT analysis of SNPs; Scores range from 0-1. Scores 0.2 – 0.85, possibly damaging, ≥0.85, probably damaging. SIFT analysis of indels; Scores range from 0-1. Scores 0 – 0.5, neutral, ≥0.5, damaging.PROVEAN analysis, scores <-2.5, predicted deleterious, PolyPhen-2 analysis; Scores range from 0 – 1. Scores ≤0.2, benign, 0.2 – 0.85, possibly damaging, ≥0.85, probably damaging. Align GV-GD analysis; GV Score ranges from 0-200, GD score ranges from 0->200. GV and GD Scores used to determine a class ranking of the effect of the amino acid substitution, with C0 corresponding to neutral, C15; moderately likely and C65 corresponding to likely deleterious. Sanger sequencing confirmation of variants Y; Yes, NC; Not confirmed.

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
ATF1	c.571C>G	P191A	rs2230674	2.376	Y - ESE signal damaged	Damaging	0.033	Deleterious	-5.265	Probably Damaging	0.999	151.88	0	Class C0	Least Likely	Y
ATM	c.2T>C	M1T	not found	<0.001	N	Damaging	0	Neutral	-1.87	Possibly Damaging	0.921	0	81.04	Class C65	Most Likely	Y
ATM	c.998C>T	S333F	rs28904919	0.134	Y- creation of new ESS site, destruction of ESE site	Damaging	0.008	Neutral	-1.592	Possibly Damaging	0.731	115.24	85.89	Class C15	Likely	NC
ATM	c.1010G>A	R337H	rs202160435	0.006	N	Damaging	0.003	Neutral	-2.431	Probably Damaging	1	160.7	0	Class C0	Least Likely	Y
ATM	c.1892C>T	P631L	Not found	0.000	Y - creation of new ESS site, destruction of ESE site	Damaging	0.02	Neutral	-2.153	Benign	0.002	209.54	94.04	Class C0	Least Likely	NC
ATM	c.2119T>C	S707P	rs4986761	0.774	N	Tolerated	0.203	Neutral	0.383	Benign	0	157.85	0	Class C0	Least Likely	Y
ATM	c.2572T>C	F858L	rs1800056	0.845	Y - creation of ESS site.	Tolerated	0.075	Deleterious	-2.574	Possibly Damaging	0.825	188.97	4.86	Class C0	Least Likely	Y
ATM	c.3161C>G	P1054R	rs1800057	1.658	Y - activation of exonic cryptic donor site	Damaging	0	Deleterious	-5.614	Probably Damaging	1	0	102.7	Class C65	Most Likely	Y
ATM	c.4258C>T	L1420F	rs1800058	1.114	Y – creation of new ESS site	Tolerated	0.061	Deleterious	-2.536	Benign	0.238	31.78	31.28	Class C0	Least Likely	NC
ATM	c.5558A>T	D1853V	rs1801673	0.443	Y - activation of exonic cryptic donor site, damage to ESE site	Damaging	0.007	Deleterious	-4.519	Possibly Damaging	0.928	93.77	77.61	Class C15	Likely	NC
ATM	c.7390T>C	C2464R	rs55801750	0.037	N	Tolerated	0.573	Deleterious	-3	Possibly Damaging	0.806	223.96	13.27	Class C0	Least Likely	NC

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
ATM	c.8305_8306insC	W2769fs	Not found	0.000	Y - activation of exonic cryptic donor site, damage to ESE site	Damaging	0.858	Deleterious	-18.06	-	-	275.49	0	Class C0	Least likely	NC
BARD1	c.1670G>C	C557S	rs28997576	1.435	N	Tolerated	0.41	Deleterious	-2.592	Benign	0.04	228.97	0	Class C0	Least Likely	NC
BARD1	c.1972C>T	R658C	rs3738888	0.731	Y - creation of ESS site	Damaging	0.003	Deleterious	-4.015	Probably Damaging	0.995	264.55	0	Class C0	Least Likely	Y
BRCA1	c.*1086A>C	3'UTR	not found	0.000	N	-	-	-	-	-	-	-	-	-	-	Y
BRCA1	c.*1288A>T	3'UTR	not found	0.000	N	-	-	-	-	-	-	-	-	-	-	Y
BRCA1	c.*1438G>A	3'UTR	not found	0.000	N	-	-	-	-	-	-	-	-	-	-	Y
BRIP1	c.517C>T	R173C	rs4988345	0.359	Y - creation of ESS site	Damaging	0.002	Deleterious	-2.542	Probably Damaging	1	231.1	0	Class C0	Least Likely	NC
BRIP1	c.2108A>T	K703I	not found	0.003	Y - ESE site damaged	Damaging	0	Deleterious	-7.07	Probably Damaging	1	244.44	0	Class C0	Least Likely	NC
CDH1	c.1004G>A	R335Q	rs373364873	0.002	N	Damaging	0	Deleterious	-3.846	Probably Damaging	0.997	188.1	0	Class C0	Least Likely	Y
CDH1	c.1493A>C	D498A	not found	<0.001	Y - ESE site damaged	Damaging	0.05	Deleterious	-4.582	Probably Damaging	0.976	186.13	0	Class C0	Least Likely	NC
CDH1	c.1774G>A	A592T	rs35187787	0.313	Y - ESE site damaged	Damaging	0	Deleterious	-2.75	Possibly Damaging	0.492	253.4	0	Class C0	Least Likely	Y
CDKN2A	c.442G>A	A148T	rs3731249	1.981	Y - ESE site damaged	Damaging	0.011	Neutral	-0.863	Possibly Damaging	0.487	241.23	0	Class C0	Least Likely	Y
CHEK1	c.601A>G	M201V	not found	0.002	N	Tolerated	0.34	Deleterious	-2.951	Probably Damaging	0.999	201.18	0	Class C0	Least Likely	NC
CHEK2	c.254C>T	P85L	rs17883862	0.240	Y-ESE site damaged	Tolerated	0.19	Neutral	0.096	Possibly Damaging	0.728	241.55	9.71	Class C0	Least Likely	Y
CHEK2	c.599T>C	I157T	rs17879961	0.425	N	Damaging	0.014	Neutral	-1.893	Possibly Damaging	0.514	187.3	0	Class C0	Least Likely	NC
CHEK2	c.1304C>T	A392V	rs1555913484	0.000	Y- activation of exonic cryptic donor, ESE site damaged	Damaging	0.01	Deleterious	-3.847	Probably Damaging	1	232.54	0	Class C0	Least Likely	NC
E2F2	c.794C>T	T265I	rs139052092	0.210	Y - ESE site damaged	Tolerated	0.184	Deleterious	-3.66	Possibly Damaging	0.943	130.23	59.51	Class C0	Least Likely	NC

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
E2F3	c.838T>A	C280S	not found	<0.001	Y- ESE site damaged	Tolerated	0.17	Deleterious	-5.478	Benign	0.006	272.33	0	Class C0	Least Likely	Y
E2F3	c.1315G>A	G439R	rs368121892	0.001	Y – Creation of new ESS site, activation of exonic cryptic donor site, damage to ESE site	Tolerated	0.089	Neutral	-1.398	Probably Damaging	0.994	353.86	0	Class C0	Least Likely	NC
E2F4	c.917_918insCAG	S307dup	not found	0.684	N	Damaging	0.667	Deleterious	-8.468	-	-	196.64	0	Class C0	Least Likely	NC
E2F4	c.918_920delCAG	S307del	rs3830472	0.426	Y - activation of an exonic cryptic donor site	Damaging	0.858	Deleterious	-8.468	-	-	-	-	-	-	NC
EP300	c.2207A>G	H736R	not found	0.001	N	Tolerated	0.25	Neutral	-0.577	Possibly Damaging	0.843	164.25	0	Class C0	Least Likely	NC
EP300	c.6627_6638delCCAGTTCCAGCA	N2209_Q2213delinsK	not found	0.174	N	Damaging	0.858	Deleterious	-15.47	-	-	268.49	0	Class C0	Least Likely	NC
EP300	c.6668A>C	Q2223P	rs1046088	2.429	N	Tolerated	0.136	Deleterious	-2.875	Benign	0	257.76	0	Class C0	Least Likely	Y
EP300	c.6964delC	H2324fs	Not found	0.000	N	Damaging	0.783	Deleterious	-10.56	-	-	239.68	0	C0	Least Likely	NC
EP300	c.6983C>T	S2328F	not found	0.001	N	Damaging	0.01	Deleterious	-4.778	Probably Damaging	0.994	260.23	28.53	Class C0	Least Likely	NC
HLTF	c.932A>G	N311S	rs2305868	2.617	Y - activation of exonic cryptic donor, ESE site damaged	Tolerated	0.064	Neutral	-2.443	Possibly Damaging	0.679	194.2	6.18	Class C0	Least Likely	Y
HLTF	c.2440C>T	P814S	rs61750364	0.333	Y - ESE site damaged, creation of ESS site	Tolerated	0.221	Neutral	-1.399	Possibly Damaging	0.935	187.2	2.75	Class C0	Least Likely	Y
HMMR	c.383C>G	S129C	rs34815524	1.026	Y - creation of ESS site	Damaging	0.007	Deleterious	-3.295	Probably Damaging	1	155.86	91.34	Class C15	Likely	Y

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
HMMR	c.2163A>C	Q705H	Not found	0.000	Y - ESE site damaged	Damaging	0.05	Deleterious	-3.08	Probably Damaging	0.965	353.86	0	Class C0	Least Likely	NC
KAT2B	c.1957C>T	R653W	rs116196143	0.312	Y - ESE site damaged	Tolerated	0.107	Deleterious	-5.619	Probably Damaging	0.996	353.86	0	Class C0	Least Likely	NC
KAT2B	c.2137C>A	P713T	rs148960024	0.869	Y - ESE site damaged	Damaging	0.042	Deleterious	-4.334	Possibly Damaging	0.901	353.86	0	Class C0	Least Likely	NC
MRE11A	c.274G>A	E92K	not found	0.001	Y - ESE site damaged	Damaging	0.02	Deleterious	-2.788	Possibly Damaging	0.619	85.44	0	Class C0	Least Likely	NC
NBN	c.1651delA	R551fs	not found	0.000	Y - ESE site damaged, creation of new ESS site	Damaging	0.858	Neutral	-0.888	-	-	183.83	0	Class CO	Least Likely	NC
NBN	c.2165G>C	W722S	not found	0.000	N	Damaging	0	Deleterious	-8.854	Probably Damaging	1	148.51	64.04	Class C0	Least Likely	NC
NQO2	c.86A>G	E29G	rs17136117	2.194	Y - creation of new ESS site	Tolerated	0.195	Deleterious	-2.877	Possibly Damaging	0.458	0	97.85	Class C65	Most Likely	NC
NQO2	c.173G>A	G58D	rs17300141	2.757	Y - creation of new ESS site	Tolerated	0.179	Deleterious	-4.416	Benign	0	60	81.64	Class C15	Likely	Y
PALB2	c.2816T>G	L939W	rs45478192	0.094	Y- ESE site damaged, creation of new ESS site	Damaging	0	Deleterious	-5.281	Probably Damaging	1	248.96	59.78	Class C0	Least Likely	NC
PALB2	c.2993G>A	G998E	rs45551636	1.615	N	Damaging	0	Deleterious	-6.233	Probably Damaging	1	199.95	0	Class C0	Least Likely	Y
PALB2	c.3116delA	N1039X	not found	0.001	Y - ESE site damaged	Damaging	0.529	Deleterious	-13.12	Probably Damaging	1	-	-	-	-	Y
PKMYT1	c.434T>C	F145S	Not found	0.000	Y - creation of new ESS site	Damaging	0	Deleterious	-6.967	Probably Damaging	1	171.82	0	Class C0	Least Likely	NC
PKMYT1	C.451C>G	R151G	Not found	0.016	Y- ESE site damaged, creation of new ESS site	Damaging	0	Deleterious	-6.104	Probably Damaging	1	158.9	0	Class C0	Least Likely	NC
PRKDC	c.3730insG	L1244P	rs11411516	2.609	N	Tolerated	0.773	Deleterious	-3.039	Possibly Damaging	0.575	182.09	0	Class C0	Least Likely	NC
PRKDC	c.5119T>A	L1707Q	rs202110076	0.218	N	Damaging	0	Deleterious	-2.647	Probably Damaging	1	199.95	0	Class C0	Least Likely	Y

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
PRKDC	c.8694C>T	R2899C	rs4278157	3.780	Y - activation of cryptic donor site, damage to ESE site	Damaging	0.013	Deleterious	-2.601	Possibly Damaging	0.929	180.39	91.34	Class C0	Least Likely	Y
PRKDC	c.11805G>A	G3935S	rs55670423	0.051	N	Tolerated	0.081	Deleterious	-5.259	Probably Damaging	1	151.88	36.4	Class C0	Least Likely	NC
PRKDC	c.11989T>C	L3996P	rs201883689	0.022	N	Damaging	0.002	Deleterious	-4.985	-	-	193.06	0	Class C0	Least Likely	NC
RAD50	c.379G>A	V127I	rs28903086	0.161	N	Tolerated	0.073	Neutral	-0.818	Probably Damaging	0.999	83.89	18.96	Class C0	Least Likely	NC
RAD50	c.980G>A	R327H	rs28903091	0.330	N	Damaging	0.022	Neutral	-2.052	Probably Damaging	0.994	125.13	4.81	Class C0	Least Likely	NC
RAD50	c.2793_2794delCAinsAC	N931_K932del/insKQ	Not found	0.000	N	Damaging	0.858	Neutral	-1.2	-	-	163.58	0	Class C0	Least Likely	NC
RAD51	c.824A>G	D275G	not found	0.000	N	Tolerated	0.5	Deleterious	-6.099	Possibly Damaging	0.938	276.16	0	Class C0	Least Likely	NC
RAD51D	c.26G>C	C9S	not found	0.041	Y - ESE site damaged	Damaging	0.04	Deleterious	-6.191	Possibly Damaging	0.948	201.58	23.3	Class C0	Least Likely	NC
RAD51D	c.383G>A	G128D	not found	0.000	N	Damaging	0.05	Neutral	-0.346	Possibly Damaging	0.488	140.07	22.66	Class C0	Least Likely	NC
RAD51D	c.497T>C	L185P	not found	0.001	N	Damaging	0	Deleterious	-5.96	Possibly Damaging	0.456	204.97	40.92	Class C0	Least Likely	NC
RAD51D	c.698A>G	E233G	rs28363284	1.144	Y - ESE site damaged	Tolerated	0.451	Deleterious	-3.368	Probably Damaging	0.973	126.08	81.06	Class C0	Least Likely	NC
RBL1	c.940G>A	G314S	not found	0.000	Y - activation of cryptic donor site, activation of cryptic acceptor site, damage to ESE site, creation of new ESS site	Damaging	0	Deleterious	-5.456	Probably Damaging	1	258.19	0	Class C0	Least Likely	Y
RBL2	c.2487A>T	R829S	rs61747629	0.719	Y - ESE site damaged	Damaging	0.048	Deleterious	-3.015	Benign	0.013	235.1	0	Class C0	Least Likely	Y
RFC3	c.246T>G	I82M	not found	0.13	Y - creation of new ESS site	Damaging	0.03	Neutral	-1.762	Possibly Damaging	0.872	125.38	0	Class C0	Least Likely	NC

Gene	Transcript Variant	Protein Variant	dbSNP ID	MAF (%)	Predicted to affect splicing	SIFT Analysis		Provean Analysis		PolyPhen-2 Analysis		Align GVGD Analysis				Variant confirmed by Sanger
						Prediction	Score	Prediction	Score	Prediction	Score	GV	GD	Score	Prediction	
RFC4	c.1034A>G	H345R	not found	0.001	Y - activation of cryptic donor site, ESE site damaged	Damaging	0.05	Deleterious	-3.335	Benign	0.002	353.86	0	Class C0	Least Likely	NC
RPA1	c.2T>C	M1T	Not found	0.000	N	Damaging	0	Neutral	-1.963	Possibly Damaging	0.501	0	81.04	Class C65	Most Likely	NC
RPS6KA1	c.1125delC	S375fs	Not found	0.000	N	Damaging	0.858	Deleterious	-11.31	-	-	157.56	0	Class C0	Least Likely	NC
RPS6KA1	c.1141delG	S378AfsTer18	Not found	0.000	N	Damaging	0.858	Deleterious	-11.31	-	-	175.88	47.63	Class C0	Least Likely	NC
SLC19A1	c.395C>T	A132V	Not found	0.024	Y - activation of cryptic donor site, creation of new ESS site	Damaging	0.01	Deleterious	-2.733	Probably Damaging	0.944	174.42	0	Class C0	Least Likely	NC
TP53	c.869G>A	R290H	rs55819519	0.015	Y - ESE site damaged	Tolerated	0.03	Neutral	-2.031	Benign	0	146.68	0	Class C0	Least Likely	Y
UIMC1	c.43C>T	R15W	rs13167812	1.057	Y - creation of new ESS site	Damaging	0.008	Deleterious	-3.33	Possibly Damaging	0.825	133.59	65.28	Class C0	Least Likely	Y
UIMC1	c.999G>T	Q333H	rs200923725	0.004	Y - creation of new ESS site, ESE site damaged	Damaging	0.013	Deleterious	-3.151	Probably Damaging	0.999	90.41	14.28	Class C0	Least Likely	NC
UIMC1	c.1690T>C	Y564H	rs115224789	0.207	N	Damaging	0	Deleterious	-3.023	Probably Damaging	1	191.17	0	Class C0	Least Likely	Y
UIMC1	c.1756G>T	A586S	rs144604125	0.047	N	Damaging	0.024	Neutral	-1.675	Probably Damaging	0.98	172.33	0	Class C0	Least Likely	NC
UIMC1	c.2045_2046delTT	F682CfsTer14	not found	0.003	N	Tolerated	0.818	Deleterious	-4.74	-	-	171.82	91.34	Class C0	Least Likely	NC
WEE1	c.628G>T	G210C	rs34412975	0.704	Y - activation of cryptic donor site, creation of new ESS site.	Tolerated	0.22	Neutral	-1.076	Possibly Damaging	0.903	190.14	97.52	Class C0	Least Likely	NC
XRCC2	c.509A>G	E170G	not found	0.003	Y - activation of cryptic donor site, creation of new ESS site, ESE site damaged.	Damaging	0.04	Neutral	-2.25	Possibly Damaging	0.883	97.85	0	Class C0	Least Likely	NC

Table 5.3: Statistical analysis of significance of identified potentially pathogenic variants within the patient cohort. A one proportion Z- test was carried out on the cohort of 131 individuals screened within this study and compared to the identified allele frequencies reported in gnomAD. Variants with statistical significance (p>0.05) indicated in red. MAF; minimum allele frequency as determined by GnomAD (%) CI; confidence interval.

Gene	Transcript Variant	Protein Variant	Number of individuals	MAF (%)	Z-statistic	Significance (P)	95% CI of observed proportion
ATF1	c.571C>G	P191A	9	2.376	1.137	0.2555	1.58% to 6.24%
ATM	c.2T>C	M1T	1	<0.001	23.341	<0.0001	0.01% to 2.11%
ATM	c.998C>T	S333F	1	0.134	1.096	0.2731	0.01% to 2.11%
ATM	c.1010G>A	R337H	2	0.006	15.827	<0.0001	0.09% to 2.73%
ATM	c.1892C>T	P631L	1	0.000	617.785	<0.0001	0.01% to 2.11%
ATM	c.2119T>C	S707P	3	0.774	0.685	0.4932	0.24% to 3.31%
ATM	c.2572T>C	F858L	3	0.845	0.531	0.5958	0.24% to 3.31%
ATM	c.3161C>G	P1054R	7	1.658	1.285	0.1988	1.08% to 5.43%
ATM	c.4258C>T	L1420F	1	1.114	1.129	0.2587	0.01% to 2.11%
ATM	c.5558A>T	D1853V	1	0.443	0.149	0.8812	0.01% to 2.11%
ATM	c.7390T>C	C2464R	1	0.037	2.901	0.0037	0.01% to 2.11%
ATM	c.8305_8306insC	W2769fs	1	0.000	617.785	<0.0001	0.01% to 2.11%
BARD1	c.1670G>C	C557S	5	1.435	0.644	0.5194	0.62% to 4.40%
BARD1	c.1972C>T	R658C	2	0.731	0.061	0.9510	0.09% to 2.73%
BRCA1	c.*1086A>C	3'UTR	1	0.000	617.785	<0.0001	0.01% to 2.11%
BRCA1	c.*1288A>T	3'UTR	1	0.000	617.785	<0.0001	0.01% to 2.11%
BRCA1	c.*1438G>A	3'UTR	1	0.000	617.785	<0.0001	0.01% to 2.11%
BRIP1	c.517C>T	R173C	2	0.359	1.094	0.2738	0.09% to 2.73%
BRIP1	c.2108A>T	K703I	1	0.003	11.191	<0.0001	0.01% to 2.11%
CDH1	c.1004G>A	R335Q	1	0.002	13.742	<0.0001	0.01% to 2.11%
CDH1	c.1493A>C	D498A	1	<0.001	23.358	<0.0001	0.01% to 2.11%
CDH1	c.1774G>A	A592T	1	0.313	0.199	0.8423	0.01% to 2.11%
CDKN2A	c.442G>A	A148T	5	1.981	0.084	0.9328	0.62% to 4.40%
CHEK1	c.601A>G	M201V	1	0.002	13.742	<0.0001	0.01% to 2.11%
CHEK2	c.254C>T	P85L	1	0.240	0.469	0.6393	0.01% to 2.11%
CHEK2	c.599T>C	I157T	1	0.425	0.108	0.9141	0.00% to 2.11%
CHEK2	c.1304C>T	A392V	1	0.000	617.785	<0.0001	0.01% to 2.11%
E2F2	c.794C>T	T265I	1	0.210	0.607	0.5438	0.01% to 2.11%
E2F3	c.838T>A	C280S	3	<0.001	92.871	<0.0001	0.24% to 3.31%
E2F3	c.1315G>A	G439R	1	0.001	19.485	<0.0001	0.01% to 2.11%
E2F4	c.917_918insCAG	S307dup	1	0.684	0.594	0.5527	0.01% to 2.11%
E2F4	c.918_920delCAG	S307del	9	0.426	7.478	<0.0001	1.58% to 6.42%
EP300	c.2207A>G	H736R	1	<0.001	30.935	<0.0001	0.01% to 2.11%
EP300	c.6627_6638delCCAGTTCCAGCA	N2209_Q2213delinsK	1	0.174	0.807	0.4199	0.01% to 2.11%`
EP300	c.6668A>C	Q2223P	6	2.429	0.146	0.8839	0.84% to 4.92%
EP300	c.6964_6964delC	H2324fs	1	0.000	617.785	<0.0001	0.01% to 2.11%
EP300	c.6983C>T	S2328F	6	0.001	117.169	<0.0001	0.84% to 4.92%
HLTF	c.932A>G	N311S	3	2.617	1.492	0.1356	0.24% to 3.31%

Gene	Transcript Variant	Protein Variant	Number of individuals	MAF (%)	Z-statistic	Significance (P)	95% CI of observed proportion
HLTF	c.2440C>T	P814S	1	0.333	0.137	0.8912	0.01% to 2.11%
HMMR	c.383C>G	S129C	4	1.026	0.804	0.4212	0.42% to 3.86%
HMMR	c.2163A>C	Q705H	1	0.000	195.356	<0.0001	0.01% to 2.11%
KAT2B	c.1957C>T	R653W	1	0.312	0.202	0.8398	0.01% to 2.11%
KAT2B	c.2137C>A	P713T	1	0.869	0.850	0.3954	0.01% to 2.11%
MRE11A	c.274G>A	E92K	1	0.001	19.485	<0.0001	0.01% to 2.11%
NBN	c.1651delA	R551fs	1	0.000	195.356	<0.0001	0.01% to 2.11%
NBN	c.2165G>C	W722S	1	0.000	195.356	<0.0001	0.01% to 2.11%
NQO2	c.86A>G	E29G	6	2.194	0.106	0.9155	0.84% to 4.92%
NQO2	c.173G>A	G58D	4	2.757	1.216	0.2239	0.42% to 3.86%
PALB2	c.2816T>G	L939W	2	0.094	3.535	0.0004	0.09% to 2.73%
PALB2	c.2993G>A	G998E	9	1.615	2.337	0.0194	1.58% to 6.42%
PALB2	c.3116delA	N1039X (Ter)	2	0.001	39.022	<0.0001	0.09% to 2.73%
PKMYT1	c.434T>C	F145S	1	0.000	195.356	<0.0001	0.01% to 2.11%
PKMYTI	c.451C>G	R151G	1	0.016	4.680	<0.0001	0.01% to 2.11%
PRKDC	c.3730insG	L1244P	1	2.609	2.262	0.0237	0.01% to 2.11%
PRKDC	c.5119T>A	L1707Q	4	0.218	4.542	<0.0001	0.42% to 3.86%
PRKDC	c.8694C>T	R2898C	14	3.780	1.327	0.1845	2.95% to 8.80%
PRKDC	c.11805G>A	G3935S	1	0.051	2.371	0.0178	0.01% to 2.11%
PRKDC	c.11989T>C	L3996P	1	0.022	3.925	0.0001	0.01% to 2.11%
RAD50	c.379G>A	V127I	1	0.161	0.891	0.3730	0.01% to 2.11%
RAD50	c.980G>A	R327H	1	0.330	0.146	0.8841	0.01% to 2.11%`
RAD50	c.2793_2794delCAinsAC	931_932NKdel/insKQ	2	0.000	390.727	<0.0001	0.09% to 2.73%
RAD51	c.824A>G	D275G	1	0.000	195.356	<0.0001	0.01% to 2.11%
RAD51D	c.26G>C	C9S	1	0.041	2.724	0.0065	0.01% to 2.11%
RAD51D	c.383G>A	G128D	1	0.000	195.356	<0.0001	0.01% to 2.11%
RAD51D	c.497T>C	L184P	1	0.001	19.485	<0.0001	0.01% to 2.11%
RAD51D	c.698A>G	E233G	5	1.144	1.163	0.2446	0.62% to 4.40%
RBL1	c.940G>A	G314S	1	0.000	19.485	<0.0001	0.01% to 2.11%
RBL2	c.2487A>T	R829S	2	0.719	0.085	0.9323	0.09% to 2.73%
RFC3	c.246T>G	I82M	1	0.013	5.234	<0.0001	0.01% to 2.11%
RFC4	c.1034A>G	H345R	1	0.001	19.485	<0.0001	0.01% to 2.11%
RPA1	c.2T>C	M1T	1	0.000	617.785	<0.0001	0.01% to 2.11%
RPS6KA1	c.1125delC	S375fs	1	0.000	617.785	<0.0001	0.01% to 2.11%
RPS6KA1	c.1141delG	S378Afster18	2	0.000	1235.601	<0.0001	0.09% to 2.73%
SLC19A1	c.395C>T	A132V	1	0.024	3.737	0.0002	0.01% to 2.11%
TP53	c.869G>A	R290H	1	0.015	4.846	<0.0001	0.01% to 2.11%

Gene	Transcript Variant	Protein Variant	Number of individuals	MAF (%)	Z-statistic	Significance (P)	95% CI of observed proportion
UIMC1	c.43C>T	R15W	3	1.057	0.139	0.8892	0.24% to 3.31%
UIMC1	c.999G>T	Q333H	1	0.004	9.666	<0.0001	0.01% to 2.11%
UIMC1	c.1690T>C	Y564H	2	0.207	2.230	0.0257	0.12% to 2.84%
UIMC1	c.1756G>T	A586S	1	0.047	2.499	0.0124	0.01% to 2.11%
UIMC1	c.2045_2046delTT	F682CfsTer14	1	0.003	11.191	<0.0001	0.01% to 2.11%
WEE1	c.628G>T	G210C	1	0.704	0.624	0.5326	0.01% to 2.11%
XRCC2	c.509A>G	E170G	1	0.003	11.191	<0.0001	0.01% to 2.11%

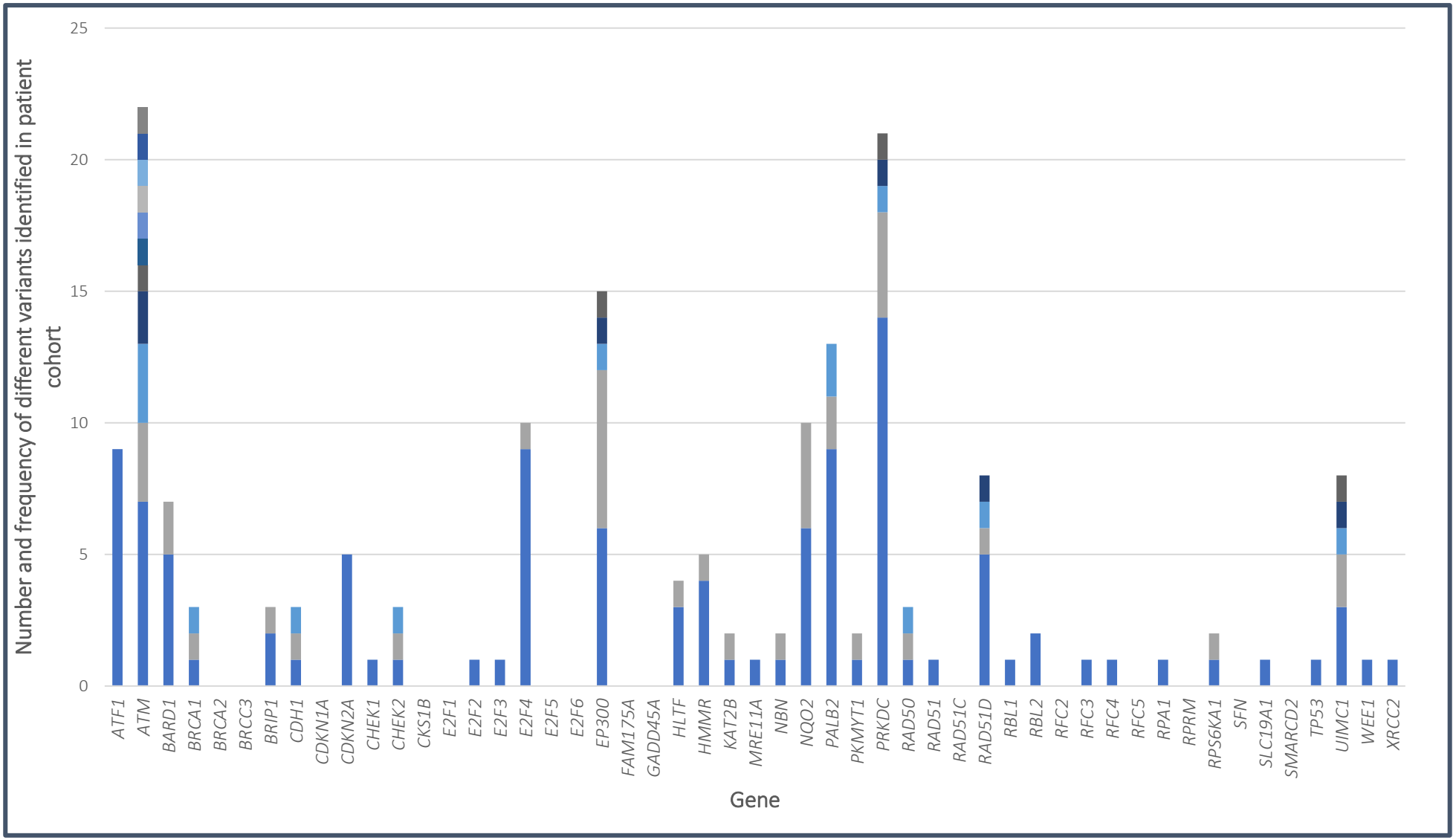


Figure 5.5: Spread and frequency of potentially pathogenic variants identified from analysis of 131 individuals. Each different coloured box indicates a different variant identified within the gene indicated. The size of each coloured box indicates the number of individuals with the identified mutation. For example, 3 different mutations were identified in *PALB2*, with 1 variant identified in 9 individuals and 2 variants identified in 2 individuals each.

As can be seen in **Figure 5.5**, potentially pathogenic variants were identified in 36 of the 51 genes analysed, with the greatest number of unique variants being identified in *ATM* (11 variants), *UIMC1* (5 variants), *EP300* (5 variants), PRKDC (5 variants) and *RAD51D* (4 variants). The same variants were identified in 2 or more individuals in multiple genes including *ATM* (5 variants), *ATF1* (1 variant in 9 samples), *CDKN2A* (1 variant in 5 samples), *NQO2* (2 variants), *PALB2* (3 variants) and PRKDC (2 variants in 14 and 4 samples respectively).

5.3.4 Predicted pathogenic variants were confirmed by sanger sequencing.

Fifty-four of the 166 predicted pathogenic variants or VUS were selected for confirmation via Sanger sequencing. Fifty-two were confirmed as true variants. An example of such confirmation is shown in **Figure 5.6** for NQO2:c.173G>A for sample SABC042.

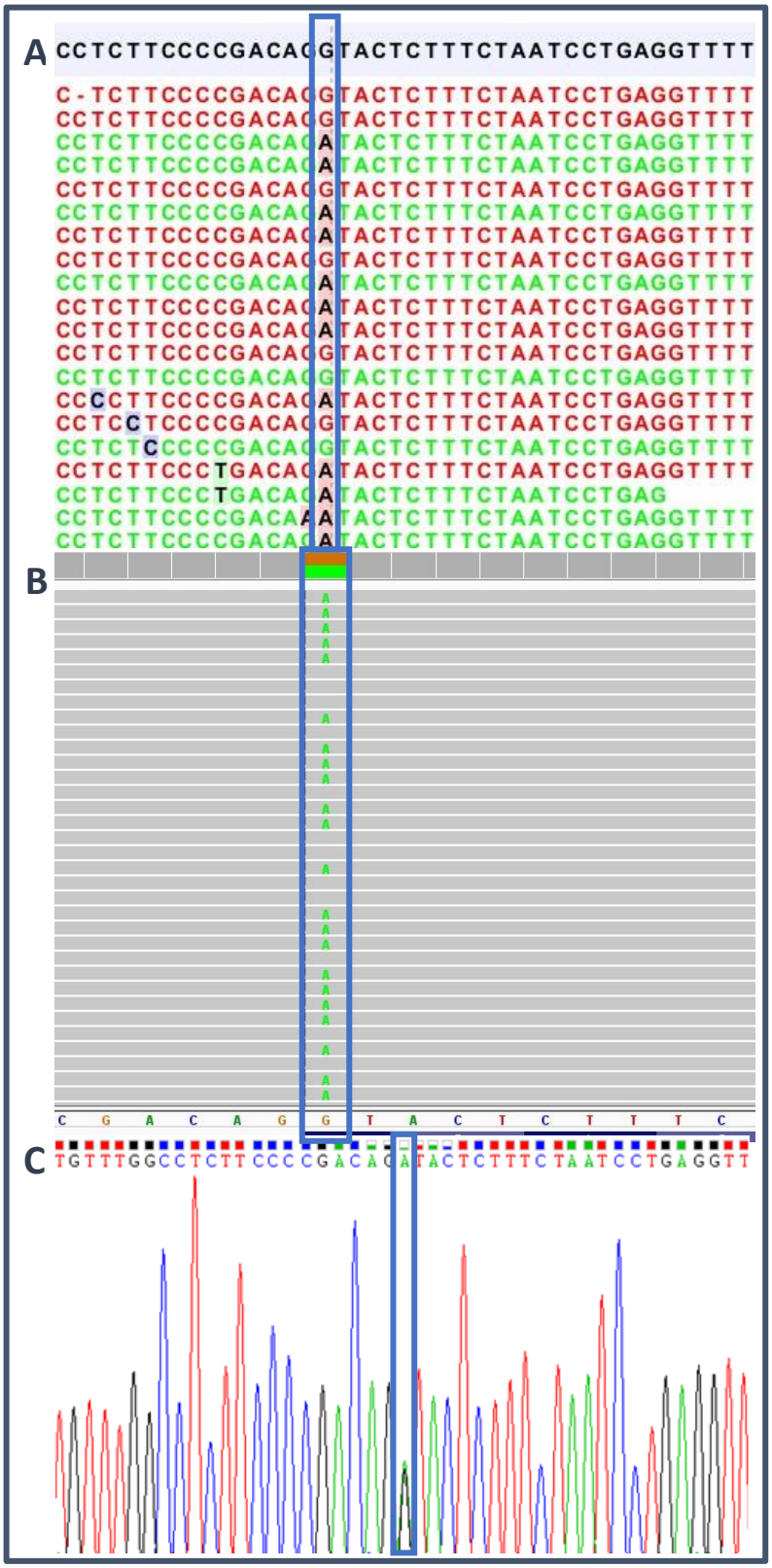


Figure 5.6: Variant Identification and confirmation of NQO2:c.173G>A in SABC042. Variant identification in both **A.** CLC Genomics Workbench; **B.** IGV **C.** Variant confirmation by Sanger sequencing. Variant shown as heterozygous in forward direction, highlighted by blue boxes. Both panels A and B include hg19 reference genome in addition to MPS sequencing data.

5.3.5 Selected variants of interest for further analysis from patient cohort

Of the 82 potentially pathogenic variants identified within the patient cohort, 3 were selected for further analysis. These variants were ATM:c.2119C>T (p.S707P), HMMR:c.383C>G (p.S129C) and UIMC1:c.1690T>C (p.Y564H).

5.3.5.1 ATM

As our department has a long-standing interest in the role of ATM, the ATM:c.2119T>C (p.S707P) variant, which was detected in 3 individuals by MPS (and confirmed by Sanger sequencing in 2 individuals; **Figure 5.7**), was selected for further analysis. Despite this variant being predicted to be benign by *in silico* analysis, it has previously been reported to be found at an increased incidence (5 times greater) in individuals with breast cancer compared to the standard population frequency (Dork *et al.*, 2001). To date no functional work has yet been carried out to determine the effect of this variant on normal ATM function.

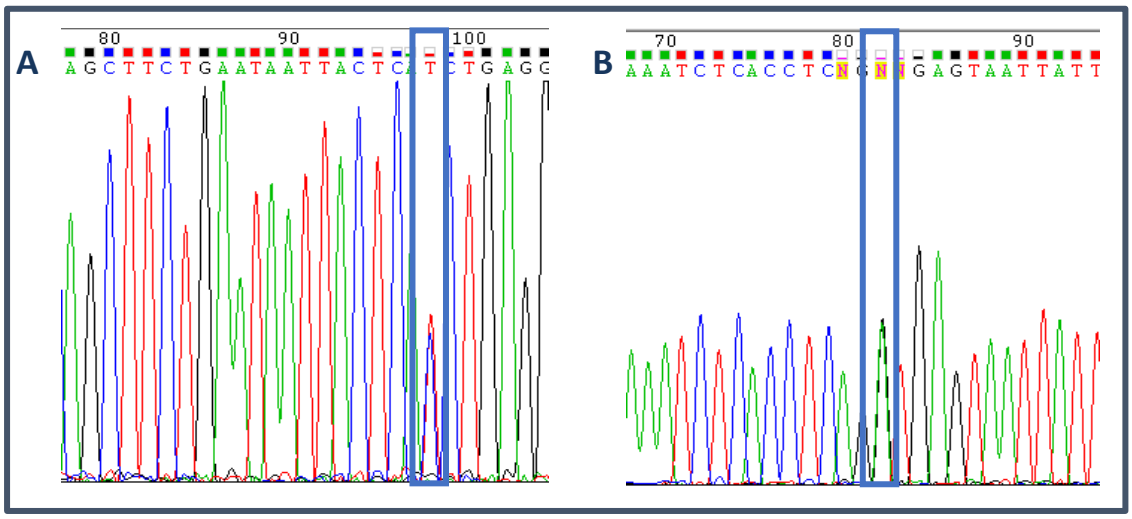


Figure 5.7: Chromatogram traces for confirmation of heterozygous ATM:c.2119T>C in two individuals. Blue box indicates the variant of interest **A.** SABC124, MPS had 564X coverage with the variant present in 49 % of reads, Sanger sequencing was carried out in the forward direction. **B.**SABC038, MPS had 35X coverage with the variant present in 60 % of reads, Sanger sequencing was carried out in the reverse direction.

5.3.5.2 HMMR

The HMMR: c.383C>G (p.S129C) missense variant was identified in 4 individuals in the patient cohort. This polymorphism was not present within COSMIC, HGMD or ClinVar. This variant was present within gnomAD, however was observed at an increased frequency within this patient cohort (4/132 individuals) than the general population (MAF=1.026%). Furthermore, it was predicted to be pathogenic/damaging by all *in silico* analyses carried out. Sanger sequencing was carried out to confirm the presence of this variant within all individuals (**Figure 5.8**).

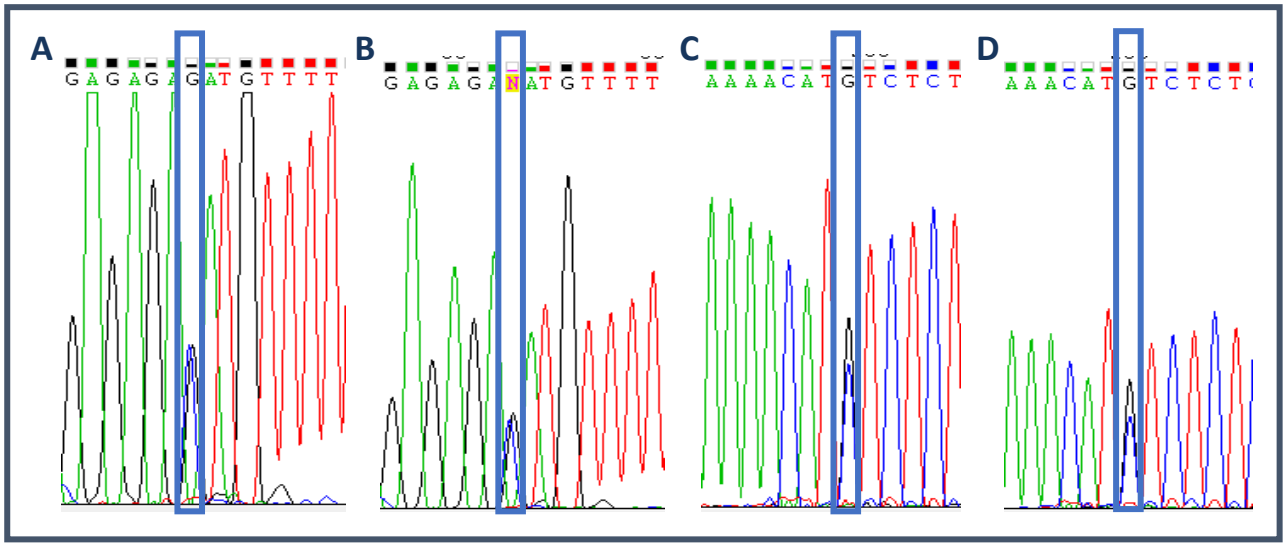


Figure 5.8: Chromatogram traces for confirmation of heterozygous HMMR:c.383C>G in four individuals. Blue box indicates the variant of interest**. A.** SABC053, MPS had 40X coverage with the variant present in 60 % of reads, Sanger sequencing was carried out in the forward direction. **B.** SABC105, MPS had 42X coverage with the variant present in 59 % of reads, Sanger sequencing was carried out in the forward direction. **C.** SABC077, MPS had 128X coverage with the variant present in 52 % of reads, Sanger sequencing was carried out in the reverse direction. **D.** SABC099, MPS had 30X coverage with the variant present in 57 % of reads, Sanger sequencing was carried out in the reverse direction.

This variant results in a change from serine to cysteine at amino acid 129 in the protein sequence. As illustrated in **Figure 5.9**, this region is highly conserved within all species included in the PolyPhen-2 analysis. From protein domain analysis, this variant was found to lie within a highly conserved Chromosome segregation ATPase domain.

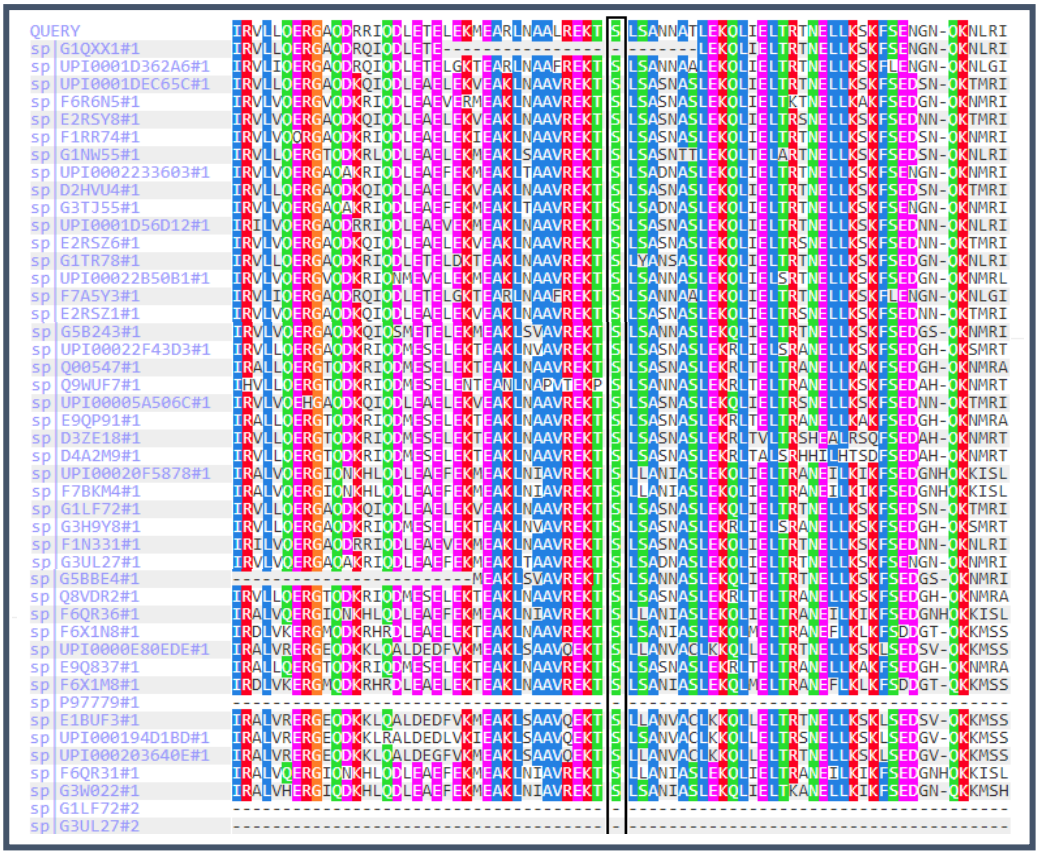


Figure 5.9: Multiple sequence alignment of amino acid sequences for multiple species analysing level of conservation for HMMR p.S129C. Variant amino acid is indicated by the black rectangle.

5.3.5.3 UIMC1

The *UIMC1:*c.1690T>C SNP results in a change from tyrosine to histidine at amino acid 564 of the full-length protein, was detected in 2 individuals and was not present within dbSNP, ClinVar, COSMIC or HGMD. The presence of this variant was confirmed in both individuals by Sanger sequencing (**Figure 5.10**). The frequency at which this variant was observed in the population screened within this study was statistically significant in comparison to the observed frequency in gnomAD (p=0.0257, **Table 5.3**). This variant was predicted to be pathogenic by all *in silico* analyses. Analysis of the protein structure indicated that this variant is present within a highly conserved zinc finger domain of the protein (**Figure 5.11**). PolyPhen-2 analysis illustrated that this amino acid change occurs within a highly conserved region of the protein (**Figure 5.12**).

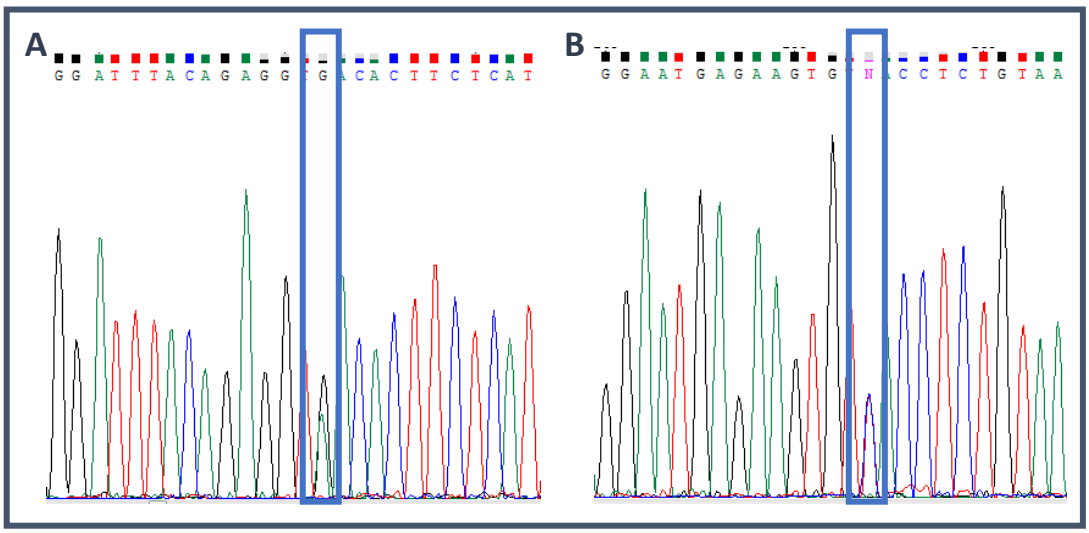


Figure 5.10: Chromatogram traces for confirmation of heterozygous UIMC1: c.1690T>C in two individuals Blue box indicates the variant of interest. **A.** SABC007, MPS had 461X coverage with the variant present in 52 % of reads, Sanger sequencing was carried out in the forward direction. **B.** SABC013, MPS had 201X coverage with the variant present in 47 % of reads, Sanger sequencing was carried out in the reverse direction.

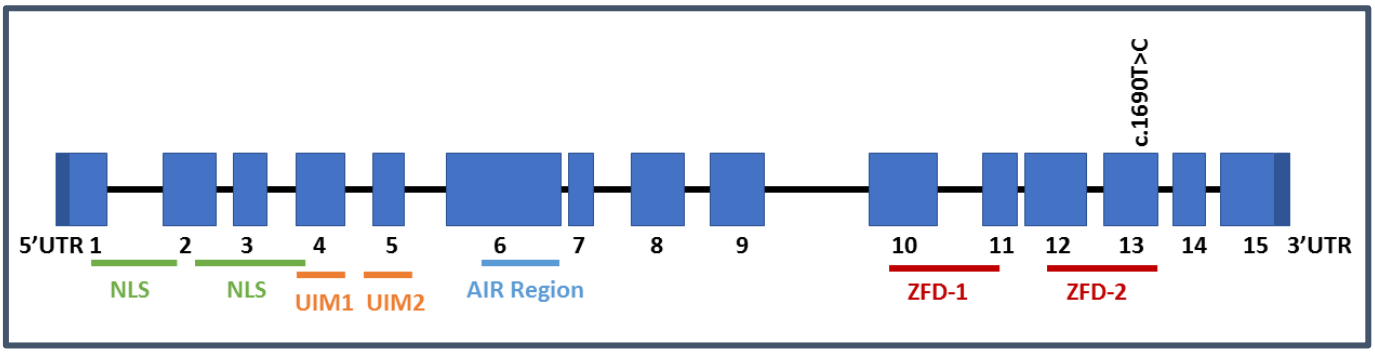


Figure 5.11: Gene structure of UIMC1 including functional domain, interacting proteins and variant of interest. Contains 15 exons, with exon 6 being the largest. Exons are indicated by light blue boxes with untranslated regions (UTRs) shown in dark blue. Identified UIMC1:c.1690T>C potentially significant variant illustrated above exon 13. Exon numbers indicated under corresponding exons. Functional domains of UIMC1 are shown under the exons. NLS; nuclear localisation signal, UIM; ubiquitin interacting motif, ZFD; Zinc finger domains

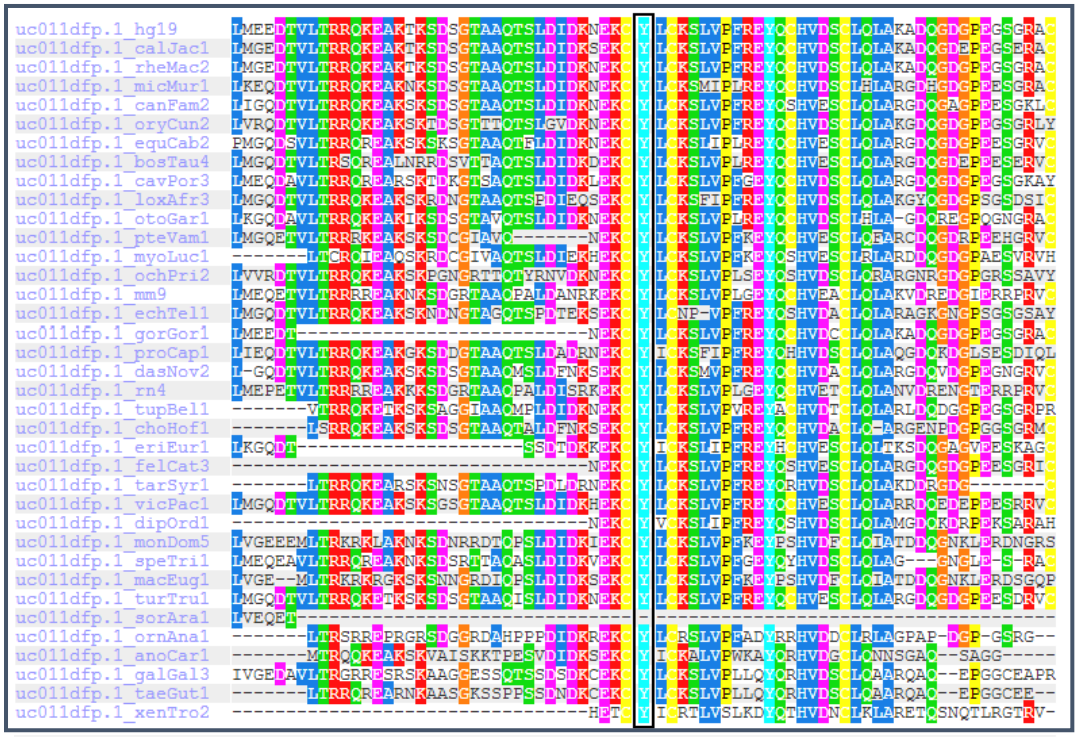


Figure 5.12: Multiple sequence alignment of amino acid sequences for multiple species analysing level of conservation for UIMC1 p.Y564H. Variant amino acid is indicated by the black rectangle.

5.4 Discussion

Through the targeted MPS sequencing of 51 genes of interest in 131 individuals, a total of 1041 unique variants were identified. Application of the bespoke bioinformatics pipeline developed in **Chapter 3** resulted in this list being narrowed down to 166 potentially pathogenic rare variants. Further analysis using functional effect prediction programs resulted in the identification of 82 potentially pathogenic variants within 84 individuals in this study. As the role of the majority of these variants have not been implicated in disease pathogenesis, further work is required to understand their functional significance. Due to the infeasibility of carrying out functional analyses on all of these identified variants, several variants became the focus of the remainder of this project.

5.4.1 MPS aproaches: genome, exome and gene panel sequencing.

There are several options for MPS approaches, ranging from whole genome sequencing to exome sequencing to specific targeted gene panels. At the commencement of this study in 2014, high throughput sequencing was a relatively costly process, and as a generalization, as the size of the regions selected for sequencing increased, so did the associated cost.

Whilst whole genome sequencing (WGS) approaches can identify all possible mutations, this is often associated with a significantly higher cost (approximately $4000 per genome in 2014) and laborious data analysis. In addition, much of the data obtained from whole genome sequencing falls within intergenic regions and therefore is clinically uninterpretable, and as a result has limited utility. Whole exome sequencing (WES) and targeted gene panel sequencing approaches provide a more cost-effective option for determining phenotype-associated mutations. WES allows for the analysis of a large proportion of all protein-coding regions of the genome (approximately 70% at the commencement of this study) and is a useful tool when opting for a discovery-based approach (Ku *et al.*, 2016). At the initial stages of this study, the cost of sequencing a single whole exome was well above $1000 and was not offered in the Flinders genomics Facility until 2015. As a result, carrying out WGS or WES were not financially viable options for this study.

This research study commenced with the hypothesis that proteins which function in the same pathways as *BRCA1* and *BRCA2* may also drive breast cancer when mutated. As the functional roles of *BRCA1* and *BRCA2* are well studied, this hypothesis-driven approach enabled the generation of a targeted sequencing panel focusing on these genes, as opposed to an unbiased genome-wide

discovery approach. The decision to sequence only a select panel of genes not only increased the proportion of interpretable sequencing data, but the smaller target size and associated reduction in costs also made it feasible to increase the number of samples that could be analyzed.

5.1.1 The utility of the Manchester scoring system (MSS)

The patient cohort had Manchester scores ranging from 5 to 53. A spread of Manchester scores was optimal for this study as it was possible that individuals may carry low to mid-penetrance breast cancer susceptibility alleles, in addition to the possibility of previously undetected *BRCA1/2* mutations or variants in high-penetrance susceptibility genes. In addition, limiting the study only to individuals with high Manchester scores would have reduced the likelihood of identifying pathogenic variants in the discovery gene set, as these genes are most likely going to exert low to mid-penetrance effects. This is because it is unlikely that additional high-penetrance breast cancer susceptibility genes remain undiscovered, given the enormous sequencing efforts that have been undertaken in this area recently (Ghoussaini *et al.*, 2012, Michailidou *et al.*, 2015, Michailidou *et al.*, 2017, Momozawa *et al.*, 2018). There are multiple other models available which calculate the probability of identifying a BRCA mutation, however the Manchester scoring system (MSS) is the most widely used and routinely re-calibrated method (Evans *et al.*, 2017). For this reason, it is currently utilised by SA Pathology to determine the suitability of screening breast-cancer individuals for mutations within *BRCA1/2*.

Manchester scores are calculated from multiple factors including the type, number and age at diagnosis of cancers observed in the family (Kast *et al.*, 2014). Although this system is designed to estimate the chance of identifying a mutation within the highly penetrant *BRCA1/2* genes. Therefore, it was determined that a range of Manchester scores would be utilised for this study, as individuals with mid-range scores may be more likely to contain a pathogenic mutation within one of the mid- to low-penetrance known susceptibility genes included on the panel, or within one of the proposed susceptibility genes not yet implicated in the development of breast cancer.

It is important to note that the MSS is only a guide, as individuals with low Manchester scores have been found to harbour pathogenic mutations within *BRCA1/2* (Farra *et al.*, 2019). From the longitudinal study included within this analysis, 12 individuals with pathogenic *BRCA1/2* mutations were analysed. The Manchester scores of these individuals ranged from 9-54, with 8 falling below

the recommended cut-off of 20 (Evans *et al.*, 2004). This suggests that the arbitrary cut-off of 20 may be too high, and individuals with lower scores should also be offered testing (as currently performed by the South Australian Familial Cancer Service). These low scores could be attributed to small family size, or lack of known familial history about cancers within the family. As the Manchester score is determined based on the number of cancers within the family and the age of onset, small family sizes or lack of knowledge pertaining to familial cancer history will result in a lower score.

Conversely, individuals with high Manchester scores have been found to be *BRCA1/2* mutation-negative. Individuals included in this study had Manchester scores ranging from 5-61; with a majority of the samples falling below the recommended score for genetic testing. Studies comparing the most commonly used risk prediction models have found that the MSS illustrates lower discriminatory accuracy in comparison to other models (Amir *et al.*, 2010). Large scale studies comparing the performance of the most commonly used models have found both BOADICEA and BRCAPRO outperform the MSS, particularly in regards to families with a low predicted risk (Antoniou *et al.*, 2008). From these studies, it has become apparent that all of these empirical models, both the MSS and the newer algorithms, tend to under-predict the number of mutations in families, especially in families with missing information on cancer diagnoses. Therefore, an ambiguity on age of diagnoses or type of cancer will have detrimental effects to accuracy in the prediction of cancer risk and/or determination of score (Antoniou *et al.*, 2008).

The MSS is an easy to use method, with probabilities generated within minutes, while the computer-based algorithms can take 20-30 minutes for data input and analysis, which can often lead to lengthy clinic visits. However, with diagnostic laboratories switching to a panel-based approach and screening additional cancer predisposition genes, it may be beneficial to switch to a model that determines cancer risk for not only *BRCA1/2*, but also other breast cancer susceptibility genes, and takes a wider range of risk factors into consideration. The BOADICEA model combines complete family history, genetic and lifestyle risk factors in a single model to provide a comprehensive approach to cancer risk prediction. Studies have illustrated that the BOADICEA outperforms the MSS in sensitivity and specificity (Antoniou *et al.*, 2008), accurately predicting the number of mutations within individuals referred for genetic screening. Furthermore, this model provides the best discrimination between mutation carriers and non-carriers in comparison to all other models

(Antoniou and Easton, 2006). This algorithm considers the likelihood of mutations within susceptibility genes other than *BRCA1* and *BRCA2* including *ATM*, *CHEK2* and *PALB2* (Lee *et al.*, 2019). As the number of cancer predisposition genes increases, it is not only necessary to screen these well documented susceptibility genes but take them into consideration when calculating the risk of identifying mutational status of individuals. It is possible to modify the MSS, as it has been recalibrated multiple times since its first inception, to include pathology and biomarker information (Evans *et al.*, 2004, Evans *et al.*, 2005, Evans *et al.*, 2009, Evans *et al.*, 2017). Therefore, it may be necessary to recalibrate the Manchester scoring system to include mutational status for these additional cancer susceptibility genes.

There was one individual included within this patient cohort that is an example of a clinical situation that would have benefited from using a more informative scoring system, such as BOADACEA. Patient SABC002 was initially selected for analysis based on their high Manchester score (53) and referred for screening in 2005. A pathogenic mutation was not identified. In 2014 an immediate family member of this individual was diagnosed with breast cancer and was subsequently screened by SA Pathology. This individual was found to harbour a pathogenic mutation within *BRCA1*, however patient SABC002 did not possess the same pathogenic mutation. Therefore, it is most likely that the initial individuals' breast cancer was due to sporadic causes and could not be attributed to the germline mutation identified in this family. As 90% of breast cancer cases are sporadic, it is possible that this individual was unfortunate enough to develop a sporadic case of cancer despite their familial history.

5.4.2 Patient cohort selected for sequencing analysis

From the screening analysis carried out on the selected patient cohort of 133 individuals, 131 patients successfully generated sequence. Additionally, throughout the longitudinal study of all individuals referred for *BRCA1/2* testing between November 2011 and October 2012, 11 out of 80 (15%) individuals screened were *BRCA1/2* mutation positive, which is in line with the 20% identified within the literature (Turnbull and Rahman, 2008, Shiovitz and Korde, 2015). Therefore, this cohort can be considered representative of the wider breast cancer population and as a result, should be informative for finding genes that have a wider applicability beyond this genetic population.

5.4.3 Bioinformatic analysis of sequencing data

Using the bioinformatic pipeline established in **Chapter 3,** a large number of SNPs and indels were identified within each individual (mean: 124, minimum: 65, maximum: 168). A third of these variants were common within the general population and were filtered out through dbSNP and gnomAD. As previously discussed, a MAF of <5% was required for variants to be analysed further. This cut-off was selected based on the literature, as previous other studies have utilised a MAF cut-off of <1% or <5%. Although this could be considered high in comparison to other studies (Damiola *et al.*, 2014, Young *et al.*, 2016, Kobayashi *et al.*, 2017) it was decided to err on the side of caution and analyse more variants in detail as the role of the majority of the sequenced genes in cancer predisposition and pathogenesis has not been established.

Additionally, when considering the MAF provided in gnomAD, the total MAF provided was used, rather than a specific allele frequency associated with ethnicity. This is due to the fact that the nationality of all individuals screened within this study was unknown, and Australia is a multicultural country. Due to this, and the lack of Australian allele frequency data within the gnomAD database, it was not possible to use one specific population for this analysis.

5.4.3.1 Assessment of potentially pathogenic variants

The functional consequences of sequence variants are often difficult to predict. As a large number of variants of low frequency were detected within each sequenced library, it was not feasible to carry out in-depth analysis of each individual variant identified. Therefore, a targeted approach aimed at identifying the variants most likely to be clinically relevant was carried out. There are a multitude of commercially available platforms which can be utilised for the analysis of sequencing data, each with their own associated cost. These programs allow for the automated analysis of sequencing data, including the majority of the *in silico* analyses included in this study (such as Ingenuity Variant Analysis). However, due to the cost associated with the purchase of a licence and cost per sample, this was deemed prohibitively expensive. This resulted in the need for manual analysis of variants within each of the listed databases and *in silico* analysis programs

From this analysis, a large proportion (65%) of these variants were predicted to be benign or had previously been illustrated to not be involved in the development of cancer. These public repositories are a useful tool for the analysis of sequence variants, especially when such large

numbers of variants of uncertain significance are identified within each sequenced individual. Of note is that through this analysis, 7 variants were predicted to be pathogenic by all approaches, however literature searches and ClinVar illustrated that functional validation of these variants had illustrated that they were not involved in the development of cancer. This illustrates that despite advances in the *in silico* prediction programs, there is still a need for confirmation of variant pathogenicity with functional data.

From the analysis of the low frequency variants with *in silico* programs and database searches, 82 variants (identified in 84 individuals) were predicted to be pathogenic. The variants selected for further analysis were predicted to be pathogenic by 3 out of 4 *in silico* programs utilised, or if they had previously reported conflicting interpretations of pathogenicity. Through the analysis carried out, it was observed that Align GVGD often did not align with the results obtained from the other analysis programs. Often it determined variants as being of least concern when predicted to be damaging by all other analyses (For example, NBN:c.2165G>C in **Table 5.2**). This analysis was carried out for all variants identified, however it has only been optimised for a limited number of genes on extensively sequenced genes (Tavtigian *et al.*, 2008a, Fortuno *et al.*, 2018). Therefore, the results that are determined for most of the genes included on this panel are not stringently optimised. Additionally, it is important to note that PolyPhen-2 is designed for the analysis of non-synonymous polymorphisms, and as such is not beneficial for the analysis of nonsense and frameshift mutations that have been identified within this study.

There is a range of programs that can be utilised for the assessment of low frequency missense variants, with new programs constantly being released, such as CADD, REVEL, and MutationTaster (Zhang *et al.*, 2018b). However, as these programs were not available at the commencement of this study in 2014, they were not included in the analysis pipeline. The pipeline utilised was based on a range of published studies conducting similar research at the time, in addition to comprehensive review articles (Duzkale *et al.*, 2013, McCarthy *et al.*, 2013, Thompson *et al.*, 2013a, Damiola *et al.*, 2014). As a result, there are multiple programs which are now routinely used for predicting pathogenic variants that have not been employed in this study. It would be interesting to carry out variant analysis on this sequencing data with these updated prediction tools, to determine if the predictive ability of these programs has significantly improved in the past 5 years, however this is beyond the scope of this study.

5.4.3.2 Identification of rare variants found in both *BRCA1/2* mutation-positive and mutation-negative individuals.

In order to eliminate variants which are less likely to be causative of breast cancer in this cohort, the sequencing results obtained for the *BRCA1/2* mutation-positive individuals were compared to those obtained for the *BRCA* mutation-negative individuals. This comparison identified 5 variants that were present in both cohorts. The rationale behind this comparison was that there is less evidence that these mutations may be causative of the breast cancer in the *BRCA1/2* mutation-negative individuals if they are also present within *BRCA1/2* mutation-positive individuals. The 5 variants identified in both cohorts included EP300:c.6668A>C, EP300:c.6983C>T, NQO2c.86A>G, NQO2:c.173G>A and PRKDC:c.8694C>T. With the exception of EP300:c.6668A>C, these variants were all present in multiple individuals and had a MAF >2%, which is similar to that observed in gnomAD. This further suggests that these variants may be rare, normal sequence polymorphisms, rather than causative pathogenic mutations.

Whilst it cannot be ruled out that these variants may be involved in the multiplicative effect of low susceptibility polymorphisms that result in hereditary breast cancer through a polygenic model, this study was not sufficiently powered to detect such effects. Therefore, these variants were not considered for further functional analysis.

5.4.3.3 Patients with no predicted pathogenic mutations

From the patient cohort included in this analysis, no predicted pathogenic mutations were identified in 36 individuals. This does not mean that their cancer is not attributed to a hereditary component, as there are additional genetic mechanisms that could be responsible for the development of their cancer which have not been analysed in this study. One possible cause is the presence of pathogenic mutations in *BRCA1* or *BRCA2* which were undetected in this screening method. Promoter mutations, 5' and 3' UTR mutations and deep intronic mutations affecting splicing would not be detected with this MPS targeted gene panel. It has been shown that promoter regions harbour functional mutations at similar frequencies to coding sequences (Rheinbay *et al.*, 2017), but are often undetected due to their presence in GC-rich sequences. These sequences are not only difficult to sequence with standard MPS approaches (Huppert and Balasubramanian, 2006, Wang *et al.*, 2011), but the downstream functional effect of any sequence changes is also difficult to predict without comprehensive functional analysis. It should be noted that although the promoter sequences of the genes were included in the sequencing panel, no pathogenic promoter mutations

were identified from the pipeline used. Additionally, mutations within UTRs may affect microRNA (miRNA) binding, either resulting in repression or over expression (Shen *et al.*, 2008, Chang and Sharan, 2012, Li *et al.*, 2012), or affect mRNA stability and the ability of mRNA transcripts to load onto ribosomes, affecting the downstream function of the produced transcript. Studies have identified that *BRCA1* and *BRCA2* are the target of over 100 miRNAs (Chang and Sharan, 2012). miRNAs have been shown to downregulate and even silence *BRCA1*, interrupting cellular processes such as DNA damage repair and cell cycle checkpoint control (Petrovic *et al.*, 2017). Three UTR mutations were identified in *BRCA1* in this study and were included in the list of 82 potentially pathogenic variants found in the patient cohort. The functional significance of these variants has not yet been investigated, however these variants may have the potential to affect *BRCA1* expression and transcript stability which could play a role in cancer predisposition.

In addition to undetected *BRCA1/2* mutations, it is also very possible that these individuals may have inherited mutations within genes that have not been included in this study. This could include, but is not limited to, genes that are also known to result in the development of a syndrome, and as a result were excluded from this panel (e.g. *PTEN* and Cowden syndrome; *MLH1, MSH2, MSH6* and Lynch syndrome). These individuals may benefit from WES or WGS for a more detailed analyses of their genetic makeup.

Furthermore, these cancers may be sporadic cases of cancer attributed to environmental factors, especially in individuals with very low Manchester scores. As previously stated, 90% of breast cancer cases are thought to be sporadic in nature. Therefore, in those individuals with very low Manchester scores, it may be that there is not a hereditary component to their cancer. From the individuals included within this patient cohort, 58% (78/133) had Manchester scores less than 20, with 19% having Manchester score ≤ 10 (25/133). There are a multitude of environmental factors and lifestyle risks that have been shown to result in an increased incidence of cancer, and these may be the cause of cancer within these individuals. As the development of cancer is a complex phenomenon, there are many mechanisms which could be responsible for cancer observed in these mutation-negative individuals, and while it is not feasible to look at all these mechanisms within one study, it is necessary to understand the limitations associated with gene panel screening.

5.4.4 Confirmation of variants by Sanger sequencing.

Validation of identified sequence variants via Sanger sequencing was carried out for the first two sequencing runs. Through this, most sequence variants were confirmed (52/54 variants) in 65 individuals. There were two variants that were not confirmed from the initial sequencing run, which were both attributed to errors associated with Ion Torrent sequencing. The two variants that were not verified were identified in the first sequencing run and were mainly identified due to user inexperience and a lack of familiarity with the sequencing issues associated with Ion Torrent sequencing. The RFC4:c.35--36TA>CT variant was identified in 2 patients in the initial sequencing run. This variant was predicted to result in a frameshift nonsense mutation, resulting in premature termination of the RFC4 protein. This variant was present following a stretch of homopolymers including 4 guanine nucleotides followed by 4 thymine nucleotides. One of the aforementioned common sequencing errors associated with Ion Torrent MPS is issues with correct base determination in stretches of homopolymer bases, which is attributed to its terminator free chemistry (Ross *et al.*, 2013). This often results in incorrect incorporation of the incorrect number of bases, with approximately 40% of the reads missing one of the G or T nucleotides within the homopolymer stretch, resulting in what appeared to be a frameshift in these individuals. However, confirmation by Sanger sequencing identified that these were false-positive variants within these individuals. The second false positive variant identified was SLC19A1:c.522delG. This frameshift mutation was identified in one individual in the initial sequencing run but was not confirmed by Sanger sequencing. This variant is present within a highly GC-rich region (67.1%), which are notoriously challenging to sequence, often with lower quality of sequencing and higher levels of background noise (Yohe and Thyagarajan, 2017). From the knowledge gained from analysis of subsequent sequencing runs, developing familiarity with the sequencing variants that appear in multiple runs and learning the issues associated with the sequencing chemistry, these variants would have been identified as most likely to be false positives.

5.4.5 Issues associated with variants of uncertain significance

The largest issue associated with these types of MPS studies is the large number of variants identified within each patient, many of which have unknown clinical significance. From the patient cohort analysed in this study, each individual had an average of 27 rare variants identified. Through the analysis pipeline depicted in **Figure 5.2,** this was narrowed down to a list of 166 low-frequency variants were predicted to be pathogenic. Further analysis condensed this list to 82 potentially

pathogenic variants or variants of uncertain significance which required further validation. Despite in-depth functional analysis, these variants remain in genetic purgatory, as their lack of clinical application means that there is no benefit to the affected individuals. In order to report identified mutations back to individuals, it is necessary for their role in cancer to be clearly demonstrated and as such, VUS must be functionally validated. New gene editing technologies such as CRISPR/Cas9 (Doudna and Charpentier, 2014) (Hsu *et al.*, 2014) and base editing (Gaudelli *et al.*, 2017) are revolutionising this field, however as it is still in its infancy, this is still a time-consuming process which is often wrought with complications (refer to **Chapter 6**). As only clinically actionable variants can be reported back to individuals, this process is not able to be carried out for the vast majority of VUS identified within samples screened.

5.4.6 Variants of interest identified within the patient cohort.

5.4.6.1 Pathogenic *PALB2* mutation.

A heterozygous *PALB2* deletion (PALB2:c.3119delA) was identified in individuals SABC042 and SABC025. This variant was predicted to be pathogenic by all *in silico* analyses carried out, in addition to being listed in both the COSMIC database and HGMD. Literature searches revealed that this deletion had been shown to result in a frameshift, resulting in premature protein truncation (Rahman *et al.*, 2007, Antoniou *et al.*, 2014). This deletion is located within the BRCA2 binding domain of the PALB2 protein and affects binding and localisation of BRCA2 to sites of DNA damage. As this variant is documented as pathogenic within the literature, this information was reported to SA Pathology, which was in turn passed onto these individuals and their families with appropriate counselling. Importantly, this discovery allows for cascade testing of the immediate family members and more vigilant monitoring for those that harbour this pathogenic *PALB2* deletion.

Interestingly, these individuals had Manchester scores of 34 and 15, with the latter falling below the arbitrary cut-off for BRCA screening as recommended in the literature (Evans *et al.*, 2004). If this cut-off was strictly adhered to within the SA Familial Cancer service, this pathogenic mutation would have been missed in this family. This emphasises the need to either lower the recommended value for screening, switch to a different screening model or modify the MSS to include the likelihood of mutations within other breast cancer susceptibility genes.

5.4.6.2 UIMC1

The missense polymorphism UIMC1:c.1690T>C was identified in two patients, SABC007 and SABC013. This variant was not listed in HGMD or COSMIC, nor was it present within the literature. This heterozygous variant was of low frequency within the general population as determined by gnomAD (0.207%). A Z-test indicated the frequency of this variant was significantly increased this patient cohort in comparison to the reference population (p=0.0257, **Table 5.3**). This sequence variant was predicted to be damaging by all analysis methods used and lies within a highly conserved zinc finger domain. This zinc finger domain has multiple functions including a role in DNA recognition, which is necessary for its protein-protein interactions in addition to its nuclear localisation (Yan *et al.*, 2002). Importantly, this protein is known to form a protein complex with BRCA1 for repair of DNA damage (Wang *et al.*, 2007), in addition to recruiting BRCA1 and specific ubiquitin structures to the sites of DNA damage (Sobhian *et al.*, 2007). Additionally, a deletion of a single amino acid within the UIM domain of *UICM1* has been illustrated to result in reduced capacity to repair DNA DSBS, leading to a significant increase in chromosomal abnormalities (Nikkilä *et al.*, 2009). Whilst this mutation was identified within a different functional domain, the significance of mutations within the remaining protein have not been investigated. Therefore, it is biologically feasible that a mutation within this zinc finger domain may have a detrimental effect on DNA damage repair and cell cycle checkpoint control. Based on this information, this variant warrants further functional research.

5.4.6.3 ATM

The variant ATM:c.2119T>C was identified in the heterozygous state in 3 patient samples (SABC023, SABC038 and SABC124). This variant was found to be listed in HGMD and COSMIC and is of a low frequency within the general population. Analysis of the polymorphism using *in silico* prediction programs showed that the sequence variant occurred within an area of low conservation across multiple sequence alignments, is not located within any functional protein domains, and was predicted to be benign or tolerated by all analyses. Despite these results, this missense variant has been shown to be 5-times more prevalent in individuals with breast cancer than in the general population (Dork *et al.*, 2001). Furthermore, this variant has been described as a predisposition mutation in both somatic and inherited breast cancers (Fletcher *et al.*, 2010). The large scale analysis carried out by Fletcher *et al.* (2010) analysed the frequency of several SNPs in ATM in 26,101 breast cancer cases and 29,842 controls. This study found that ATM:c.2119T>C along with 4 other SNPs in

ATM, can explain a small proportion of familial cancer risk. To date, no further large-scale studies genotyping ATM in breast cancer cases and controls has been carried out, indicating that further work is required to understand the role of this variant in breast cancer development.

Whilst there is circumstantial evidence to suggest that this ATM variant may be involved in breast cancer predisposition, to date there has been no functional validation carried out on this variant. The department of Molecular Medicine and Pathology has a long-standing interest in ATM and has well-established wet-lab assays to determine ATM functionality. Therefore, this variant was selected as one of potential future interest for functional studies.

5.4.6.4 HMMR

Mutations in HMMR have previously been associated with the development of inherited breast cancer (Pujana *et al.*, 2007). Previous work has demonstrated that HMMR associates in protein complexes with BRCA1 and BRCA2 to control centrosome number and chromosome segregation. Furthermore, HMMR is a substrate for BRCA1-BARD1 mediated polyubiquitination, in addition to the BRCA1-HMMR interaction required for normal cell structure (Pujana *et al.*, 2007).

The heterozygous HMMR:c.383C>G variant was identified and confirmed in 4 individuals (SABC053, SABC077, SABC099 and SABC105). This variant was not present within any of the population frequency or clinical significance databases and was predicted to be pathogenic by all *in silico* analysis programs used. This variant lies within a highly conserved chromosome segregation ATPase domain, which is required for accurate replication and segregation of chromosomes during cellular division. However, there has been loose support for the role of HMMR in the development of cancer, with several studies illustrating cumulative effects from an *HMMR* mutation in individuals with *BRCA1* mutations (Maxwell *et al.*, 2011), whilst others have not found any support for *HMMR* as causative gene in hereditary cancer (Kalmyrzaev *et al.*, 2008). Furthermore, studies have indicated that perturbed HMMR function may be implicated in sporadic breast cancer, due to the dysregulation of normal cell growth and motility (Maxwell *et al.*, 2011). Due to the conflicting evidence surrounding the involvement of this gene in cancer development, the number of individuals with this VUS in the cohort, and the lack of presence in any databases, this variant is a good candidate for functional validation.

5.4.7 Conclusions

In summary, through the targeted sequencing of 51 genes in 131 individuals, a large number of variants of low frequency were identified. Multiple *in silico* programs were used to identify the clinical significance of these variants, however many variants of uncertain significance remained. Whilst the evidence of pathogenicity of some variants was more compelling than others, it is still not feasible to functionally validate the large number of mutations identified. Therefore, a small subset of variants was selected for functional analysis.

Initially, the *UIMC1, HMMR* and *ATM* missense variants were all selected for functional validation, however due to time constraints, functional validation was ultimately limited to *UIMC1* (Refer to **Chapter 6**)**.** The predicted pathogenic nature of this variant, in conjunction with the function of UIMC1 in normal cellular processes indicates that a loss of normal function could be associated with the development of inherited cancer and requires functional analysis through cellular models for further understanding.

Chapter 6:
Functional validation of predicted pathogenic *UIMC1* variant

6.1 Introduction

As described in **Chapter 5**, it is clear that there is a need to functionally validate predicted pathogenic variants to determine their role in cellular function and in the development of cancer. Two approaches were selected for functional validation in this study; mammalian expression plasmids and CRISPR/Cas9 gene editing. As the process of functional validation is both time- and resource-intensive, only one variant was selected for functional analysis, UIMC1 c.1690T>C.

6.1.1 UIMC1

UIMC1 (also known as RAP80) is a central component of the BRCA1-A complex along with Abraxas and BRCA1 (Wang *et al.*, 2007). This complex is required for regulating DNA damage repair and cell cycle checkpoint control within the cell nucleus. UIMC1 contains several ubiquitin interaction motifs (UIM), which interact with ubiquitinated proteins at the sites of DNA damage (Sobhian *et al.*, 2007). Furthermore, UIMC1 is also required for the recruitment of the BRCA1-Abraxas complex to the site of DNA damage, where it further ubiquitinates additional proteins and is speculated to amplify ubiquitination within the damaged region (Wang *et al.*, 2007). As ubiquitination is a central mechanism in the DNA damage response pathway, it is hypothesised that UIMC1 plays a pivotal role to the maintenance of cellular integrity. Mutations within this DNA damage response pathway are often critical events in carcinogenesis, and UIMC1 is a key member of this pathway (Ali *et al.*, 2017, Jin *et al.*, 2019), with numerous studies showing that UIMC1 is integral for the accumulation of BRCA1 at sites of DNA damage (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a).

In addition to the involvement of UIMC1 in DNA damage repair, this variant was selected for functional validation because it was identified in two individuals in this patient cohort. Whilst this could support the hypothesis that this variant is associated with breast cancer susceptibility, as this study was carried out on a small population within South Australia, it is also possible that these individuals were related, and this variant is merely a rare polymorphism unique to this family. The potential segregation of this variant was assessed through analysis of linked polymorphic STS markers.

6.1.2 CRISPR/Cas9

Revolutionising the field of genome engineering, the CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR associated 9) system is a relatively recent development in molecular biology which can be utilised for the precise editing of mammalian genomes (Jinek *et al.*,

2012, Cong *et al.*, 2013). Initially discovered as an adaptive immune response by bacteria and Archaea, (Ishino *et al.*, 1987), CRISPR-containing organisms acquire DNA fragments from invading bacteriophages and plasmids before transcribing them into CRISPR RNAs (crRNAs). This results in the generation of a sequence-specific fragment which is utilised for future resistance against infection, using the crRNAs as a guide to direct cleavage of complementary invading DNA through the nuclease activity of the Cas protein also encoded by the CRISPR loci. (Ishino *et al.*, 1987, Marraffini and Sontheimer, 2008, Hsu *et al.*, 2014, van der Oost *et al.*, 2014) The complex CRISPR immune system functions through the cooperation of numerous diverse Cas proteins, which have been divided into 2 major classes based on their mechanisms of action and their composition. Class I systems involve RNA-guided target cleavage through a large complex of several effector proteins (types I, III and IV), whereas class 2 systems (type II) only require one RNA-guided endonuclease for cleavage (i.e. Cas9 in type II) (Makarova *et al.*, 2015).

6.1.2.1 Functionality of CRISPR/Cas9

The Cas9 nuclease of the type II CRISPR system is the most widely used for genomic editing amongst all Cas proteins. Target cleavage is guided by a duplex of two RNAs; the crRNA that recognises the invading DNA through a 19bp complementary region and the tracrRNA that hybridises with the crRNA and is unique to the type II CRISPR system (Garneau *et al.*, 2010, Jinek *et al.*, 2012). Revolutionary studies have shown that that the Cas9 nuclease, along with the crRNA-tracrRNA duplex can be repurposed for genome editing, with the crRNA-tracrRNA duplex fused into a chimeric single guide RNA (sgRNA) (Jinek *et al.*, 2012, Cong *et al.*, 2013, Jiang *et al.*, 2013). This cas9-sgRNA complex has the ability to bind DNA that complementary base pairs with the sgRNA and is adjacent to a protospacer-adjacent motif (PAM) sequence (**Figure 6.1a**). Upon binding, the Cas9-sgRNA complex induces cleavage 3bp downstream of the PAM sequence. Therefore, Cas9 is easily able to be re-programmed to edit any genomic location containing a PAM sequence through modification of the sgRNA sequence. There is a plethora of Cas9 orthologs that are present within a variety of type II CRISPR systems. The most commonly used Cas9 for genome editing is the CRISPR system adapted from *Streptococcus pyogenes*. The SpCas9 is 1368 AA in length and has a simple PAM sequence of NGG, or a weaker NAG, where N is any nucleotide (Wang *et al.*, 2016).

As illustrated in **Figure 6.1a**, the Cas9 contains two nuclease domains. The HNH domain cleaves the target strand of DNA (which has a sequence complementary to the sgRNA) and the RuvC nuclease domain that cleaves the non-target DNA strand. Inserting a mutation into either of these domains

results in a nickase Cas9 (nCas9) which is only capable of cleaving one strand of DNA (**Figure 6.1b**). For more precise genome editing, pairs of nCas9s are able to be targeted to adjacent DNA sites, resulting in a DSB only if both complexes are present at the target site (Ran *et al.*, 2013a).

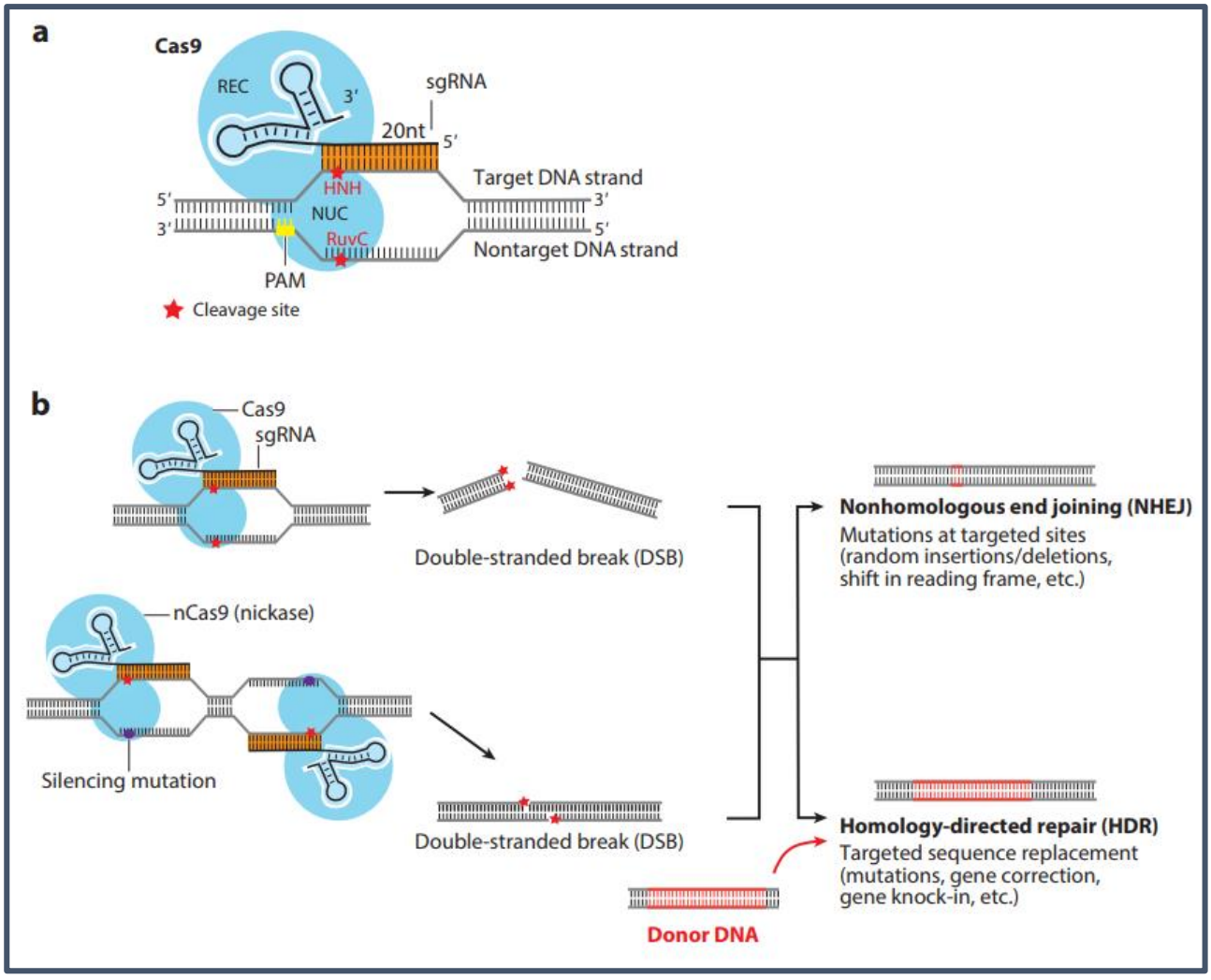


Figure 6.1: CRISPR/Cas9 sequence specific genome editing A. Schematic of the Cas9 nuclease system modified for targeted genomic editing. Recognises target DNA by 20 nucleotide (nt) complementary base-pairing interaction between a sing guide RNA (sgRNA) and the targeted DNA strand. Cas9 also interacts with the protospacer-adjacent motif (PAM) of the DNA target through the PAM-interacting domain at the c-terminus. Cas9 utilises two nuclease domains (HNH and RuvC) to cleave double stranded DNA 3bp downstream of the PAM site, creating a DSB. The Cas9 nuclease lobe (NUC) contains the RuvC, HNH and PI domains, while the recognition love (REC) of Cas9 contains other regions that interact with the sgRNA-DNA duplex. **B. The use of CRISPR/Cas9 in genomic editing.** (Top) The DSB generated by Cas9 activates the non-homologous end joining (NHEJ) or homology directed repair (HDR) DNA repair pathways. NHEJ results in random indels at the target site, whilst HDR can be used for targeted indels or desired mutations through homologous recombination with donor DNA. (Bottom) A mutation within a nuclease domain of Cas9 results in a cas9 based nickase (nCas9) that cleaves only one strand of DNA. The specificity of Cas9 genome editing can be enhanced significantly through using a pair of nCas9s that target each strand of DNA at adjacent sites as both nCas9-sgRNA complexes must be present at the target site for generation of DSBs (Modified from Wang *et al.* (2016)).

6.1.2.2 Genome engineering with CRISPR/Cas9

Since its initial discovery, Cas9 has been used extensively in genome editing via two main processes; DNA cleavage and DNA Repair (**Figure 6.1b**). The sgRNA directs Cas9 to a specific genomic locus, where Cas9 results in a DSB, triggering DNA repair through cellular mechanisms such NHEJ and homology-directed repair (HDR). NHEJ causes random indels at the site of the DSB and may result in gene knockout through causing a shift in the reading frame or mutating a crucial region of the encoded protein. HDR can be utilised to generate the desired sequence replacement at the site of the DSB, through the use of a repair DNA template (Wang *et al.*, 2016). This system has been used in a variety of reverse genetics studies, allowing easy analysis of the role of various genes by selectively disrupting its function with targeted modifications.

Retargeting the Cas9 protein is simple, via the creation of a new sgRNA that pairs with the desired DNA targeting site adjacent to a PAM site (Doudna and Charpentier, 2014, Hsu *et al.*, 2014). In the instance of the *S.pyogenes*, the NGG PAM motif allows it to target, on average, every 8bp within the genome, allowing the modification of almost any gene to be carried out (Cong *et al.*, 2013, Doudna and Charpentier, 2014, Hsu *et al.*, 2014).

Genome engineering with the use of the CRISPR/Cas9 system has become such an incredibly fast-paced field, with laboratories worldwide utilising this technology to further elucidate disease mechanisms. The ability to introduce DSBs and specific mutations at defined positions has made it possible to generate cell lines and primary cells containing deletions and point mutations resembling those described in cancers (Doudna and Charpentier, 2014). This rapid modelling of genetic events also allows for functional analysis of mutations of uncertain significance that are identified though screening studies (Sánchez-Rivera and Jacks, 2015). The CRISPR/Cas9 system enables permanent modification of single or multiple loci through either the stable or transient delivery of the required CRISPR components. Mammalian cell cultures have been edited through transient transfection of plasmid DNA encoding Cas9 and sgRNAs (Cho *et al.*, 2013, Cong *et al.*, 2013, Mali *et al.*, 2013b), or Cas9 ribonucleoprotein complexes (RNPs) (Kim *et al.*, 2014, Lin *et al.*, 2014b). Alternatively, CRISPR components can be delivered via retroviruses or lentiviruses (Malina *et al.*, 2013, Shalem *et al.*, 2014). Loss of function mutations rely on NHEJ, which often results in indels near the Cas9 cleavage site, frequently leading to nonsense mutations. However, the introduction of a gain of function, or specific point mutation requires the inclusion of an HDR template containing the desired mutation. Once generated, cell lines carrying one or more mutations can then be tested using a multitude of

in vitro assays to examine the effect of the mutations on cancer associated phenotypes. Examples of this have successfully been carried out on cancer cell lines (Kuscu *et al.*, 2017), primary cell lines (Xu *et al.*, 2018), patient derived xenographs (Behrmann *et al.*, 2017), organoid cultures (Matano *et al.*, 2015) in addition to animal models and human embryos (Kang *et al.*, 2016).

6.1.3 Aims and hypotheses

As UIMC1 is a key component of the BRCA1-genome surveillance complex and plays a key role in recruiting BRCA1 to the site of DNA damage, it is hypothesised that mutations within this gene which render the protein non-functional will result in an increased susceptibility to the development of breast cancer. It is hypothesised that cells lacking functional *UIMC1* will be unable to, or show a reduced ability to, repair DNA double stranded breaks, in addition to having altered cell proliferation.

The aims of this chapter are to:

1. Determine if the two individuals with the UIMC1:c.1690T>C mutation are related via genetic linkage analysis.
2. Generate CRISPR/Cas9 plasmids for the generation of the potentially pathogenic UIMC1c.1690T>C variant, and for knockout of UIMC1 function in mammalian cell lines.
3. Create *UIMC1*-modified HEK293 and MCF10A cell lines for functional analysis, including proliferation and ability to repair DNA double stranded breaks.

6.2 Methods

6.2.1 Analysis of microsatellite repeats

The STS marker tool on the UCSC genome browser (https://genome.ucsc.edu/) was utilised to select 4 di- or tri-nucleotide repeats to be used for linkage analysis (**Figure 6.2**). Synthesised primers were provided lyophilised (Integrated DNA Technologies (IDT), Singapore) and were subsequently resuspended in sterile water at a final concentration of 100µM and stored at -20°C. Marker sequences are listed in **Appendix F.** Markers were optimised using standard PCR conditions (**Appendix G**). Once optimised, one oligonucleotide was replaced with a fluorescently labelled (FAM or HEX) version. Samples were amplified and visualised using gel electrophoresis to confirm amplification of a single amplicon. Samples were diluted 1:50 and sent to the SA Pathology DNA Sequencing Facility for Fragment Analysis.

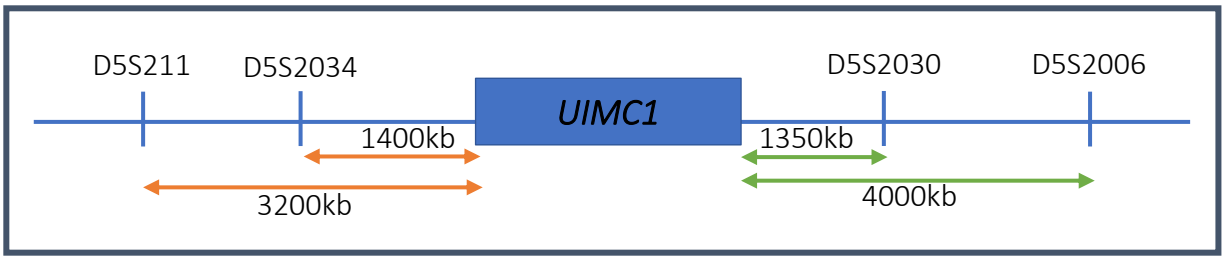


Figure 6.2: Location of STS Markers selected for linkage analysis of individuals with identified *UIMC1* polymorphisms. Approximate distance from *UIMC1* is indicated.

6.2.2 Fragment analysis

For each sample, 1 µL of PCR product was combined with 0.15 µL ROX500 Size Standard (Life Technologies) and 8.85 µL Hi-Di Formamide (Life Technologies). Samples were then resolved using POP-7 polymer on the 3130xl Genetic Analyser (Life Technologies). Generated data was analysed using Peak Scanner (v1.0, ThermoFisher Scientific). All fragment analysis was carried out by Mr. Oliver Van Wageningen at the Flinders Sequencing Facility.

6.2.3 Functional validation of UIMC1

Functional validation of the loss of UIMC1 and the UIMC1:c1690T>C variant was carried out through CRISPR/Cas9 editing. CRISPR/Cas9 Modification work was approved by the Flinders University Biosafety Committee (Exempt Dealing #2017-02).

6.2.4 Cell Culture Methods

For all cell culture experiments, HEK293 cells were seeded 24 hours prior to experimentation and cultured using DMEM low glucose media (Sigma Aldrich) with 10 % FCS, L-Glutamine and Penicillin and Streptomycin, unless specified otherwise. For all MCF10A experiments, cells were seeded 72 hours prior to experimentation and cultured using the MEGM Bullet kit (Lonza) with 100 ng/mL Cholera toxin (Sigma-Aldrich) unless specified otherwise.

6.2.4.1 Puromycin kill curve

Puromycin concentrations for both MCF10As and HEK293s were optimised to determine the concentration that would effectively kill all non-transfected cells within 72 hours. A puromycin kill curve was carried out for each cell line with 0.1–5 µg/mL puromycin. Cells were seeded at a density of 50000 cells and 100000 cells for HEK293 and MCF10A respectively. Cells were plated in triplicate and incubated for 48 hours and 72 hours prior to the addition of puromycin media for HEK293 and MCF10A cells respectively. Media containing varied concentrations of puromycin was added to the cells and they were incubated in the IncuCyte® System (Essen Bioscience, Michigan, USA). Changes in cell growth were captured every 2 hours for a 7-day period and overall confluence was measured. Media containing puromycin was changed every 48 hours.

6.2.5 CRISPR Plasmids

Functional validation of the loss of UIMC1 and the UIMC1:c1690T>C variant was carried out through CRISPR/Cas9 editing. CRISPR/Cas9 Modification work was approved by the Flinders University Biosafety Committee (Exempt Dealing #2017-02). Three different plasmids were used for the modification of mammalian cell lines. All plasmids were a generous gift from Professor Feng Zhang (Broad Institute, MIT, USA) and were provided by Addgene (Massachusetts, USA)

6.2.5.1 Knockout plasmid

PX330-U6-Chimeric_BB-CBh-hSpCas9 encodes a Cas9 from *S.pyogenes* generated a double stranded cut in target DNA. The plasmid map is provided in **Appendix I** and was provided in an agar stab (Plasmid #42230) (Cong *et al.*, 2013).

6.2.5.2 Nickase plasmid with puromycin selection

pSpCas9n(BB)-2A-GFP (PX461) is a cas9n (D10A nickase mutant) from *S.pyogenes* with the addition of puromycin resistance for selection of transfected cells. This plasmid generated a single stranded

cut in the target DNA. The plasmid map is provided in **Appendix I** and was provided in an agar stab (Plasmid #62987) (Ran *et al.*, 2013b)

6.2.5.3 Nickase plasmid with GFP selection

pSpCas9n(BB)-2A-GFP (PX461) is a cas9n (D10A nickase mutant) from *S.pyogenes* with the addition of green fluorescence protein (GFP) for selection of transfected cells. This plasmid generated a single stranded cut in the target DNA. The plasmid map is provided in **Appendix I** and was provided in an agar stab (Plasmid #481040) (Ran *et al.*, 2013b).

6.2.6 sgRNA and repair template design

An online CRISPR design tool was used to determine suitable target sites and assess predicted off-target sites (www.crispr.mit.edu, last accessed 16 January 2018). Guide RNAs (sgRNAs) were designed for two purposes:

1. for gene knockout, and
2. to introduce the identified variant.

The PX330 plasmid was used for gene knockout and two sgRNAs were selected for each locus. The PX461 and PX462v2.0 nickase plasmids were utilised as a pair for the precise cutting of single strands of DNA, resulting in repair via a provided homology directed repair template. This template allowed for the introduction of the specific variants of interest and modified the PAM sequence post-editing to ensure the DNA was not cut again.

The following criteria were addressed when designing sgRNAs:

1. Presence of a 5'-NGG PAM sequence (for *S.pyogenes*) immediately preceding the 20 nucleotide sgRNA sequence.
2. Analysis for off-target editing sites.
3. The addition of a G nucleotide to sgRNAs lacking a 5' G nucleotide which is required for U6 transcriptional initiation.

Guide RNAs were ordered with the following overhangs to enable ligation following digestion with *BpiI* (*BbsI*).

5' – CACCGNNNNNNNNNNNNNNNNNN – 3'
3' – CNNNNNNNNNNNNNNNNNNNCAAA – 5'

Targeted DNA modifications were carried out via a single stranded oligodeoxynucleotide (ssODN) with a flanking sequence of at least 40 bp on each side that are homologous to the target region. The ssODN was designed to introduce the desired base pair change, including the following criteria:

1. The site of modification is no further than 15-20 bp away from the nick site.
2. Homology arms are at least 50 bp in length either side of the site of modification.

All designed sgRNAs and ssODNs were ordered from IDT (Singapore) and provided lyophilised.

6.2.7 Generation of CRISPR plasmids to be used for targeted modifications

One microgram of required plasmid was digested with 2 µL FastDigest *Bpil* (concentration not provided, ThermoFisher Scientific) in 1X FastDigest Buffer with the addition of 10 U Alkaline Phosphatase (Calf Intestinal; New England Biolabs; NEB) in a final volume of 20 µL. Samples were incubated at 37 °C for 45 minutes, followed by heat inactivation @ 80 °C for 20 minutes.

Digested plasmids were electrophoresed on a 0.8 % agarose gel containing GelRed and bands were excised from the gel by visualisation on the Chemi-Doc Imaging System using the XcitaBlue™ filter (Bio-Rad, California USA). Digested plasmids were purified with the QIAQuick Gel Extraction Kit (QIAGEN, Refer to **Chapter 2, Section 2.3.5.2**). Purified products were quantitated using the Nanodrop 1000.

Forward and reverse sgRNAs were diluted 1:1000 in annealing buffer (10 mM Tris-Buffer, 50 mM NaCl, 1 mM EDTA), followed by combining at equimolar concentrations. Serial dilutions of the sgRNA pairs were carried out at concentrations of 500 nM, 5 nM and 50 pM. sgRNAs were heated at 95 °C for 10 minutes, following which the heating block was switched off and samples were equilibrated to room temperature (4-12 hours). Eight and a half microlitres of annealed sgRNAs were phosphorylated through the addition of 100 U T4 Polynucleotide Kinase (PNK; NEB) and 1X T4 ligation buffer (NEB) in a final volume of 10 µL. Samples were incubated at 37 °C for 30 minutes.

Two hundred nanograms of *Bpil*-digested gel-purified plasmid, 3 µL annealed and phosphorylated sgRNAs (at various dilutions), 1 X Quick Ligase Buffer (NEB), and 1 µL Quick Ligase (concentration not provided; NEB) were combined in a final volume of 15 µL. Samples were incubated at 16 °C for 12-16 hours in a Veriti Thermocycler then stored at -20 °C until required.

DH5α competent cells (50 µL) were defrosted on ice and added to pre-chilled 1.5 mL microfuge tubes, along with 7.5 µL of ligation reaction. Reactions were gently mixed and incubated on ice for 20 minutes. Samples were heat shocked at 42 °C for 45 seconds and placed back on ice for a subsequent 5 minutes. One hundred microlitres of pre-chilled SOC media was added to each reaction, and samples were incubated at 37°C in shaking incubator (@200 rpm) for 45 minutes. Following incubation, the entire 150 µL reaction was streaked onto LB agar plates containing 100 µg/mL Carbenicillin (Sigma-Aldrich). Plates were incubated overnight at 37 °C. The number of colonies were counted for each treatment post-incubation.

Standard PCR was carried out (refer to **Chapter 2, Section 2.3.3**) in order to screen colonies for insertion of the sgRNA (Refer to **Appendix F** for primer sequences, and **Appendix G** for cycling conditions). PCR template consisted of colony cells resuspended in 10 µL of MilliQ water. Amplified products were electrophoresed on a 1.5 % agarose gel containing GelRed. The presence of a product in the guide specific PCR combined with the absence of a product in the *BbsI* cut site-specific PCR was used to indicate the successful incorporation of the sgRNA. Colonies which were positive by PCR were then Sanger sequenced to ensure sgRNAs were incorporated in the correct orientation.

Cultures were generated by inoculation with positive colonies and were incubated overnight in LB media with 100 µg/mL Carbenicillin at 37 °C with shaking (200 rpm). Plasmid DNA was extracted with the QIAGEN Plasmid DNA purification Kit (Midi) as per the manufacturer's protocol.

6.2.8 Targeted modification of mammalian cells with CRISPR/Cas9

6.2.8.1 Transfection with Lipofectamine 2000

HEK293 cells were seeded into 24 well plates (Corning) 24 hours prior to transfection at a density of 140,000 cells per well. MCF10A cells were seeded into 24 well plates (Corning) 72 hours prior to transfection at a density of 250,000 cells per well. Cells were transfected with 500 ng plasmid and 10 µM HDR template with Lipofectamine 2000 (Life Technologies) following the manufacturer's recommended protocol. Cells were then incubated at 37 °C, 5 % CO_2 for 24 hours and media changed to DMEM or MEGM (- antibiotics) post-transfection.

At 48 hours post-transfection, cells were selected for successful transfection with CRISPR/Cas9 plasmids through either fluorescence and cell sorting, or puromycin selection. Subsequently, cells were screened for gene modification as outlined in **Section 6.2.9**

6.2.8.2 Transfection via Nucleofection

HEK293 cells underwent electroporation-based transfection method (termed nucleofection) using the Amaxa 4D-Nucleofector™ X Unit (Lonza). Nucleofection protocols were optimised as per the manufacturer's instructions. $1x10^6$ cells per cuvette were transfected with the Amaxa 4-D Nucleofection Kit V (Protocol number D4XC-2002_2011-09). Cells were transfected with a total of 1 µg or 3 µg DNA (for pmaxGFP and PX461/PX462/PX330 plasmids respectively) and incubated in a 6-well plate for 24 hours. Cells underwent puromycin selection and were screened for gene modification as outlined in **Section 6.2.9**. Cells were transfected with a range of protocols, predominantly the pulse protocols A-023 and D-023.

6.2.9 SURVEYOR™ assay and sequencing analysis for confirming gene modification

Following puromycin selection, the viable cell population was screened for successful gene modification using the SURVEYOR assay (**Figure 6.3**, IDT, Singapore). One thousand cells were taken from each treatment and washed with PBS and resuspended in 30 µL MilliQ water. An aliquot of cells was incubated at 98 °C for 10 minutes to lyse cells prior to PCR amplification. Genomic regions containing the CRISPR target sites for both the knockout and point mutation were PCR amplified. Products were visualised on 1.5 % agarose gel containing GelRed and were combined with the PCR product generated from the wildtype sequence for the region and subjected to re-annealing (incubation at 95 °C for 10 minutes and allowed to cool to room temperature) to enable heteroduplex formation. Following re-annealing, products were treated with SURVEYOR nuclease and SURVEYOR enhancer S (IDT) following the manufacturers recommended protocol. Samples were analysed on a 2 % agarose gel containing GelRed. Quantification was based on relative band intensities. Samples which displayed multiple bands were subjected to Sanger sequencing for confirmation.

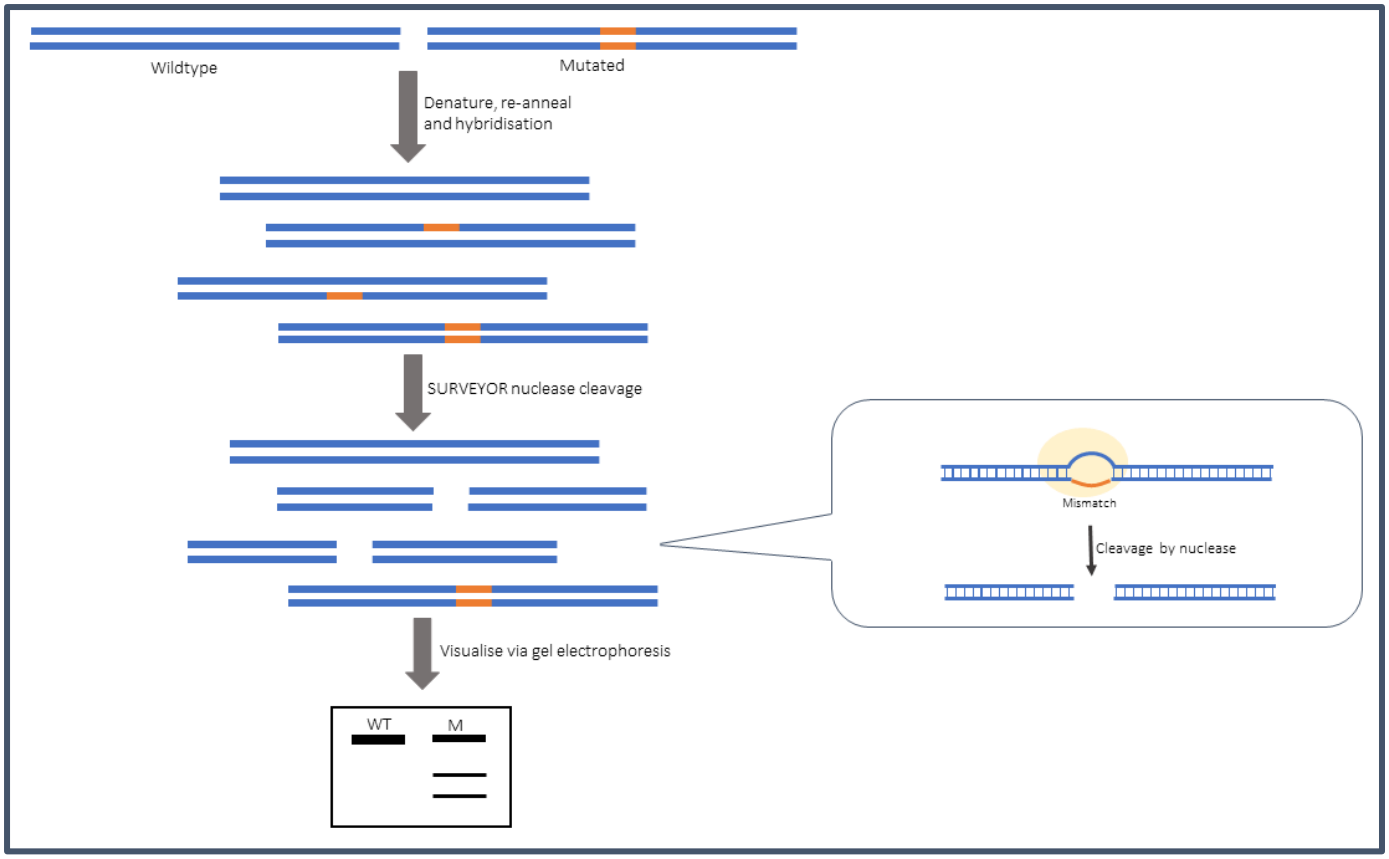


Figure 6.3: Workflow of the SURVEYOR mismatch cleavage assay. Genomic DNA was extracted from a population of cells that were subjected to CRISPR modification. PCR amplification was carried out using primers that flank the target site of modification. Denaturation and re-annealing of the PCR products results in the generation of a mixed population of homo- and hetero- duplexes. The fragments were treated with SURVEYOR nuclease and SURVEYOR enhancer S which cuts only the heteroduplexes. Cleavage products were visualised by agarose gel electrophoresis.

6.2.10 Generation of monoclonal cell lines

Cell populations were serially diluted and plated at a density of 75 cells/10 mL in a 96 well plate with conditioned DMEM media. Cells were incubated at 37 °C, 5 % CO_2 for 2-3 weeks, and visually examined for the growth of monoclonal cell populations, taking note of any wells with multiple populations.

Cells were passaged at 3-4 weeks and transferred to 24 well plates before being taken for gene modification analysis. Cells were continually cultured until confluent in 6 well plates and then frozen down following the methods outlined in **Section 2.4.3**.

6.2.11 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting

The recipes for all buffers required for SDS-PAGE and western blotting are detailed in **Section 2.5.4** of this thesis.

6.2.11.1 Preparation of total cell lysates for SDS-PAGE

To prepare a total cell lysate for analysis of protein expression by western blot, $1x10^7$ cells were placed in a sterile microfuge tube, washed twice in 1 mL PBS (500 *x g*, 5 minutes) then resuspended in 1 mL Pierce's RIPA Buffer (ThermoFisher Scientific) and 1X HALT Protease Inhibitor (ThermoFisher Scientific). Cells were incubated on ice for 10 minutes and then lysed using 3 cycles of sonication (75 % power; 5 second bursts with a 1-minute rest on ice between bursts) using a Branson B12 Sonifier (Branson Sonic Power, Danbury USA). The lysate was incubated for 15 minutes then centrifuged at 13000 x *g* for 5 minutes at 4 °C to remove debris. The supernatant containing the soluble proteins was removed and transferred to a sterile microfuge tube.

6.2.11.2 Protein concentration determination

The protein concentration of individual samples was determined using the EZQ™ Protein Quantitation kit (ThermoFisher Scientific) as per the manufacturers' instructions. A standard curve using BSA was used to determine the sample concentration.

6.2.11.3 1D-SDS PAGE

Total protein from each sample (10 µg) was combined with 1X Reducing Laemmli Buffer + 0.05 % β-mercaptoethanol. Samples were heated to 95 °C for 5 minutes in a heating block and were centrifuged at 13,300 x *g* for 5 minutes. The supernatant was loaded into a Mini-PROTEAN SDS PAGE stain free gel (Bio-Rad) and electrophoresed in 1X SDS-PAGE running buffer. A broad range (10-250 kDa) Precision Plus Protein standard (Bio-Rad) was used to estimate the size of products. The gel was electrophoresed at 220 V for 30 minutes in a Bio-Rad Mini Protean II gel electrophoresis system. Gels were removed from the support and visualised using the Bio-Rad ChemiDoc™ Touch (Bio-Rad) prior to transfer.

6.2.11.4 Transfer and western blotting

A 0.2 µm PVDF-transfer membrane (Bio-Rad) was soaked in methanol and the membrane and blotting pads were soaked in 1X transfer buffer for 3 minutes before beginning transfer. The membrane was placed over the polyacrylamide gel within the transfer cassette and was sandwiched

between 2 blotting pads. Excess buffer, overhanging gel and any bubbles were removed prior to locking the cassette and placing it in the Trans-Blot® Turbo™ Transfer system (Bio-Rad). Transfer was performed at 25 V, 1.0 A for 30 minutes. Once transferred, both the gel and membrane were visualised using the ChemiDoc Touch to ensure successful transfer had occurred.

Following transfer, the membrane was placed in blocking buffer and was incubated at 4 °C overnight with shaking. The blocking buffer was subsequently removed, and the membrane was incubated with Anti-UIMC1 rabbit monoclonal primary antibody (AbCam; ab124763) diluted 1:10000 in Ab-diluent. The membrane was incubated overnight at 4 °C with shaking in the dark. The membrane was washed (3 x 10 minutes) in wash buffer and incubated with donkey anti-mouse horseradish peroxidase secondary antibody (Sigma-Aldrich) diluted 1:10000 in Ab-diluent for 1 hour with rocking at RT in the dark. The membrane was then washed with wash buffer (1 x 15 minute and 2 x 5-minute washes) then 1 x 5-minute wash in TBS. Following the final wash, the membrane was incubated with 2 mL SuperSignal West Pico Chemiluminescent substrate (ECL reagent; ThermoFisher Scientific) for 5 minutes in the dark. Antibody-antigen complexes were visualised on the ChemiDoc Touch System and analysed using the ImageLab software (Bio-Rad).

6.2.12 Irradiation of cells

Cells were passaged and plated 48 hours prior to irradiation. Cells were plated at $5X10^6$ cells per T75 flask, with 7 flasks plated per cell line. Cells were irradiated using the X-RAD 320 (Precision X-Ray, Connecticut, USA) in the Flinders Medical Centre Animal House by Ms. Isabell Bastian (College of Medicine and Pubic Health). Cells were either exposed to 2 Gray (Gy) irradiation, (65cm from source, 300 Kiloelectronvolts (KeV), 13 milliampere (mA)) or sham irradiated over a period of 52 seconds. Following irradiation, cells were incubated at 37 °C, 5 % CO_2 until the appropriate time point for analysis (1, 4- and 24-hours post irradiation).

6.2.13 Analysis of dsDNA damage repair capabilities through γH2AX analysis

The ability to repair DNA double stranded breaks in cells was assessed via phosphorylation of the H2A histone family member X (γH2AX) using a modified version of the method developed by Ms. Marie Lowe (College of Medicine and Public Health, Flinders University). In brief, cells were rinsed with PBS and trypsinised as outlined in **Section 2.4.2**. Cell viability was assessed via Trypan blue exclusion assay. Cells were centrifuged at 700 x *g* and supernatant was aspirated. Cell pellets were resuspended in pre-chilled fixation solution and incubated for 20 minutes on ice. Cells were

centrifuged at 700 x *g,* resuspended in PBS and gently vortexed to remove PBS. Cells were resuspended in PBS and stored at 4 °C for a maximum of 3 days to allow for batch processing of cells at all time points.

Cells were pelleted and resuspended in permeabilization buffer, vortexed gently and incubated at RT for 15 minutes. Cells were spun at 700 x *g* and supernatant was aspirated, resuspended in blocking solution and incubated at 37 °C for 30 minutes. Cells were centrifuged at 700 *x g*, supernatant was aspirated, and cells were washed with PBS twice. Cells were incubated in 30 ng FITC conjugated anti-phospho-histone H2AX (serine^{139}; Millipore) antibody, diluted in blocking buffer for a minimum of 1 hour in the dark. Excess antibody was removed though washing with PBS. Cells were prepared for imaging flow cytometry by resuspension in PBS containing 5 mM EDTA. Prior to imaging, cells were subjected to needle aspiration and 20 ng 4',6-Diamidino-2-phenylindole dihydrochloride (DAPI; Sigma Aldrich) was added for staining of cell nuclei. Cells were imaged using the ImageStream®X Mark I (ISX; Amnis Corporation, Merck-Millipore, Seattle USA). Images between 500 and 1000 cells were acquired at 60X magnification with extended depth of field (EDF) using the 405 nM and 488 nM excitation lasers set to 50 mW and 100 mW respectively.

6.2.14 Analysis of cell images and calculation of γH2AX foci number in cells.

Gamma H2AX foci were quantified in 500 – 1000 images of cells captured with the Inspire™ imaging flow cytometry software using method outlined in Parris *et al.* (2015). In brief, the Ideas™ software applies a series of pre-defined building blocks which first identifies cells that are in focus, followed by single cells based upon cell area and aspect ratio. The in-built spot counting wizard requires user defined populations of cells with very few foci (<3) and high foci numbers (>10). These defined populations are used to count the number of foci in each cell for the whole population of cells. Histograms detailing the number of observed foci, plus the mean +/- standard deviation was obtained for each biological replicate.

6.2.15 Analysis of cell proliferation

For assessment of the growth rate of CRISPR/Cas9 cell lines, cell growth and cell proliferation were analysed. Cells were seeded at a starting density of 5000 in quadruplicate in a 24 well plate. Cells were visualised using an Olympus CKX400 inverted microscope at 100X magnification. The number of cells observed within a defined 25 mm^2 window in each well was counted, and cell confluence was estimated. Cells were counted every 24 hours for 7 days.

6.2.16 Statistical Analyses

All statistical analysis was performed using GraphPad Prism version 8.0 for windows (GraphPad Software, California USA). All statistical tests used are indicated in the appropriate figure legend, with the Two-way ANOVA and multiple t-test analyses corrected with Tukey adjustments for multiple comparisons. A p-value of less than 0.05 was considered significant for all statistical calculations.

6.3 Results

6.3.1 Individuals carrying the same UIMC1 variant are most likely not related

The c.1690T>C variant detected in *UIMC1* is present at a higher frequency in this population (0.833%) than the allele frequency as expected from gnomAD (0.207%). Given that these samples are de-identified, and all individuals came from a relatively small population within South Australia, it is possible that the individuals are related, and this variant may be inherited through a common ancestor. This would not necessarily invalidate this variant as being involved in the development of the breast cancer in both individuals, but it would be informative to establish the likelihood of relatedness between these individuals.

To determine if these individuals are related, analysis of 4 polymorphic markers was carried out for both the *UIMC1*-mutated individuals and a subset of control individuals who do not carry the *UIMC1* variant. Examples of marker zygosity are indicated in **Figure 6.4** and **Figure 6.5** for homozygous and heterozygous samples respectively as determined by the Peak Scanner program. Results from analysis of four markers surrounding *UIMC1* (as illustrated in **Figure 6.2**) are shown in **Table 6.1**.

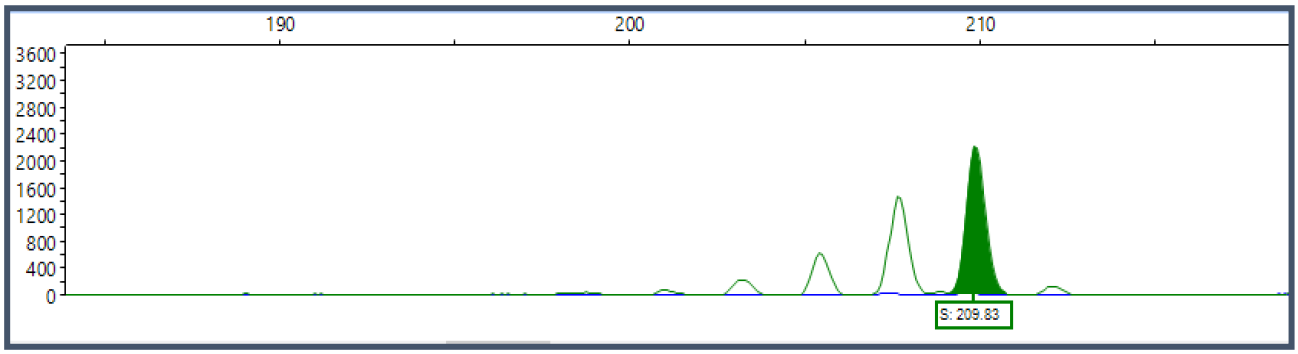


Figure 6.4: Peak Scanner image generated from the fragment analysis of STS marker D5S2034 for individual SABC091. The single shaded green peak indicates a homozygous allele size of 210bp for both alleles. The surrounding peaks are stutter bands. Size (bp) indicated along the x-axis, with intensity indicated along the y-axis.

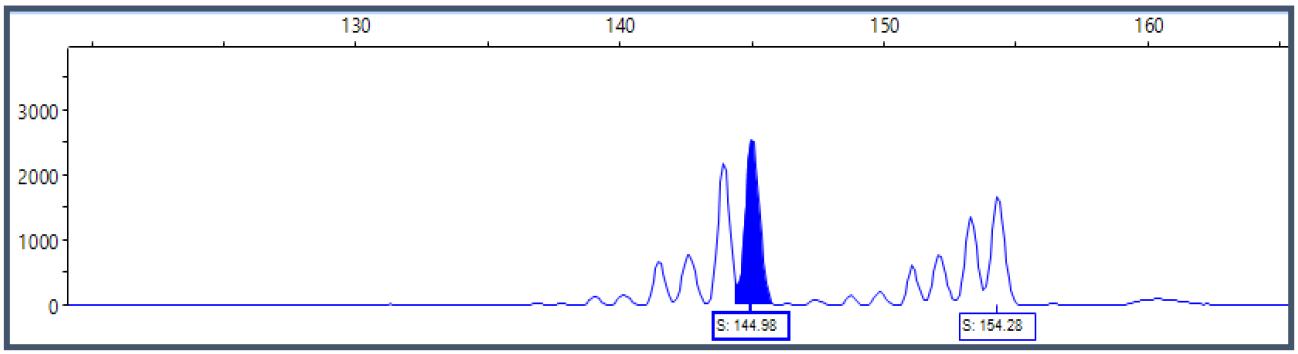


Figure 6.5: Peak Scanner image generated from the fragment analysis of STS marker D5S2006 for individual SABC091. This individual is heterozygous, with two different sized alleles detected in this sample. The generated products were 145bp and 154bp. The surrounding peaks are stutter bands. Size (bp) indicated along the x-axis, with intensity indicated along the y-axis.

Table 6.1: STS Marker Analysis for D5S211, D5S2034, D5S2030 and D5S2006 for individuals carrying the same polymorphism in *UIMC1*. Two individuals with the same predicted pathogenic polymorphism were screened, in addition to control individuals in which the variant was not detected. Sizes are indicated in base pairs (bp). Sizes that are present in multiple individuals (<3 samples) are highlighted in blue and green, with the different colours grouping the different marker sizes together.

Patient ID	D5S211		D5S2034		D5S2030		D5S2006	
	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2	Allele 1	Allele 2
UIMC1:c.1690T>C								
SABC007	194	194	216	208	175	177	152	152
SABC013	190	194	217	210	173	175	157	145
Controls								
SABC023	190	194	210	210	171	173	154	154
SABC076	186	190	210	205	173	173	154	145
SABC124	190	194	210	205	173	175	152	145

The individuals with the same detected polymorphisms shared similarities in only 2 of the 4 markers analysed. Both individuals with the *UIMC1* variant carried a chromosome containing a repeat length of 194bp for marker D5S211, as did 2 of the 3 controls analysed. A similar situation was observed for the other marker that the individuals carrying the *UIMC1* variant had in common, a 173bp allele at marker D5S2030. Of significance is that the individuals carrying the *UIMC1* variant did not have any alleles in common for markers D5S2034 and D5S2006, therefore it is unlikely that these individuals carry the same chromosome 5 (on which *UIMC1* is located). From these results it is most likely that these individuals are not closely related, and that the *UIMC1* variants arose independently. Further work, including analysis of phase is required to determine if the markers of identical length are located on the same chromosome in both individuals. Additionally, analysis with more polymorphic markers in conjunction with analysis of microsatellites that are located closer to

UIMC1 would also be beneficial for this analysis in order to clearly rule out relatedness between these individuals.

6.3.2 sgRNA design for CRISPR/Cas9 gene editing

6.3.2.1 UIMC1:c.1690T>C

A pair of sgRNAs were designed through the MIT Zhang Lab sgRNA design tool to introduce the identified *UIMC1*:c.1690T>C variant identified in exon 13 of *UIMC1*. The pair of sgRNAs, HDR template and introduced mutation are illustrated in **Figure 6.6.**

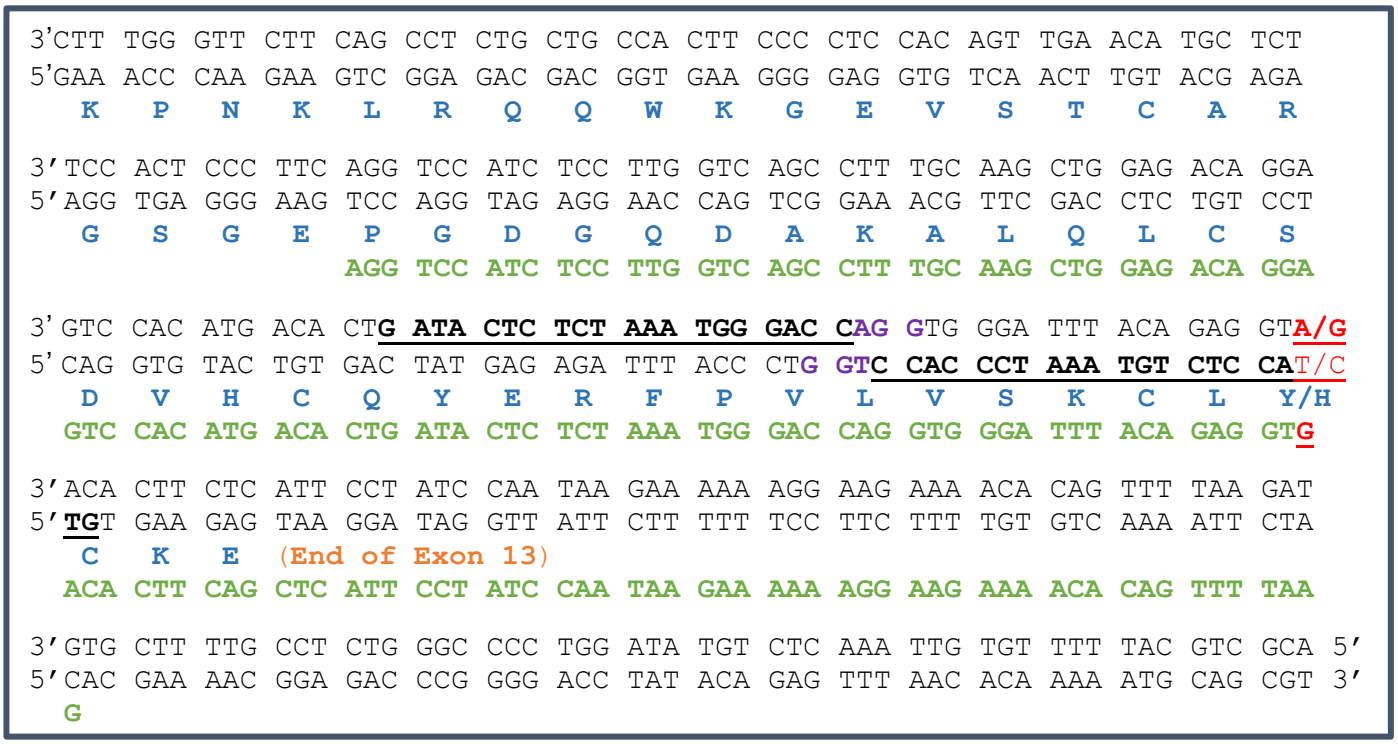


Figure 6.6: *UIMC1* exon 13 and neighbouring intron sequence annotated with paired sgRNAs for CRISPR/Cas9 editing with PX461/PX462v2.0 plasmids Sequence reads in 3' to 5' orientation, designed sgRNA sequences indicated in bold and underlined, PAM sequence indicated in purple UIMC1:c.1690T>C variant indicated in red with reference allele and mutation indicated respectively. Homology directed repair sequence indicated in green, and protein sequence indicated in blue.

6.3.2.2 UIMC1 knockout

Guide sgRNAs were designed using the Zhang Lab sgRNA design tool (www.crispr.mit.edu, last accessed 16 January 2018) to introduce a nonsense mutation early within the coding sequence. *UIMC1* has 15 exons, and the sgRNAs for gene knockout were designed to lie within exon 2 of the gene. In addition, a knockout sgRNA was also generated for exon 13 of *UIMC1* to determine the effects of a variant resulting in premature truncation of the last 2 exons. The identified

UIMC1:c.1690T>C variant is present within a zinc finger like domain of *UIMC1*, the role of which is poorly understood. Whilst this variant is not expected to result in the premature truncation of the UIMC1 protein, it does introduce an amino acid change within a highly conserved zinc finger region. This method may allow us to further understand the functional significance of not only the identified variant, but also this zinc finger like region within *UIMC1*. The location of the designed sgRNAs within *UIMC1* are illustrated in **Figure 6.7**.

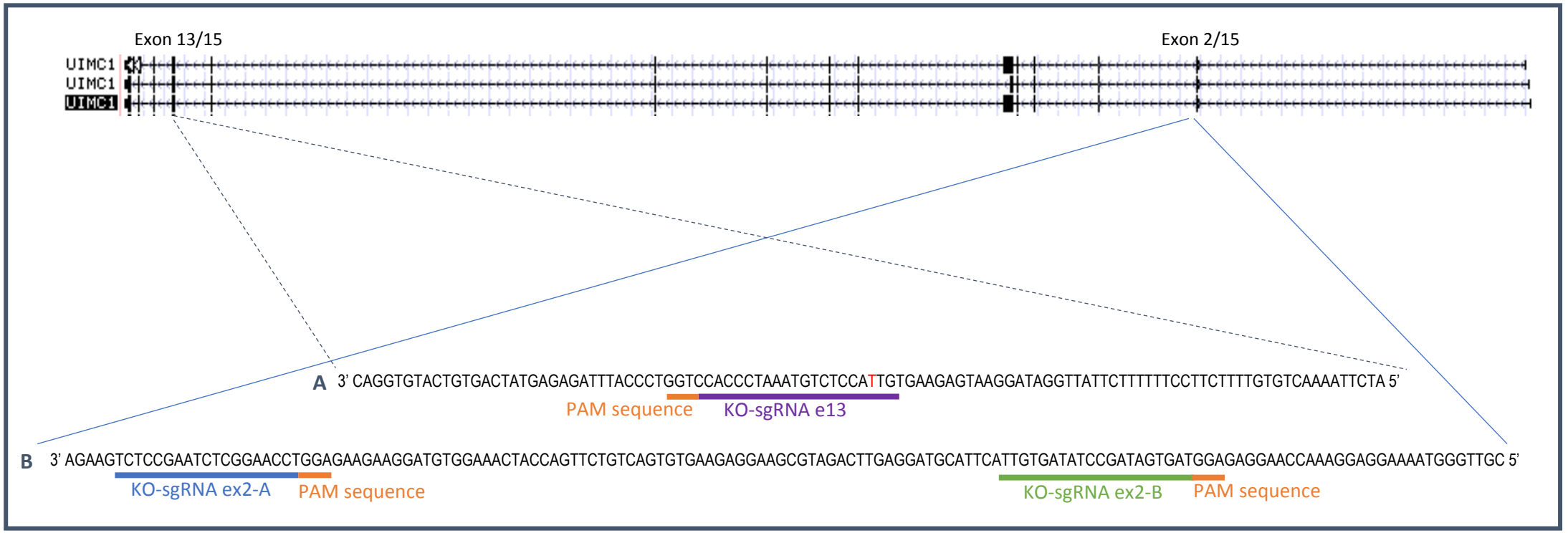


Figure 6.7: sgRNAs designed for knockout of UIMC1 function using Zhang Lab CRISPR/Cas9 tool. Ideogram of UIMC1 from UCSC genome browser indicated in panel above with exons annotated as solid bars, and introns as dashed lines. Several splice variants of UIMC1 can be seen, with the gene containing 15 exons. UIMC1 knockout carried out within both exons 2 and 13. The location of these two exons is indicated at the very top of the image. **A**) Knockout sgRNA designed for exon 13 shown in purple. Variant of interest is highlighted in red and all PAM sequences are underlined in orange. **B)** UIMC1 knockout sgRNAs designed for exon 2, shown in blue and green.

6.3.3 Generation of CRISPR/Cas9 plasmids with successful incorporation of sgRNAs.

The methodology for generating CRISPR/Cas9 plasmids containing the sgRNA inserts required extensive optimisation, including:

- Utilising various isoschizomers of the *BbsI* restriction enzyme (*BbsI*, *BpiI*, FastDigest *BpiI*)
- Variable incubation times for plasmid digestion (15 mins – 12 hours)
- Two different methods of plasmid purification (Gel extraction, Ethanol precipitation)
- Multiple annealing buffers (T4 Ligation Buffer, Sigma Annealing Buffer and IDT Duplex Buffer)
- Variable heating and cooling ramping times for the annealing of sgRNA complexes
- Various ligation kits, concentrations of reagents and ligation times for ligation of sgRNA complexes and digested plasmids; and
- Different batches and types of competent cells for transformations (Stbl3 and DH5α competent cells)

Following this extensive optimisation process, plasmids containing the sgRNAs of interest were obtained (**Figure 6.8**). PCR products were Sanger sequenced to verify the sgRNA was incorporated into the plasmid in the correct orientation (**Figure 6.9**).

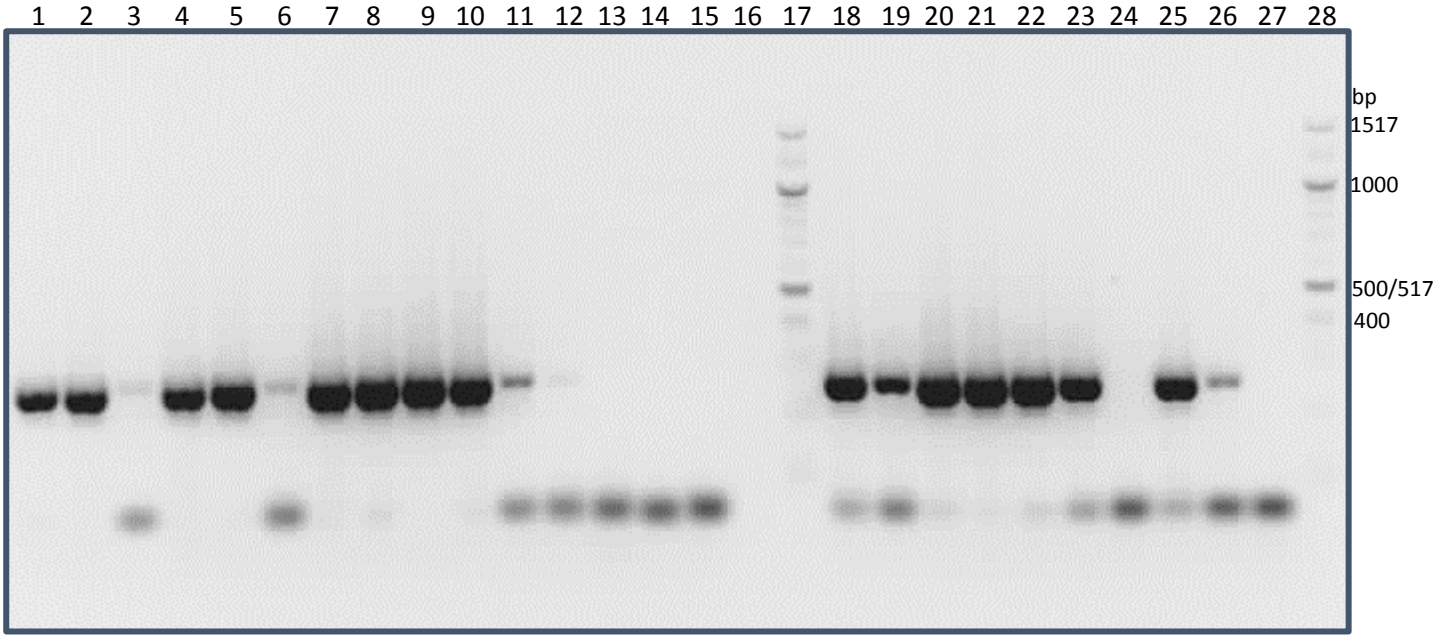


Figure 6.8: Colony PCR to screen CRISPR plasmids for incorporation of sgRNAs into PX461 and PX462v2.0 plasmids. Lanes 17,28; NEB 100bp ladder. Lanes 1-6; PX461 plasmid, sgRNA-A, Lanes 7-13; PX462v2.0 plasmid, sgRNA-A, Lane 14; PX461 No insert, Lane 15; PX462v2.0 no insert, Lane 16; Water, Lanes 18 -20; PX461 plasmid, sgRNA-B, Lanes 21-26; PX462v2.0 plasmid, sgRNA-B, Lane 27; Water

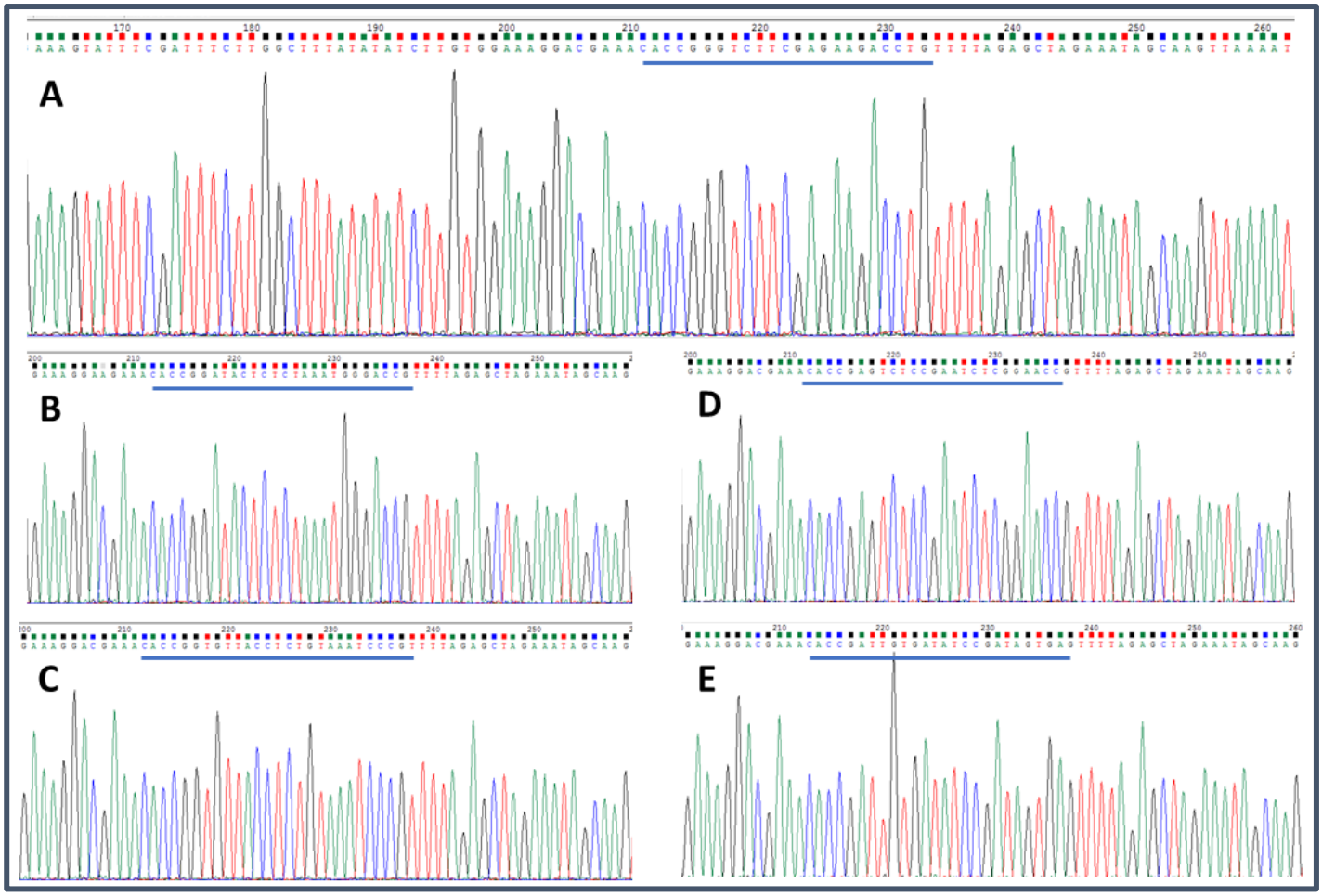


Figure 6.9: Chromatogram traces indicating sequence confirmation of incorporation of sgRNAs into CRISPR/Cas9 PX462 and PX330 plasmids. The 19bp location of edit in underlined with blue line in all panels**.** All traces in the forward direction and sequenced using the U6 promoter primer. Identical results obtained for the reverse sequence (data not shown) **A**. Wildtype PX462 plasmid with *bbsI* cut sites intact. **B.** PX462 plasmid with guide A introduced for edit in exon 13 of *UIMC1*. **C** PX462 plasmid with guide B introduced for edit in exon 13 of *UIMC1*. **D**. PX330 plasmid with guide A introduced for frameshift/knockout mutation in exon 2 of *UIMC1*. **E.** PX330 plasmid with guide B introduced for frameshift/knockout mutation in exon 2 of *UIMC1.*

6.3.4 Assessment of normal proliferation of mammalian cell lines

Prior to performing any manipulations on both the HEK293 and the MCF10A cells, 'normal' growth curves were established to determine the effect of introduced mutations on cell proliferation rates.

6.3.4.1 HEK293 growth curve

HEK293 cells were initially seeded at a density of 120000 cells and were in exponential phase until reaching a plateau at 72 hours (**Figure 6.10**). Cells for all experiments were plated 48 hours prior to experimentation as determined by the optimal time for cell growth from this analysis.

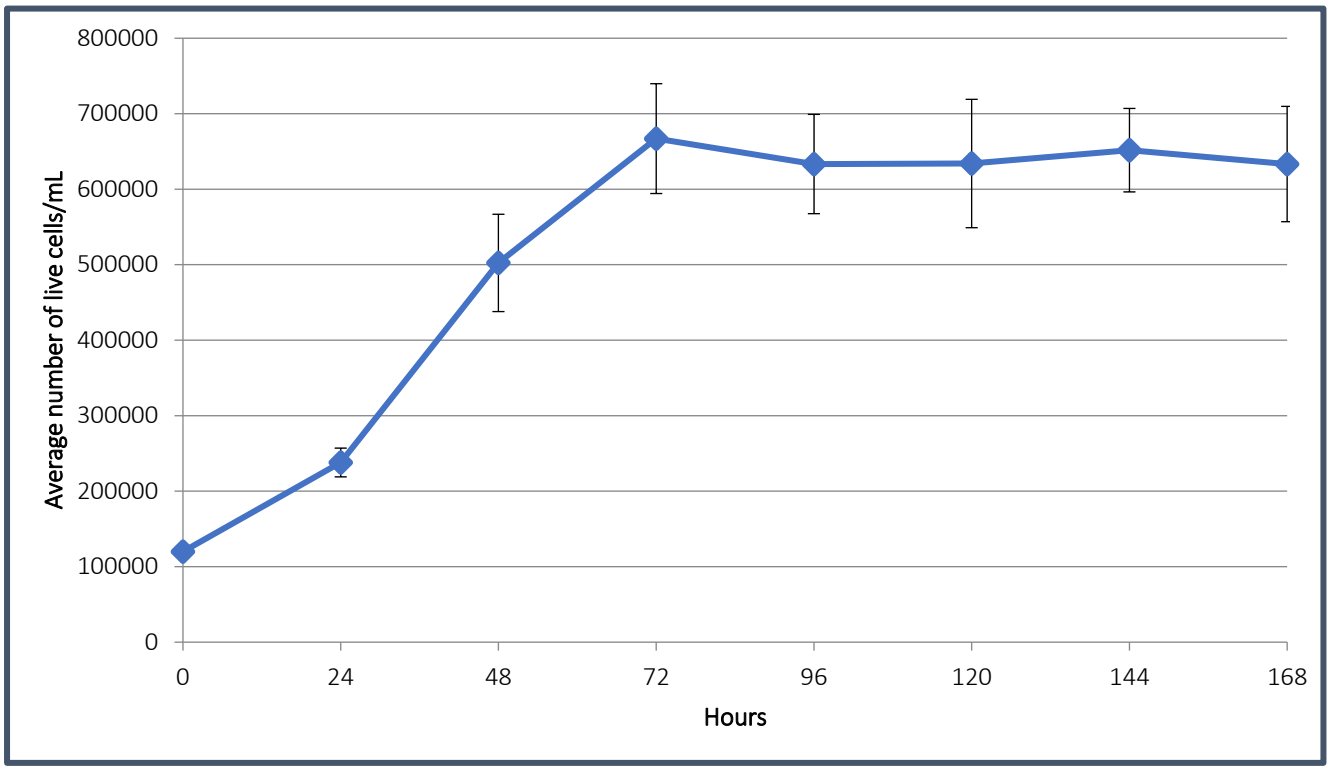


Figure 6.10: HEK293 growth curve over 7 day as determined by trypan blue staining. n=3. Mean cell count +/- standard deviation. Cells were initially seeded at a density of 12000 cells per well and counted every 24 hours.

6.3.4.2 MCF10A growth curve

MCF10As are a slow growing cell line, with cells not reaching a plateau, even after being cultured for 17 days (**Figure 6.11**). Hence, cells were seeded at a higher starting density for all experiments (double that of HEK293 cells) and cultured for 72 hours prior to all experiments.

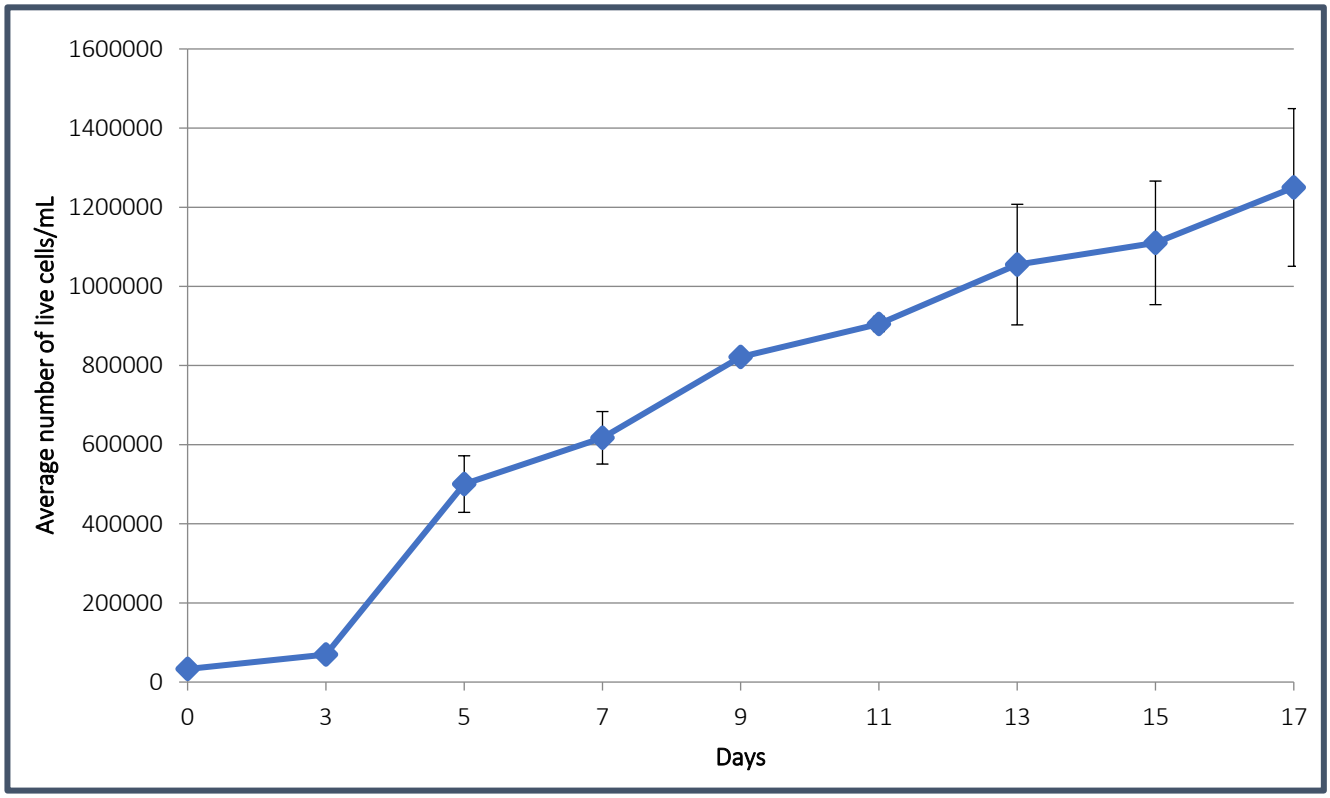


Figure 6.11: MCF10A growth curve over 17 days as determined by Trypan blue staining. n=3. Mean cell count +/- Standard deviation. Cells were initially seeded at a density of 12000 cells per well and counted every 48 hours.

6.3.5 Assessment of optimal concentration of Puromycin for selection post CRISPR/Cas9 transfection

The PX462v2.0 CRISPR/Cas9 plasmid allows for the selection of successfully transfected cells through resistance to puromycin. Therefore, it was necessary to determine the optimal concentration of puromycin to kill cells which do not have puromycin resistance within 72 hours.

6.3.5.1 HEK293 puromycin kill curve

Based on the literature, a varying range of puromycin concentrations were recommended for selection of HEK293 cells (0.5 µg/mL – 2 µg/mL puromycin) over a period of 2-7 days. Cells were seeded at a density of 50000 cells 48 hours prior to applying various concentrations of puromycin onto cells. Cell confluence was measured via the IncuCyte. From the growth curve presented in **Figure 6.12,** 1 µg/mL, 2 µg/mL and 3 µg/mL puromycin all resulted in high numbers of cell death by 72 hours (Range: 7-15 % confluent cells) by 72 hours. The concentration of 3 µg/mL was selected as it resulted in the greatest decrease in cell confluence by 72 hours.

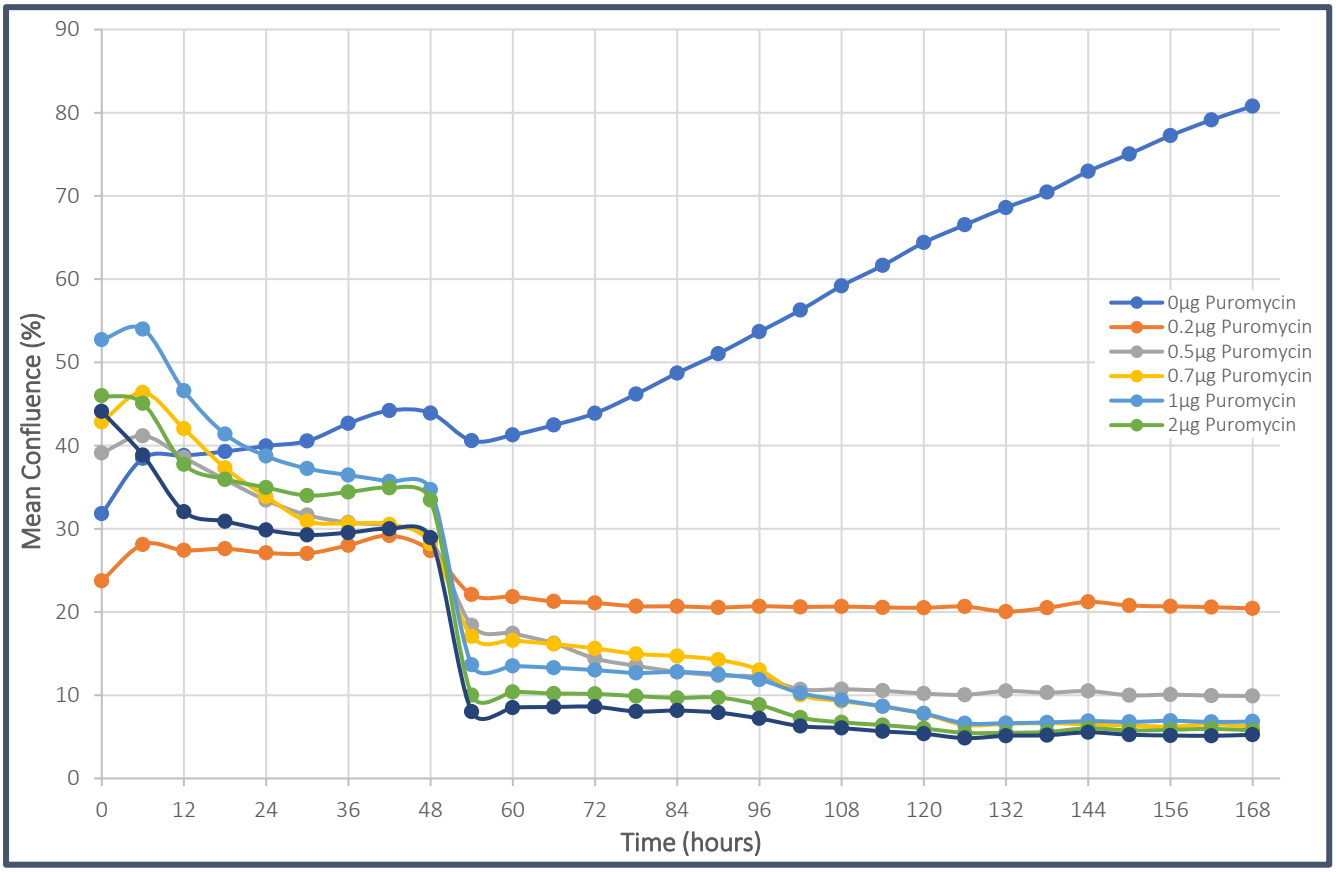


Figure 6.12: HEK293 Puromycin curve over 7-day period. n=12. Cells seeded at initial density of 50000 cells and treated with various concentrations of puromycin ranging from 0 µg/mL to 3 µg/mL in media as recommended by literature. Media changed every 48 hours. Cell confluence measured every 2 hours using the IncuCyte.

6.3.5.2 MCF10A puromycin kill curve

From literature searches, a range of puromycin concentrations (0.2 µg/mL – 3 µg/mL) were recommended for the selection of MCF10A cells (4-7 days). Cells were seeded at a starting density of 100000 cells 72 hours prior to applying various concentrations of puromycin (0 µg/mL – 5 µg/mL) and cell death was visualised via the IncuCyte for 7 days. The growth curve illustrated in **Figure 6.13** showed that 5 µg/mL puromycin was a toxic concentration and resulted in rapid death of the MCF10A cells (as measured by a decrease in cell confluence). Furthermore, 1-3 µg/mL puromycin

resulted in a decrease in cell confluence to >10 % following 72 hours selection. Due to the sensitive nature of the cells, and the similar confluence values observed per concentrations, 1 µg/mL concentration was selected for future work.

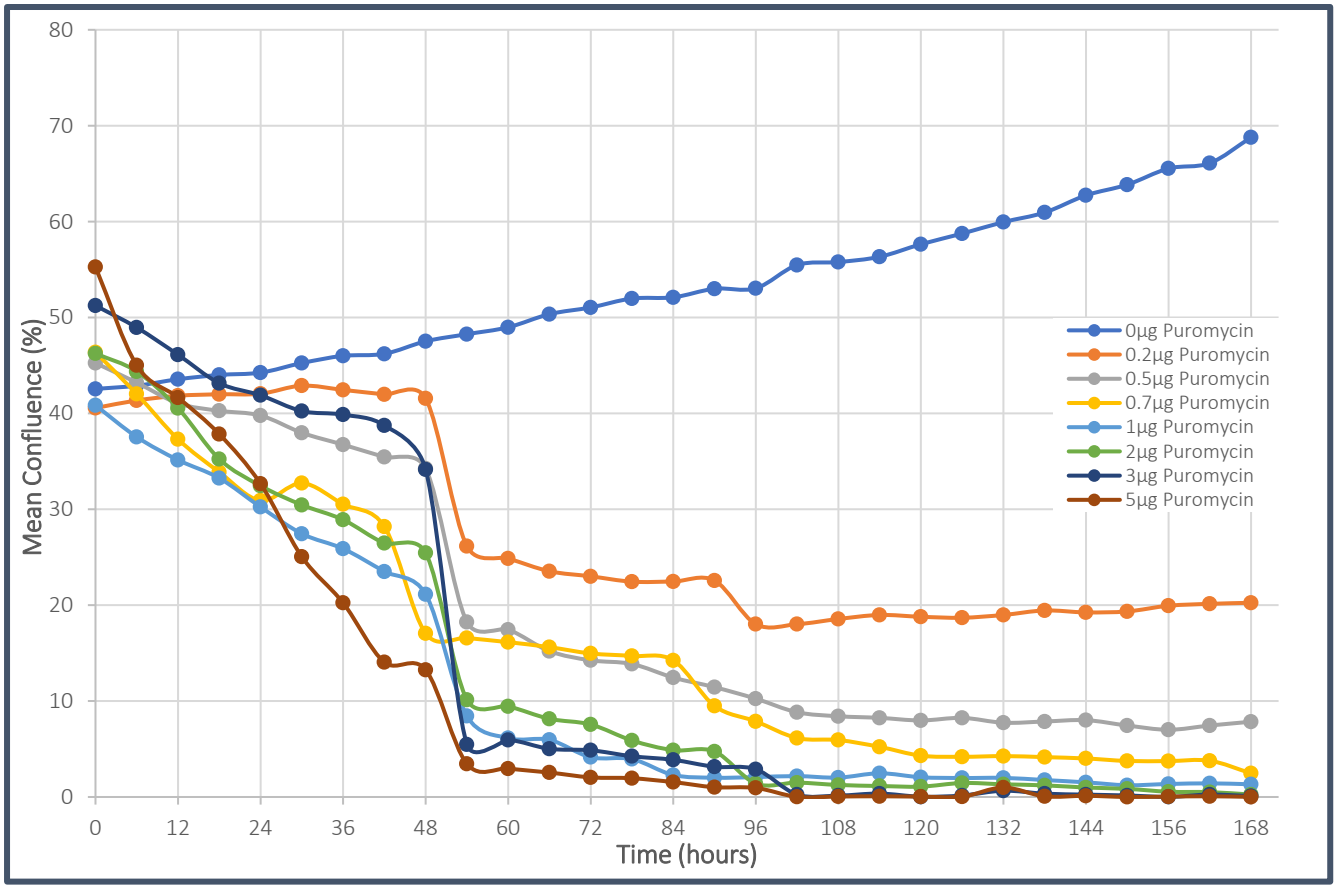


Figure 6.13: MCF10A Puromycin kill curve over 7-day period. n=12. Cells seeded at initial density of 100000 cells and treated with various concentration of puromycin ranging from 0 µg/mL to 5 µg/mL in media as recommended by literature. Media changed every 48 hours. Cell confluence measured every 2 hours using the IncuCyte.

6.3.6 Determination of transfection efficiencies of mammalian cell lines using GFP plasmids in conjunction with Lipofectamine2000

Transfection efficiencies for both MCF10A and HEK293 cell lines were determined with the PMAX plasmid (3486bp, Lonza) and PX461 Plasmid (9289bp).

6.3.6.1 HEK293

HEK293 cells were utilised as a positive control for all CRISPR/Cas9 experiments as they are known to have a high transfection efficiency. Cell transfection concentrations and timepoints were optimised using the pmaxGFP plasmid (Lonza) optimising both Lipofectamine 2000 concentrations and plasmid concentration, in addition to incubation length. The transfection efficiencies (**Figure 6.14**) and cell viabilities following transfection (**Figure 6.15**) were determined via fluorescence and trypan blue respectively.

From the analysis of the various time points, it was evident that an increase in the length of transfection resulted in an increase of GFP+ cells (**Figure 6.14**; 87.5% GFP+ cells after 24-hour transfection). It was determined that 24 hours was the ideal length of time for successful transfection, with minimal cell death in HEK93 cells. However, greater concentrations of plasmid for the increased period of time (24 hours) resulted in greater cell death (**Figure 6.15**). Therefore, 0.5 µg plasmid was used for future transfections to prevent excessive cell death. Minimal differences in cell viability or transfection efficiency were observed between the two Lipofectamine concentrations, therefore the 2 µL Lipofectamine 2000 concentration was utilised for all future experiments following the manufacturer's recommendations.

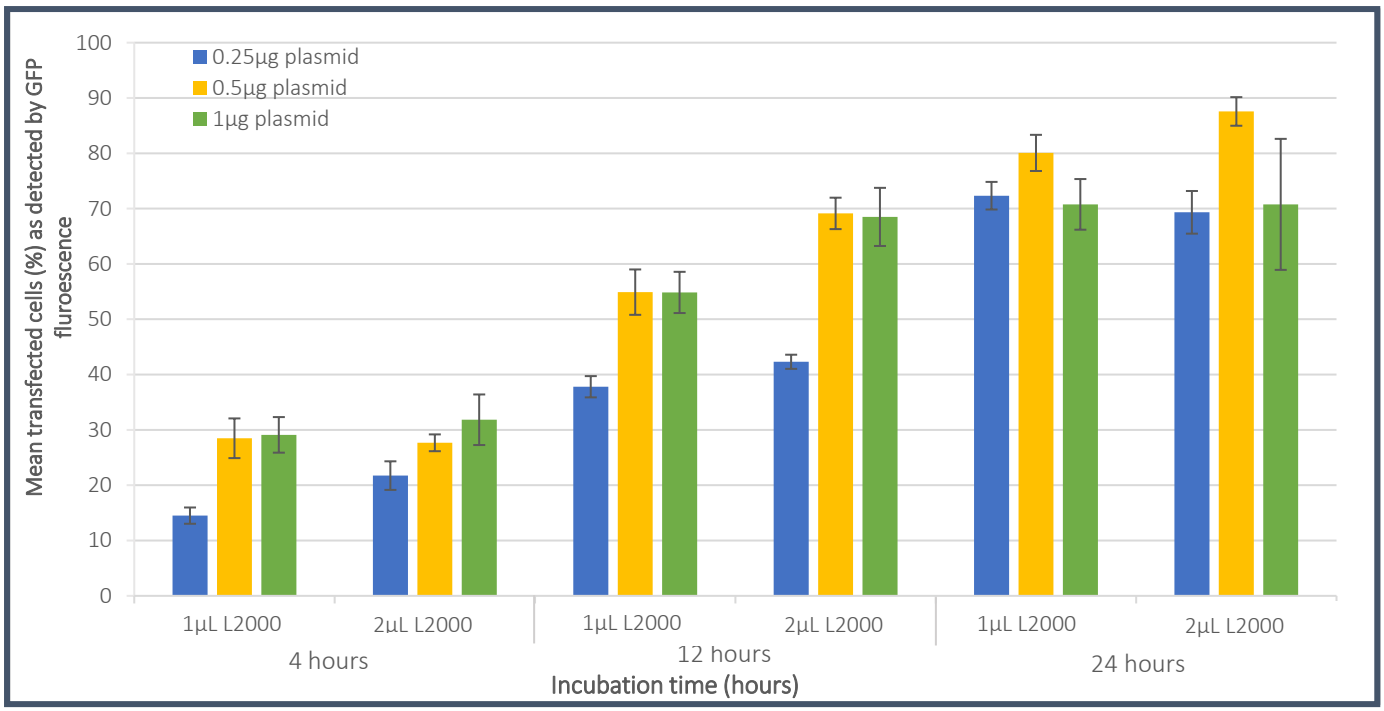


Figure 6.14: Determination of optimal transfection protocol for HEK293 cells using Lipofectamine2000 and pmaxGFP plasmid. n=3. Incubation time (4, 12 and 24 hours), plasmid concentration (0.25 µg, 0.5 µg and 1 µg) and Lipofectamine volumes (L2000; 1 µL and 2 µL) altered to determine optimal conditions for HEK293 transfection. Cells seeded at density of 140000 and plated for 24 hours prior to transfection. Transfection efficiencies determined by % of GFP expressing cells as visualised using the EVOS fluorescent microscope and quantified using ImageJ software. Mean +/- standard deviation.

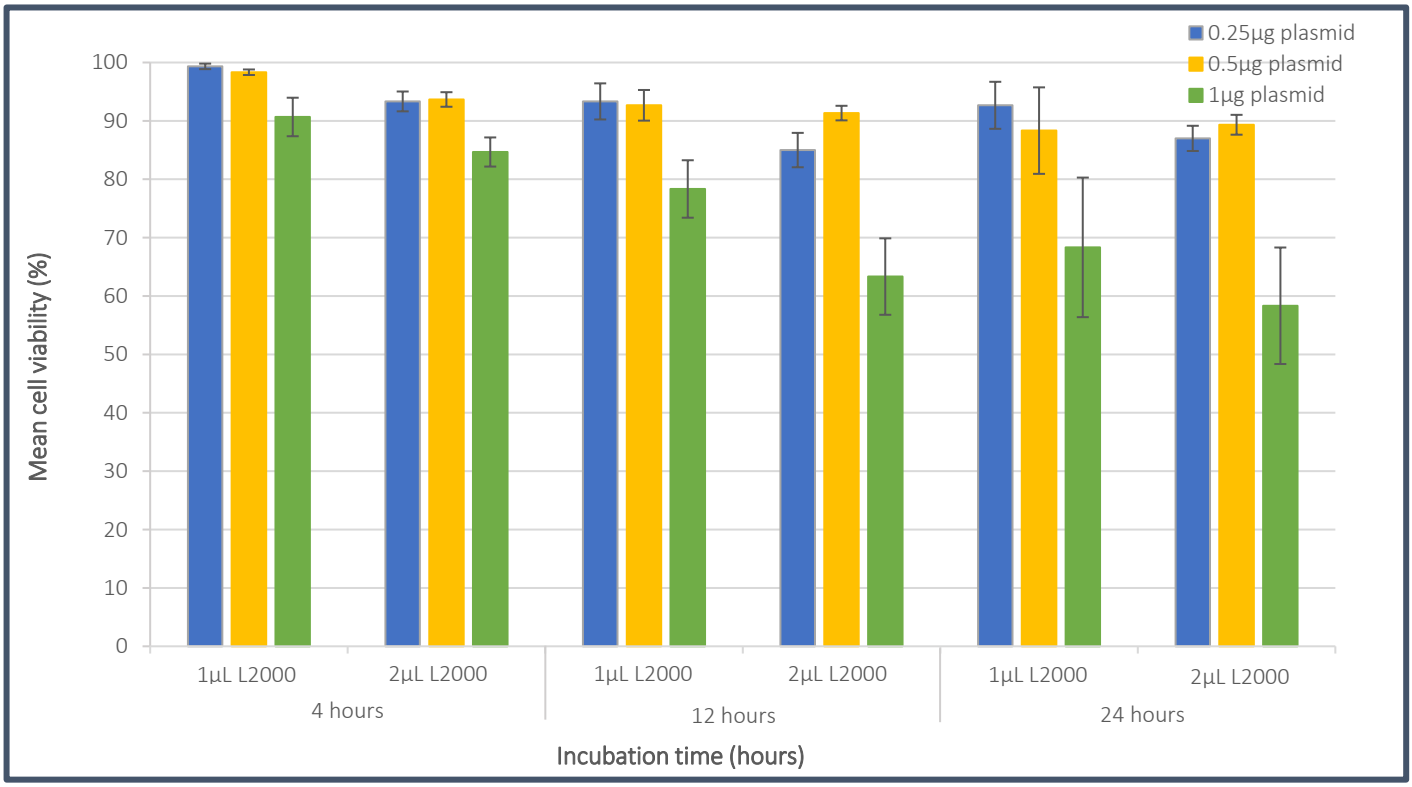


Figure 6.15: Cell viability of HEK293 cells following transfection with Lipofectamine 2000. n=3. Incubation time (4, 12 and 24 hours), plasmid concentration (0.25µg, 0.5µg and 1µg) and Lipofectamine volumes (L2000; 1µL and 2µL) altered to determine optimal conditions for HEK293 transfection. Cells seeded at density of 140000 and plated for 24 hours prior to transfection. Cell viability determined by Trypan blue exclusion method. Mean +/- standard deviation

Following the optimised transfection protocol for HEK293 cells (0.5 µg plasmid, 24-hour transfection using 2 µL Lipofectamine 2000) a comparison of transfection efficiencies and cell viabilities when transfected with pmaxGFP and PX461 plasmids was carried out. The number of GFP positive cells and the cell viability was compared between both plasmids (**Figure 6.16**). As observed in both Panels A and B of **Figure 6.16,** HEK293 cells were dense, with a significantly higher proportion of cells transfected with the pmaxGFP plasmid (Panel A) than the PX461 plasmid (Panel B). Quantification of cell viability and transfected cells indicated that there was a statistically significant decrease in transfection efficiency ($p < 0.0001$) when the PX461 plasmid was used in comparison to pmaxGFP (**Figure 6.20**). This could be attributed to the size of the plasmid and the method of transfection used. There was no significant decrease in cell viability between transfection with either plasmid.

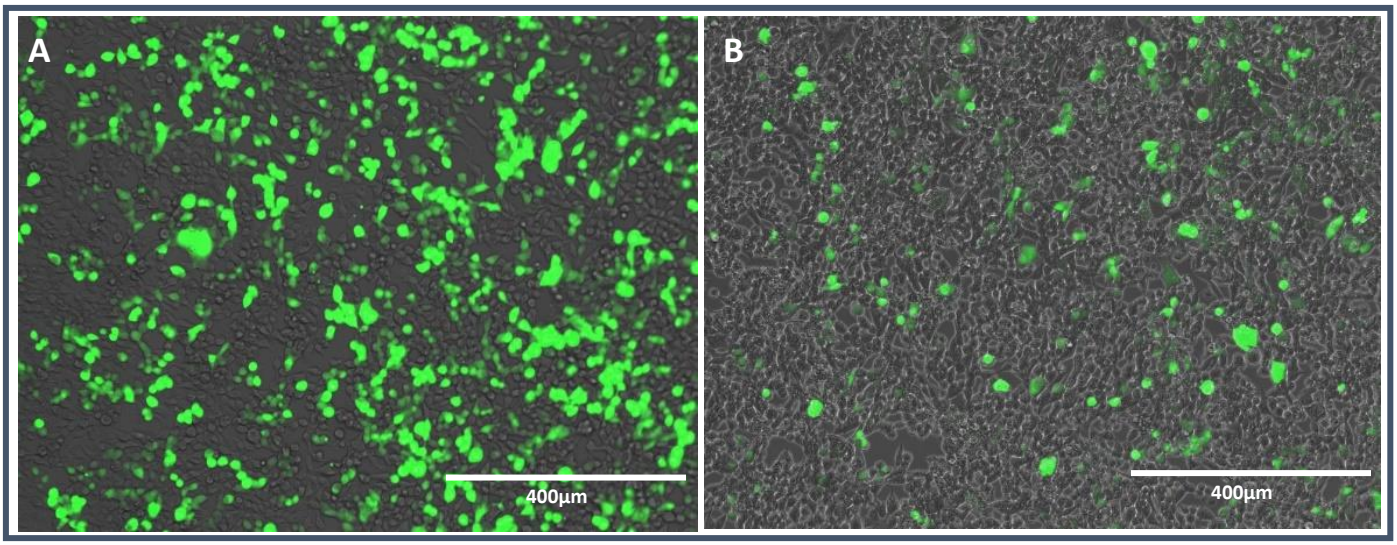


Figure 6.16: Transfections of HEK293 cell lines using Lipofectamine 2000 A. 500 µg PMAX at 100X magnification**, B.** 500 µg PX461 A+B Plasmid at 100X magnification.

6.3.6.2 MCF10A

Transfection optimisation was also carried out with the MCF10A cell line. Incubation times were determined based on the literature and differed from those used for the optimisation of HEK293 transfections. The transfection efficiencies (**Figure 6.17**) and cell viabilities following transfection (**Figure 6.18**) were determined via fluorescence microscopy and trypan blue respectively.

From the analysis of the various time points, it was evident that an increase in the transfection time resulted in an increase of transfected cells (**Figure 6.17**; Maximum 22.3% GFP+ cells after 24-hour transfection with 1 µg plasmid). However, the concentration of plasmid that resulted in the highest

transfection efficiency also resulted in greater cell death, with only 29% of cells remaining viable after a 24-hour transfection. Shorter incubation times were also carried out, in order to try to preserve cell viability. However, this resulted in minimal GFP+ cells (>10 % transfection efficiency, where cell viability was <50% for transfected MCF10As at 3, 6 and 12 hours). Unfortunately, as the length of transfection was increased, the cell viability rapidly decreased, with minimal cells remaining adherent in wells following incubations.

Furthermore, it was evident that an increase in Lipofectamine 2000 concentration did not correlate with an increase in GFP+ cells (**Figure 6.17**), but rather resulted in an increase in cell death (**Figure 6.18**). Therefore, in order to prevent toxicity, cells were transfected with a lower dose of Lipofectamine 2000 (1 µL) for future experiments. A significant difference was observed in cell viability between the concentration of plasmid used for transfection, and as a result, 0.5 µg plasmid was used for future experiments.

Overall, **Figure 6.17** and **Figure 6.19** indicates the poor transfection efficiencies of the MCF10A cell line. Altering the concentration of plasmid and Lipofectamine 2000 as well as incubation times did not yield a transfection efficiency greater than 25%, with those parameters resulting in greater transfection efficiency also resulting in greater cell death (**Figure 6.18**). As mentioned, minimal cells remained adherent to the wells (as indicated by the low number of cells observed in both panels of **Figure 6.19**).

Following the optimal transfection protocol for MCF10As (0.5 µg plasmid, 24-hour transfection using 1 µL Lipofectamine 2000) a comparison of transfection efficiencies and cell viabilities when transfected with pmaxGFP and PX461 plasmids was carried out. The number of GFP positive cells and the cell viability was compared between both plasmids (**Figure 6.19**). As observed in both Panels A and B of **Figure 6.19**, minimal cells remained post-transfection, due to high numbers of cell death. However, greater numbers of cells were successfully transfected with the pmaxGFP plasmid (Panel A) as compared to the PX461 plasmid (Panel B). Quantification of cell viability and transfected cells indicated that there was a statistically significant decrease in transfection efficiency ($p=0.0043$) when the PX461 plasmid (19.3% GFP+ cells) was used in comparison to pmaxGFP (5.33% GFP + cells; **Figure 6.20**). Furthermore, there was a significant decrease in cell viability between transfection with the pmaxGFP plasmid in comparison to the PX461 plasmid ($p=0.0123$).

Due to the low transfection efficiency of MCF10A cells, the slow growing nature of the cell line, the high numbers of cell death associated with Lipofectamine 2000 and plasmid concentration toxicity and issues associated with reagent availability (refer to **Section 6.4**), this cell line was not utilised further.

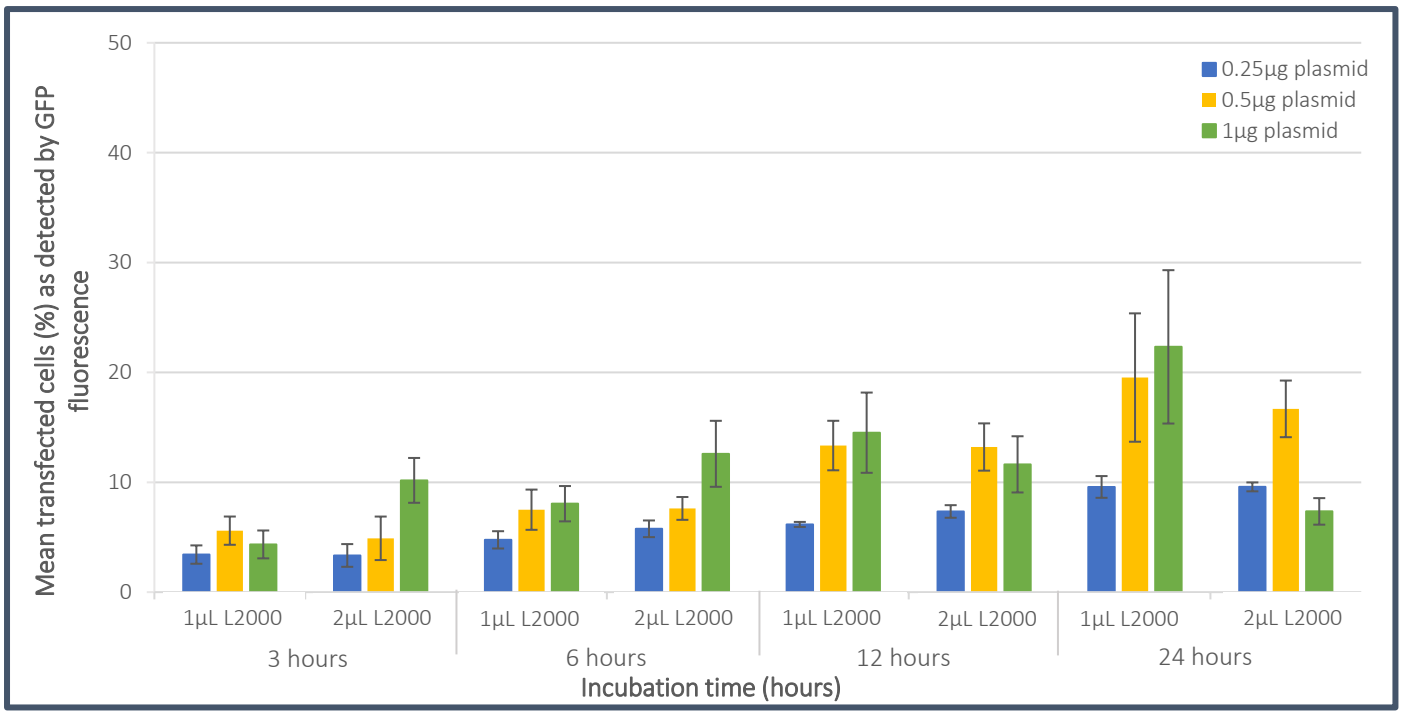


Figure 6.17: Determination of optimal transfection protocols for MCF10A cells using Lipofectamine 2000 and pmaxGFP plasmid. n=3. Incubation time (3, 6, 12 and 24 hours), plasmid concentration (0.25 µg, 0.5 µg and 1 µg) and Lipofectamine volumes (L2000; 1 µL and 2 µL) altered to determine optimal conditions for MCF10A transfection. Cells seeded at density of 250000 and plated for 72 hours prior to transfection. Transfection efficiencies determined by % of GFP expressing cells as visualised using the EVOS fluorescent microscope and quantified using ImageJ software. Mean +/- standard deviation

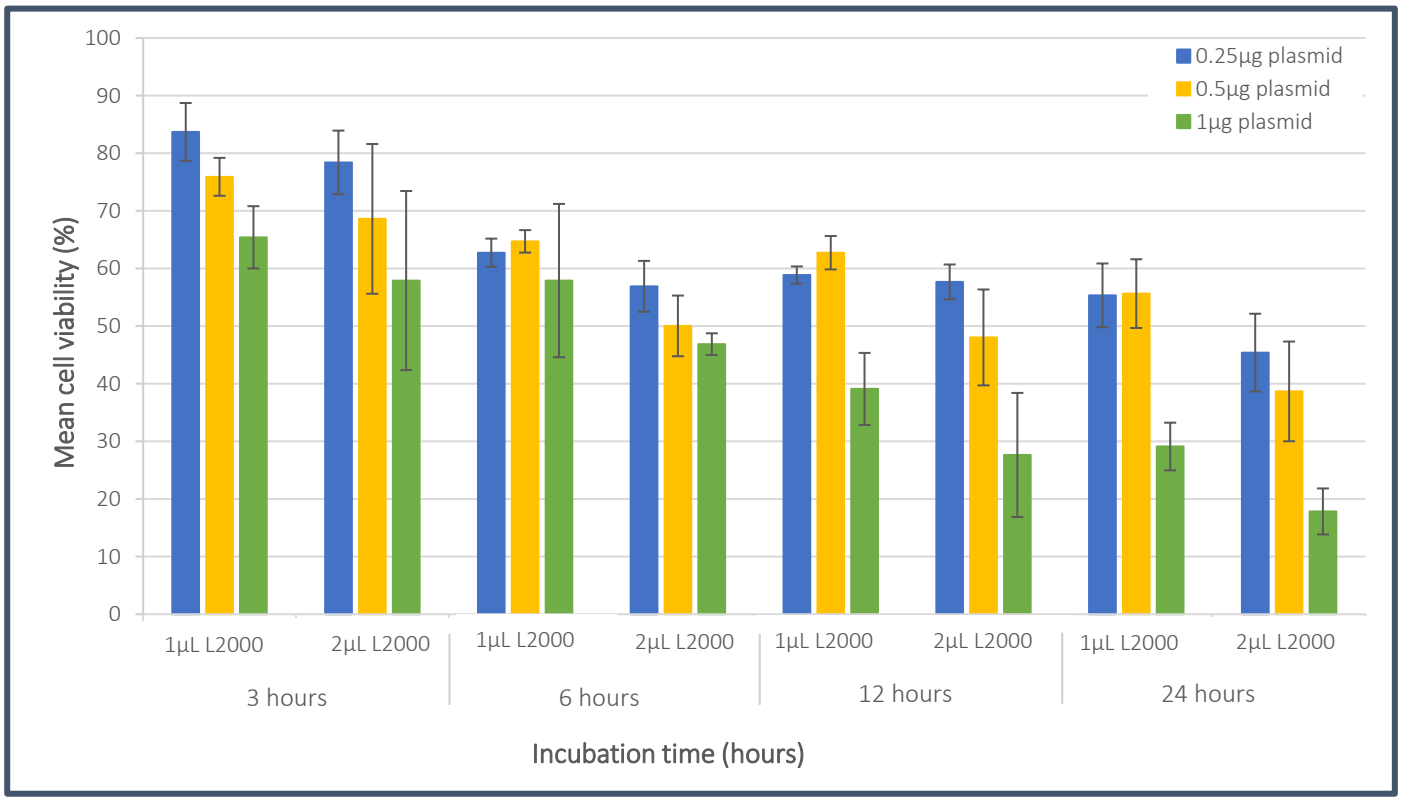


Figure 6.18: Cell viability of MCF10A cells following transfection with Lipofectamine 2000. n=3. Incubation time (3, 6, 12 and 24 hours), plasmid concentration (0.25 µg, 0.5 µg and 1 µg) and Lipofectamine volumes (L2000; 1 µL and 2 µL) altered to determine optimal conditions for MCF10A transfection. Cells seeded at density of 250000 and plated for 72 hours prior to transfection. Cell viability determined by Trypan blue exclusion method.

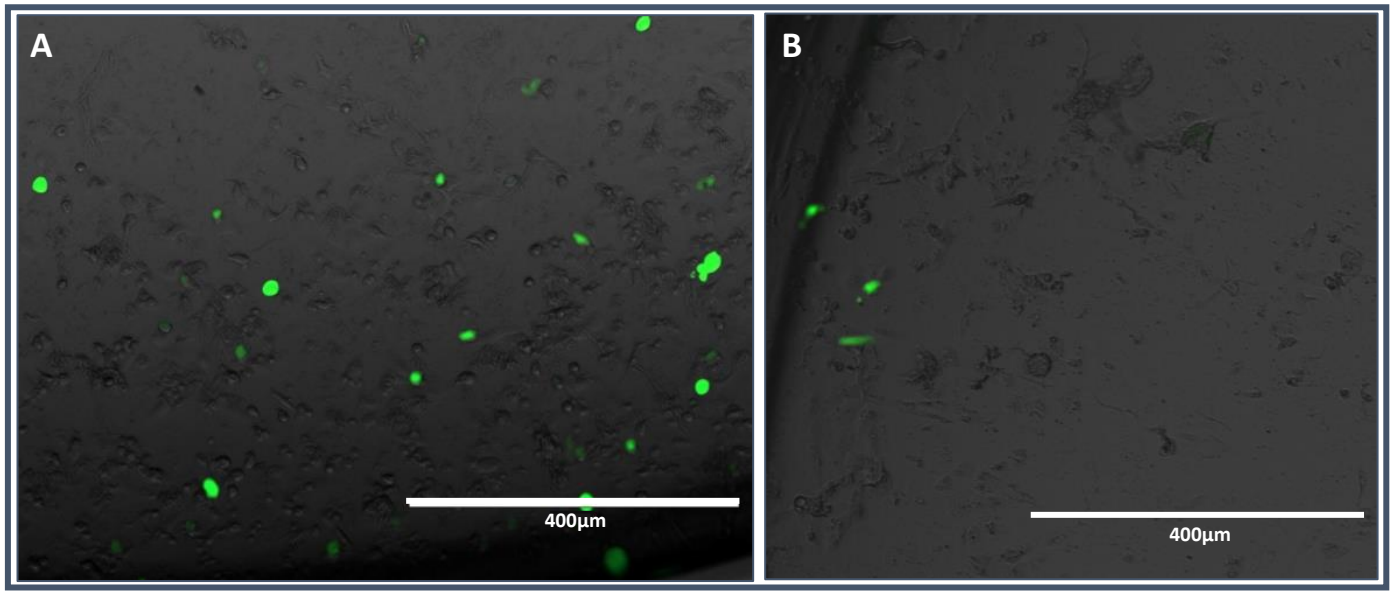


Figure 6.19: Transfections of MCF10A cell lines using Lipofectamine 2000. Cells seeded at density of 250000 and visualised 24 hours after transfection. **A)** 500 µg pmaxGFP at 100X Magnification**, B)** 500 µg PX461 A+B Plasmid at 100X magnification.

.

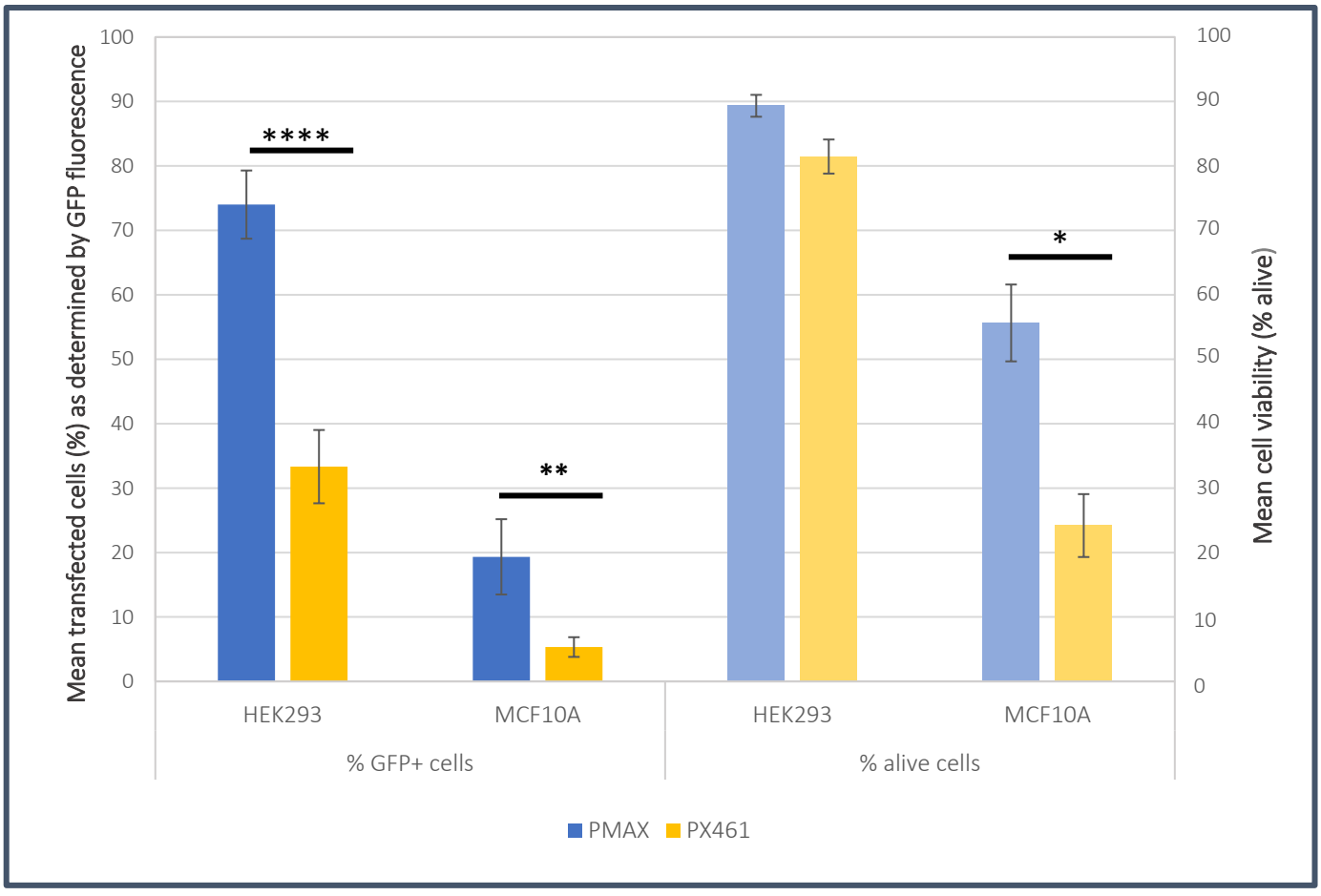


Figure 6.20: Comparison of transfection efficiency and cell viability following transfection of pmaxGFP and PX461 on HEK293 and MCF10A cell lines using Lipofectamine 2000. n=4. Cells transfected with optimised protocols as determined in Figure 6.14 and Figure 6.17. Transfection efficiencies determined by % of GFP expressing cells as visualised using the EVOS fluorescent microscope and quantified using ImageJ software. Cell viability determined by Trypan blue exclusion method. Mean +/- standard deviation. Statistical significance determined using Welch's t-test, with **** indicating p <0.0001, ** indicating p <0.005 and * indicating p <0.05

6.3.7 Few HEK293 cells indicate signs of gene editing with Lipofectamine 2000 transfection

CRISPR/Cas9 transfection was carried out on HEK293 cells using the optimised protocol for Lipofectamine 2000. Cells were treated with 3 µg/mL puromycin for 72 hours post transfection and a sample of the cell population was taken from each treatment. These cells were screened for edited UIMC1 with the SURVEYOR assay (**Figure 6.21**). A distinct band can be observed at 200 bp in the positive control in Lane 2, indicating the formation of a heteroduplex that has been cut by the SURVEYOR nuclease. Very faint bands can also be observed in lanes 3, 4 and 8 for the screened cell populations, indicating that genome editing has occurred within a small sample of the cell population screened. Monoclonal cell lines were then generated from these cell populations.

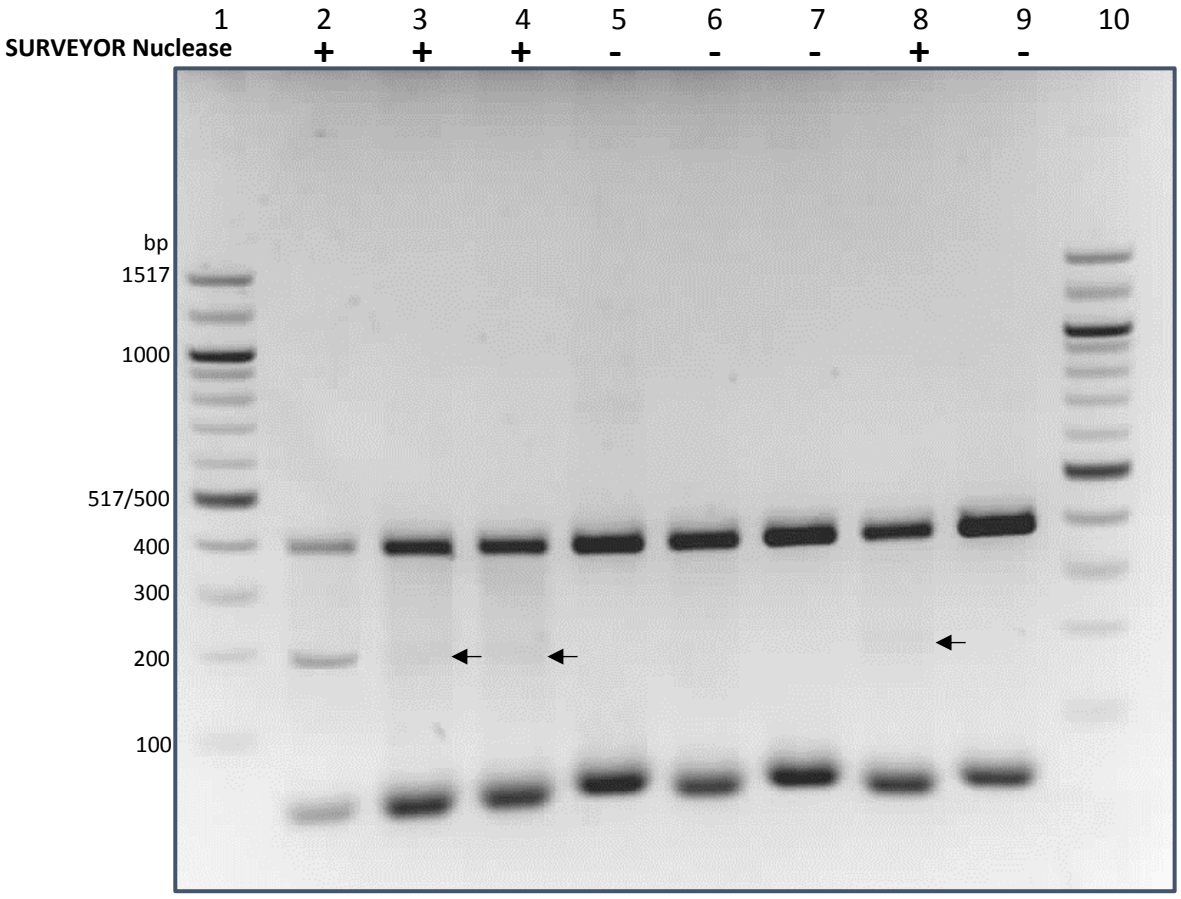


Figure 6.21: SURVEYOR assay on Lipofectamine2000 transfected CRISPR/Cas9 cell populations post puromycin selection. Products run on 2.5 % agarose gel containing GelRed. Lanes 1+ 10; NEB 100bp Ladder, Lane 2 + 5; Positive control (SABC013, heterozygous for UIMC1c.1690T>C variant), Lanes 3 + 6; PX330 exon 13 guide A transfected cell population, Lanes 4 + 7; PX330 exon 13 guide B transfected cell population, Lanes 7 + 8; PX462A+B Transfected cell population. + indcates presence of SURVEYOR enhancer and SURVEYOR nuclease. Presence of faint bands at approximatley 200bp indicated by arrows.

6.3.8 All monoclonal cell lines generated from Lipofectamine 2000 transfected cells were found to have a wildtype *UIMC1* sequence.

Over 150 monoclonal cell lines were generated for the 3 different plasmids used for gene editing. The majority (135 cell lines) were screened for edits within *UIMC1* with the SURVEYOR assay and/or Sanger sequencing. Unfortunately, all cell lines generated were found to be wildtype at the regions of interest (**Figure 6.22** is representative of the 135 screened cell lines). In an attempt to improve editing efficiency, nucleofection was then used to transfect the CRISPR/Cas9 plasmids.

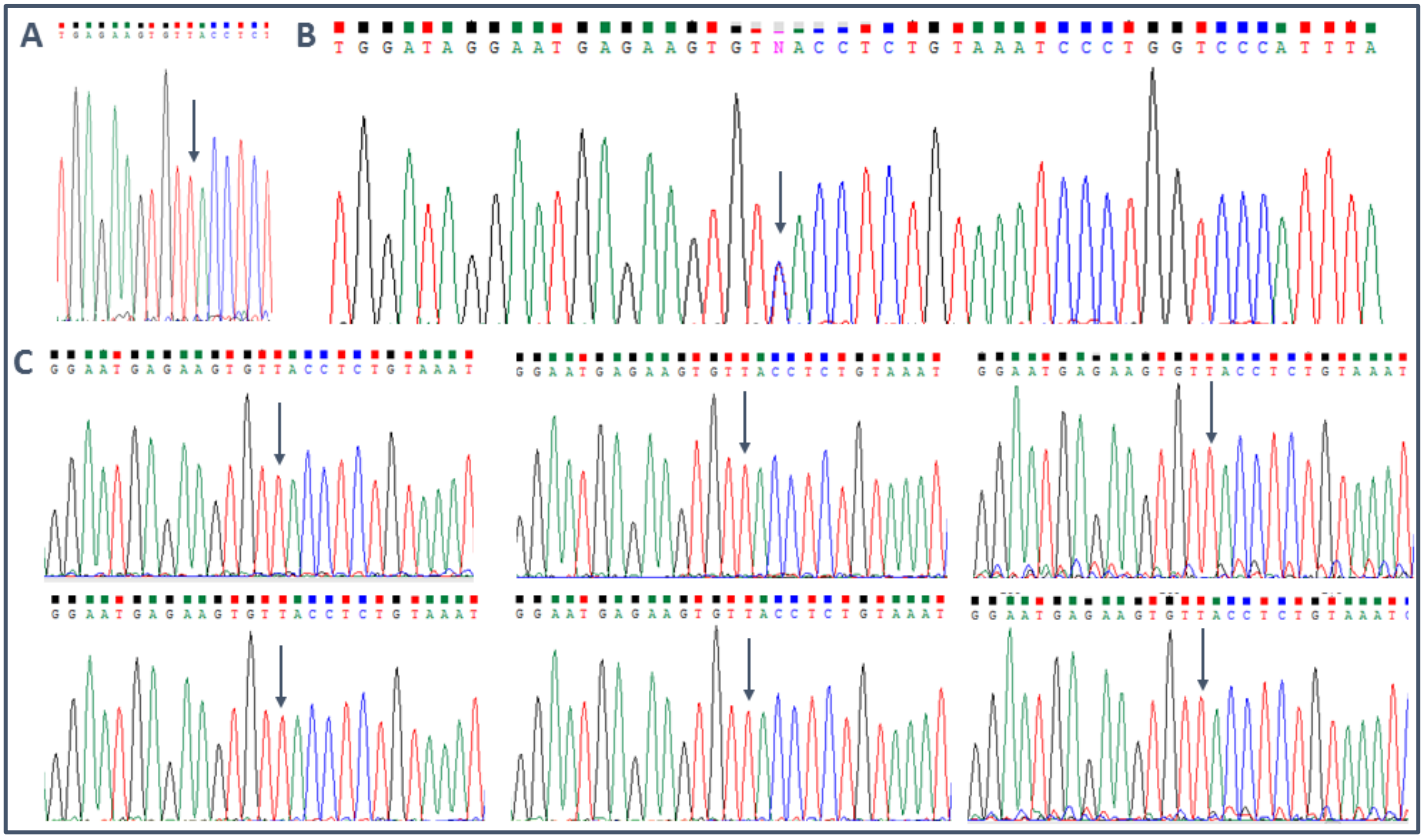


Figure 6.22: Chromatogram traces of monoclonal cell lines generated from Lipofectamine 2000 transfected cell populations. All traces shown in the reverse direction. Identical results were obtained for the forward sequence (results not shown). Blue arrow indicates the site of the 1690T>C variant. **A.** Wild-type HEK293 cells. **B.** Heterozygous 1690T>C variant in SABC007 **C.** Six monoclonal cell lines screened for the incorporation of the T>C variant, all of which were wildtype

6.3.9 Optimisation of nucleofection protocols.

Prior to performing nucleofection, it was necessary to optimise the protocol for the HEK293 cell line. Four different pulse protocols were recommended by Lonza for HEK293 cells, and a control plasmid was provided (pmaxGFP Plasmid). Pulse protocols A-023 and Q-001 (Panels A and C respectively in **Figure 6.23**) resulted in the greatest number of cells expressing GFP protein 24 hours post nucleofection (**Figure 6.25**). However, due to the size difference between the control plasmid and the CRISPR plasmids, it was also necessary to carry out optimisation using the PX461 plasmid.

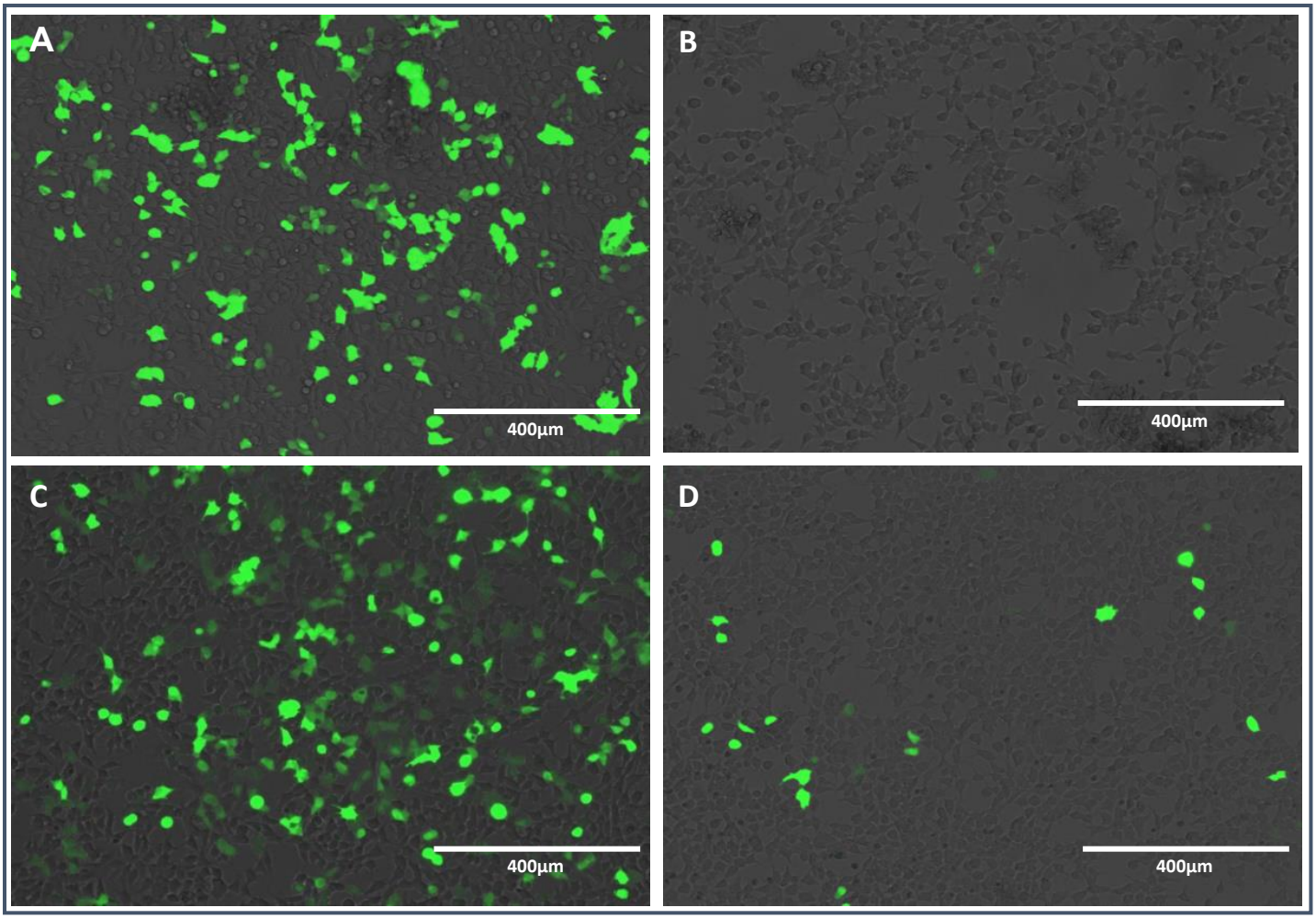


Figure 6.23: Transfection efficiencies of pmaxGFP plasmid with HEK293 for optimisation of Nucleofection Pulse Protocol (Lonza). 10^6 cells per cuvette, cells visualised using EVOS fluorescence microscope, viewed at 100X magnification. Each panel represents a different nucleofection pulse protocol **A)** A-023 **B)** D-023 **C)** Q-001 **D)** X-001. Each cuvette contained 1 µg pmaxGFP plasmid

The provided control pmaxGFP plasmid was approximately 3,500 bp in size, and as such, experiments must be scaled accordingly in order to accurately reflect efficiency for the plasmid to be transfected (according to the manufacturers recommendations). As the control plasmid was 3x smaller than the PX461 plasmid (approximately 10,000 bp), the same optimisation experiment was conducted using 3 times more GFP expressing CRISPR/Cas9 plasmid (3 µg; **Figure 6.24**).

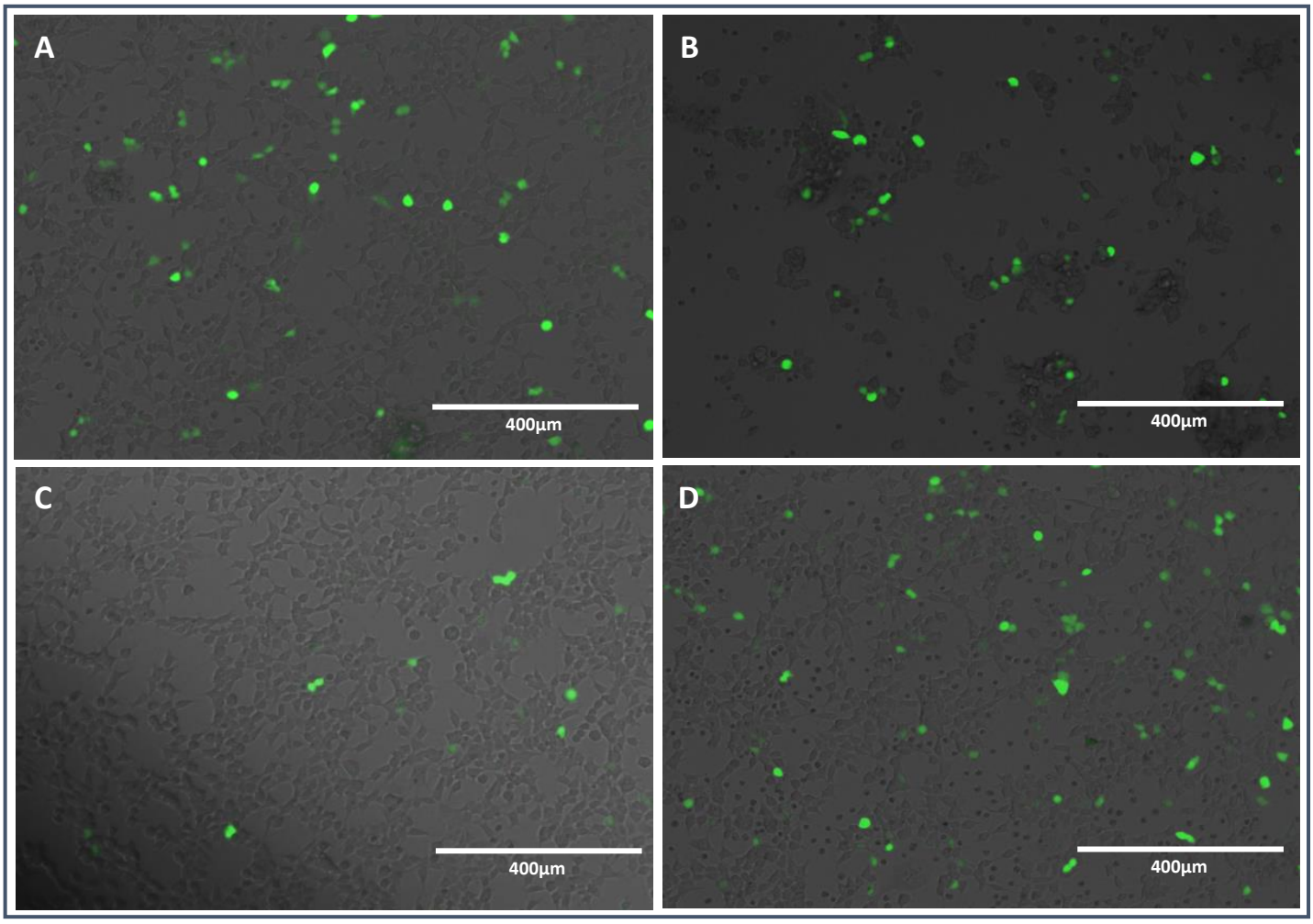


Figure 6.24: Transfection efficiencies of PX461 plasmid on HEK293 for optimisation of Nucleofection Pulse Protocol (Lonza). 10^6 cells per cuvette, cells visualised using EVOS fluorescence microscope, viewed at 100X magnification. Each panel represents a different nucleofection pulse protocol **A)** A-023 **B)** D-023 **C)** Q-001 **D)** X-001. Each cuvette contained 3 µg PX461 plasmid.

Interestingly, pulse protocols A-023, D-023 and X-001 resulted in the greatest level of GFP expression with the PX461 plasmid, however, significantly less cells were expressing the GFP plasmid. This contrasts with the previous results, as the Q-001 protocol resulted in the lowest number of GFP-expressing cells.

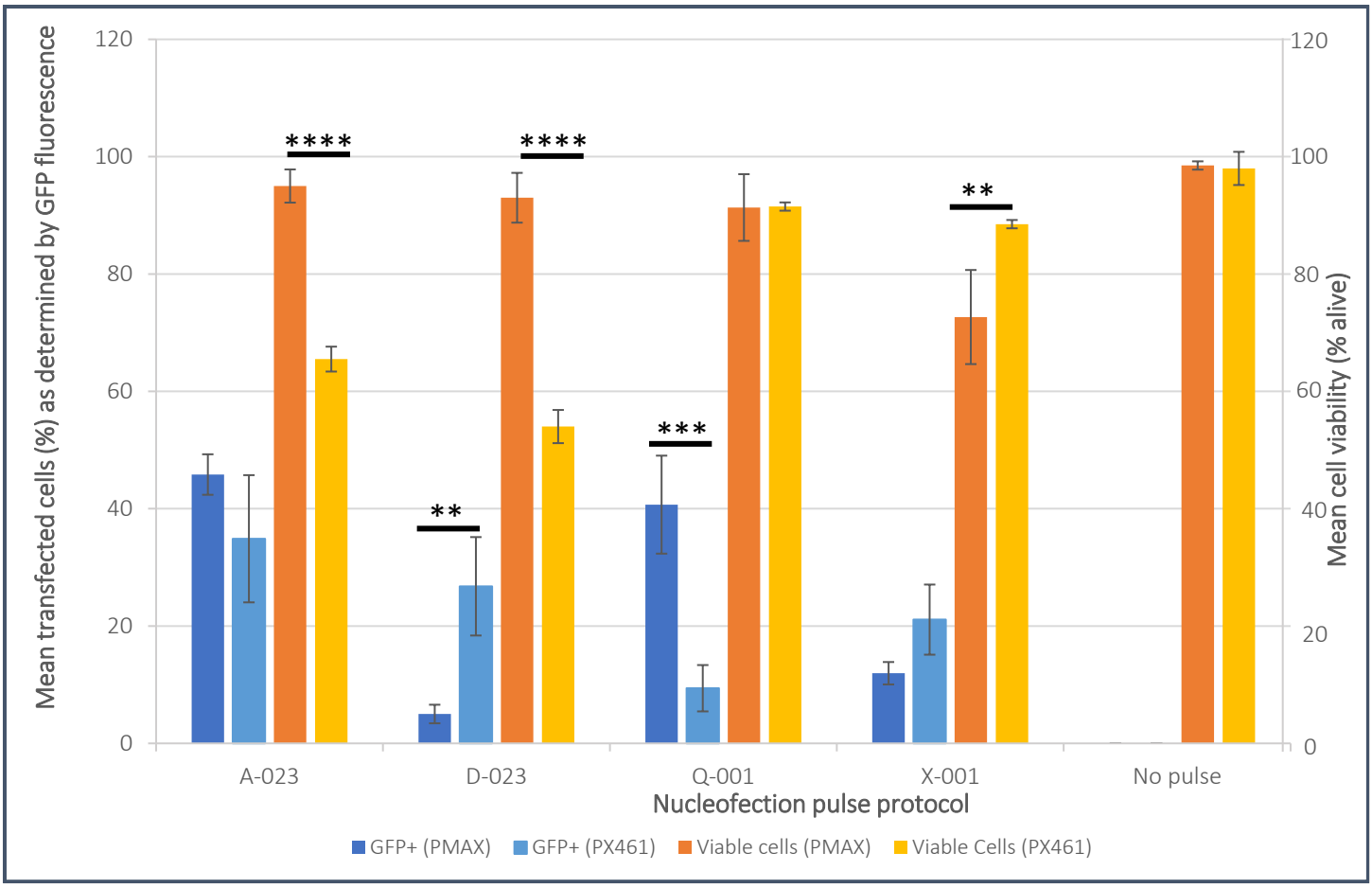


Figure 6.25: Mean transfection efficiency and cell viability of HEK293 cells using pmaxGFP and PX461 plasmids 24 hours post nucleofection. n=4. Optimisation of the most effective transfection protocol was carried out as per the manufacturer's instruction (Lonza). Transfection efficiencies determined by % of GFP expressing cells as visualised using the EVOS fluorescent microscope and quantified using ImageJ software. Cell viability determined by Trypan blue exclusion method. 1000000 cells per cuvette, with 1 µg pmaxGFP or 3 µg PX461 plasmid used in each cuvette. Mean +/- standard deviation. Statistical significance determined using multiple t-tests, with **** p <0.0001, **** p <0.001 and ** p <0.005

Following transfection, cell viability and transfection efficiency were analysed to determine the optimal pulse protocol for future experiments. Cells were cultured for 24 hours post-nucleofection and cell viability was assayed using Trypan Blue (**Figure 6.25**). There was a statistically significant increase in cell transfection between the pmaxGFP and PX461 plasmid using the D-023 pulse protocol (p=0.002), and interestingly, a highly significant decrease in GFP+ cells when the Q-001 pulse protocol was used (p=0.0005). No statistical difference was observed between the A-023 and X-001 protocols (p=0.102 and p=0.026 respectively).

Unfortunately, the pulse protocols with the greatest transfection efficiency also resulted in a statistically significant increase in cell death (protocols A-023 and D-023 for PX461 plasmid where p<0.0001), with cell viability down to 60 % and 55 % respectively (**Figure 6.25**). Based on the

combination of transfection efficiency and cell viability, A-023 pulse protocol was used for all further experiments (unless stated otherwise).

6.3.10 Transfection via nucleofection resulted in a greater proportion of cells containing a successful *UIMC1* edit

Following optimisation of the nucleofection protocol, modification of *UIMC1* was carried out using both knockout plasmids and nickases to create the specific *UIMC1*c.16090T>C variant. Cell populations were screened 24 hours post-nucleofection for successful gene modification.

6.3.10.1 Screening cell populations for successful edits in exon 2 of *UIMC1*

Figure 6.26 shows the results of the SURVEYOR assay conducted on cell populations that were transfected by nucleofection with the PX330 plasmids, which should result in a double stranded break in exon 2 of *UIMC1.* As shown in the positive lanes for samples 3 and 4, a distinct band can be observed, indicating the formation of heteroduplexes which have been cleaved in the presence of the SURVEYOR nuclease. These bands are more prominent than previous attempts at screening cell populations.

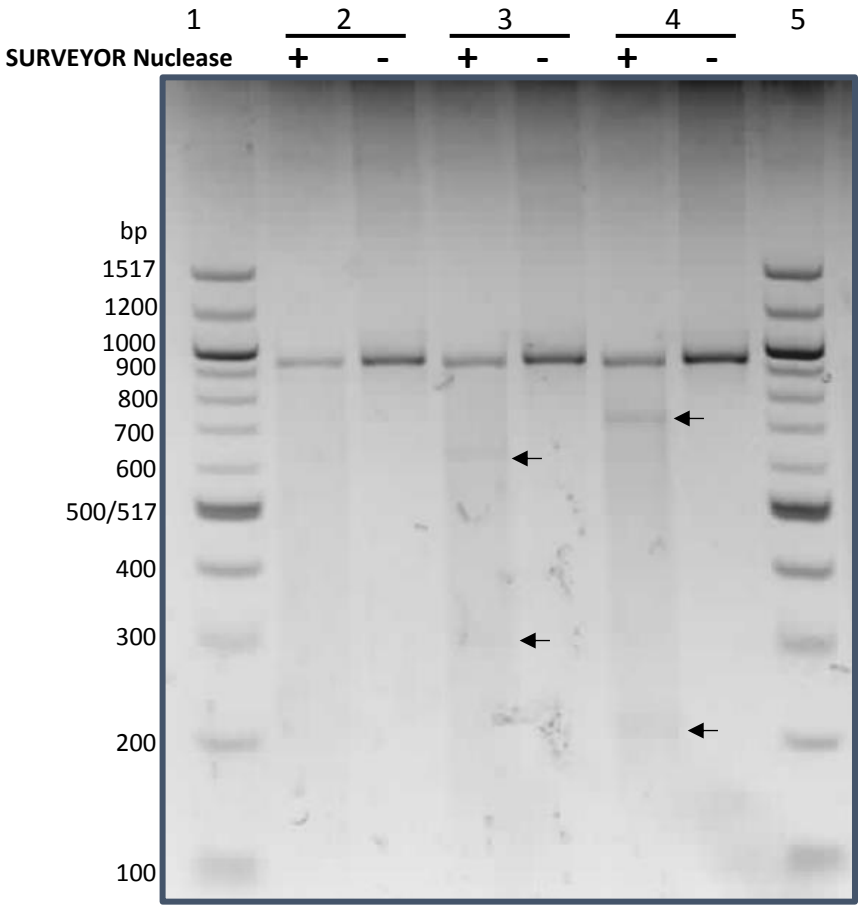


Figure 6.26: SURVEYOR assay on HEK293 cells transfected by nucleofection to introduce mutations within exon 2 of *UIMC1*. Heteroduplexes formed between WT and edited HEK293 cell population PCR products and analysed using SURVEYOR assay for mismatches in DNA templates. Lanes 1,5; NEB 100 bp ladder, Lane 2; PX330- (CRISPR sham), Lane 3; PX330-A knockout (KO) plasmid, Lane 4; PX330-B KO plasmid. + indicates presence of SURVEYOR enhancer and nuclease, - indicates no nuclease or enhancer included in reaction. Faint bands observed and their presence is indicated by arrows.

6.3.10.2 Screening transfected cell populations for introduction of the UIMC1:c.1690T>C variant.

The SURVEYOR assay was carried out on the cell populations for the *UIMC1* exon 13 premature truncation plasmid and the nickase plasmid pair. Minimal signs of editing were observed (**Figure 6.27**). A faint band was observed in lane 2 (as indicated by the arrow) indicating successful editing of a small proportion of the cell population with the PX330 plasmid. However, no heteroduplex formation was observed for the cells edited with the nickase plasmid pair (Lane 3). As this is a low sensitivity assay, both cell populations were plated for amplification of monoclonal cell lines. Initially the same amplicons were used for analysis of this region using the optimised primers utilised in **Figure 6.21**, however these primers began to result in the generation of multiple bands which could

not be resolved through troubleshooting. New primers were acquired however the issues persisted. As a result, new, smaller amplicons were designed to determine if successful modification of exon 13 in *UIMC1* occurred.

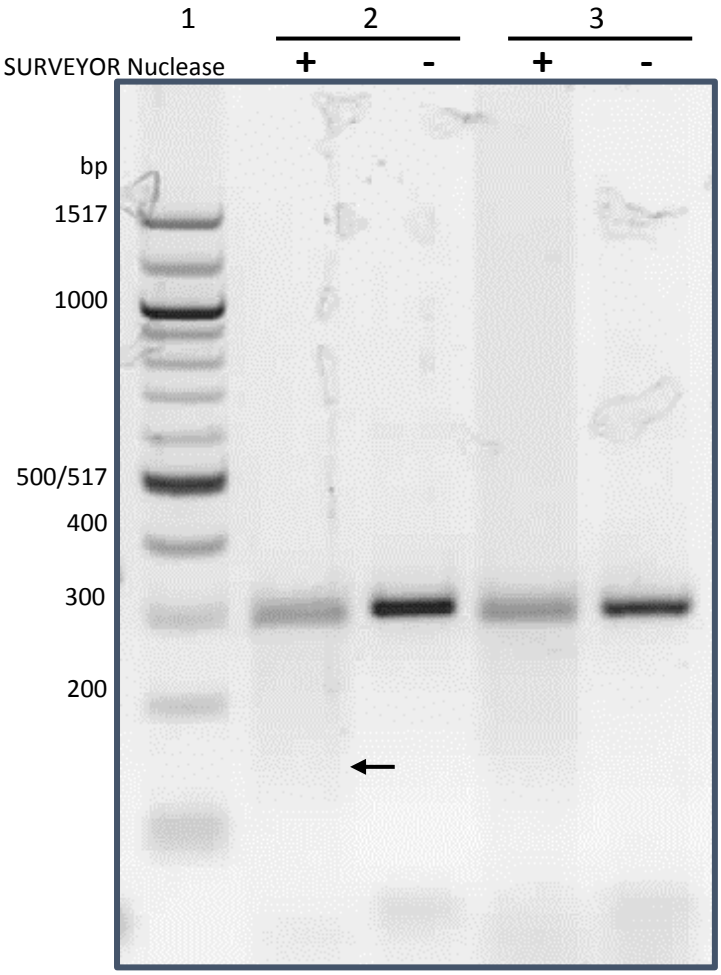


Figure 6.27: SURVEYOR assay on HEK293 cells transfected by nucleofection to generate mutations within exon 13 of *UIMC1*. Heteroduplexes formed between WT and edited HEK293 cell population PCR products and analysed using SURVEYOR assay for mismatches in DNA templates. Lane 1; NEB 100 bp ladder, Lane 2; PX330-KO exon 13 plasmid, Lane 3; PX462A+B +HDR template for introduction of UIMC1:c.1690T>C. + indicates presence of SURVEYOR enhancer and nuclease, - indicates no nuclease or enhancer included in reaction.

6.3.11 Monoclonal cell lines containing mutations within UIMC1 were successfully generated.

6.3.11.1 Paired nickases and HDR were not able to successfully introduce the potentially pathogenic UIMC1: c.1690T>C variant into HEK293 cells.

Overall, a total of 87 monoclonal cell lines were generated and cultured for the introduction of the C>T missense variant in *UIMC1*. The SURVEYOR assay was carried out on all 53 of these cell lines (analysis of 10 cell lines is shown in **Figure 6.28**). These results indicate that none of the analysed cell lines appeared to contain sequence differing from that of the wildtype sequence, as the SURVEYOR enzyme has not cut at the site of any mismatches formed in the heteroduplex. However, there was a decrease in band intensity associated for some samples (**Figure 6.28;** Lanes 5 and 7), indicating the possibility of heteroduplex formation. As a result, all samples were also analysed via Sanger sequencing. However, this indicated that all screened monoclonal cell lines were wildtype at the *UIMC1*:c.1690 location (data not shown). Unfortunately, the remaining cell lines (34 monoclonal cell lines) developed a fungal infection which could not be treated despite the use of fungizone and cells were disposed of.

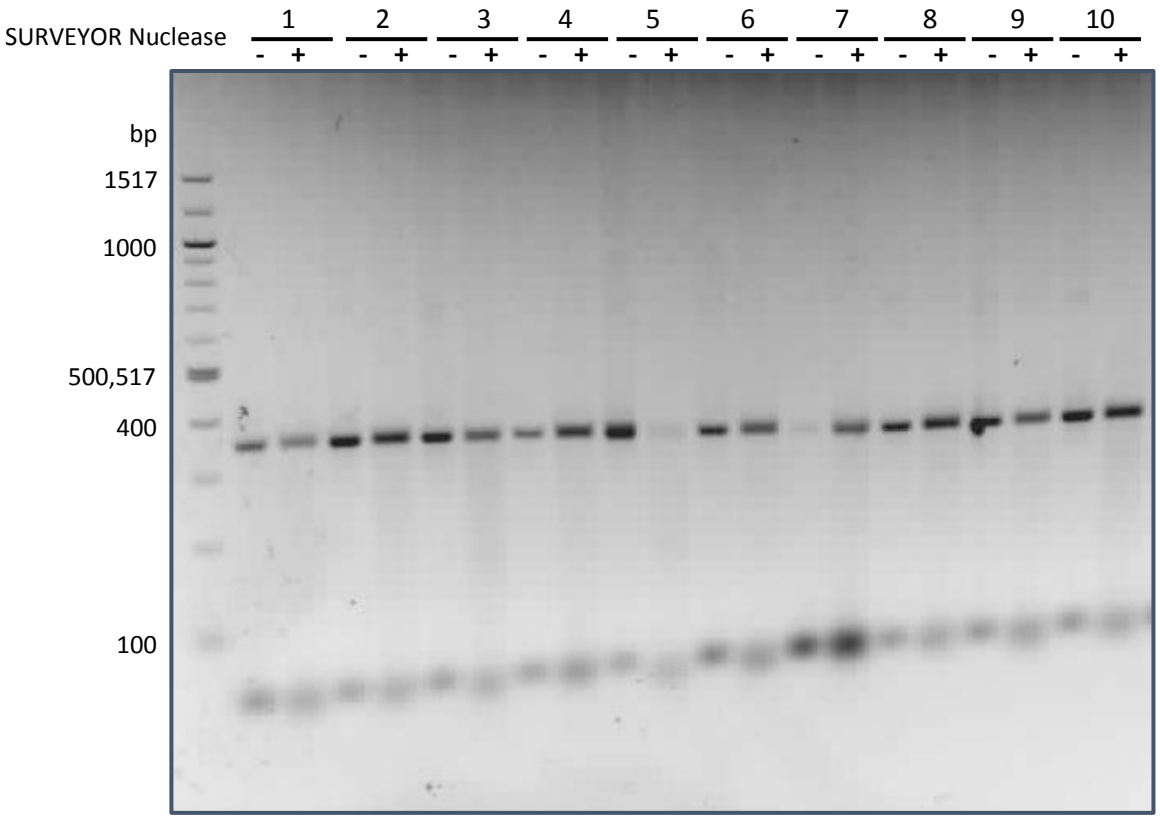


Figure 6.28: SURVEYOR assay of monoclonal cell lines generated through Nucleofection with the PX462A+B plasmids with the use of a HDR template. Each lane pair indicates a monoclonal cell line that was PCR amplified and analysed using the SURVEYOR assay, with the first lane indicating the presence of no SURVEYOR nuclease, and the second lane illustrating the results with SURVEYOR enhancer and nuclease.

Multiple attempts were made to introduce the 1690T>C mutation into *UIMC1*. Four different transfections via nucleofection attempts were carried out, with altering plasmid concentration and nucleofection pulse protocols (using pulse protocols A-023, D-023 and X-001) in an attempt to successfully edit the specific base in the HEK293 cell line. Throughout this process, distinct visible bands in the presence of the SURVEYOR nuclease could not be seen within any of the generated cell populations (very faint bands were often seen, indicating a low proportion of the cells were edited). Further attempts to generate monoclonal cell lines containing this edit were unsuccessful. Therefore, it appears that a method with a higher efficiency is required for the introduction of this missense mutation into cell lines.

6.3.11.2 Incorporation of nonsense and frameshift mutations into exon 2 of *UIMC1.*

Eighty-nine monoclonal cell lines were developed and screened for the knockout of *UIMC1* in exon 2. Twenty four cell lines were generated and maintained for KO exon2-A, and 65 cell lines were generated and maintained for KO exon2-B. Cells were screened for modification within exon 2 of *UIMC1* using the SURVEYOR assay (**Figure 6.29**). Cell lines that indicated the presence of *UIMC1* editing (as indicated by multiple bands observed in Panel **B**; Lanes 9, 13, 16, 21 and 22 of **Figure 6.29**) were subjected to Sanger sequencing for identification of mutagenesis (**Figure 6.30**) and to determine the potential effect on the resultant protein.

The chromatograms illustrated in **Figure 6.30** indicate the successful modification of exon 2 of *UIMC1* in the selected cells. Panel **A** indicates the sequence of the region of interest in the wildtype HEK293 cells with no modification. Comparison of this region in panels **B, C** and **D** of **Figure 6.30** illustrate the generation of *UIMC1* mutant cell lines, with the location of the introduced mutation varying slightly due to the nature of CRISPR/Cas9 editing and the NHEJ repair mechanism utilised for double stranded DNA cuts. All presented cell lines were subjected to protein quantification via western blot analysis before any further work was commenced (**Figure 6.33**)

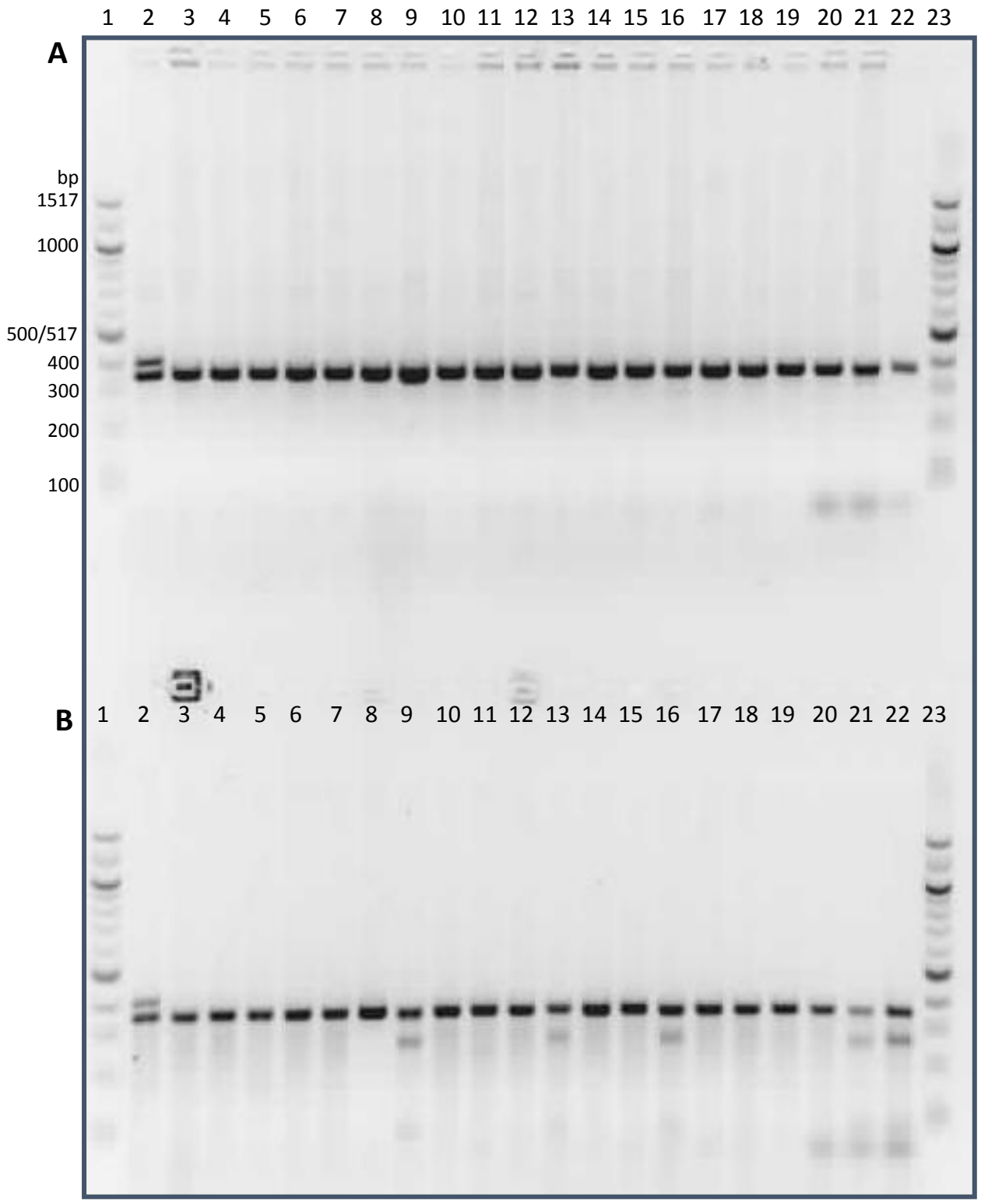


Figure 6.29: SURVEYOR assay carried out on monoclonal cell lines generated through nucleofection with PX330 ex2-B knockout plasmids. **A.** SURVEYOR assay run with no SURVEYOR nuclease or enhancer. **B.** SURVEYOR assay run with the inclusion of SURVEYOR nuclease and enhancer. Lanes 1 + 23; NEB 100 bp ladder. Lane 2; polyclonal cell population, Lanes 9, 13, 16, 21, 22 indicate three distinct bands in panel **B**.

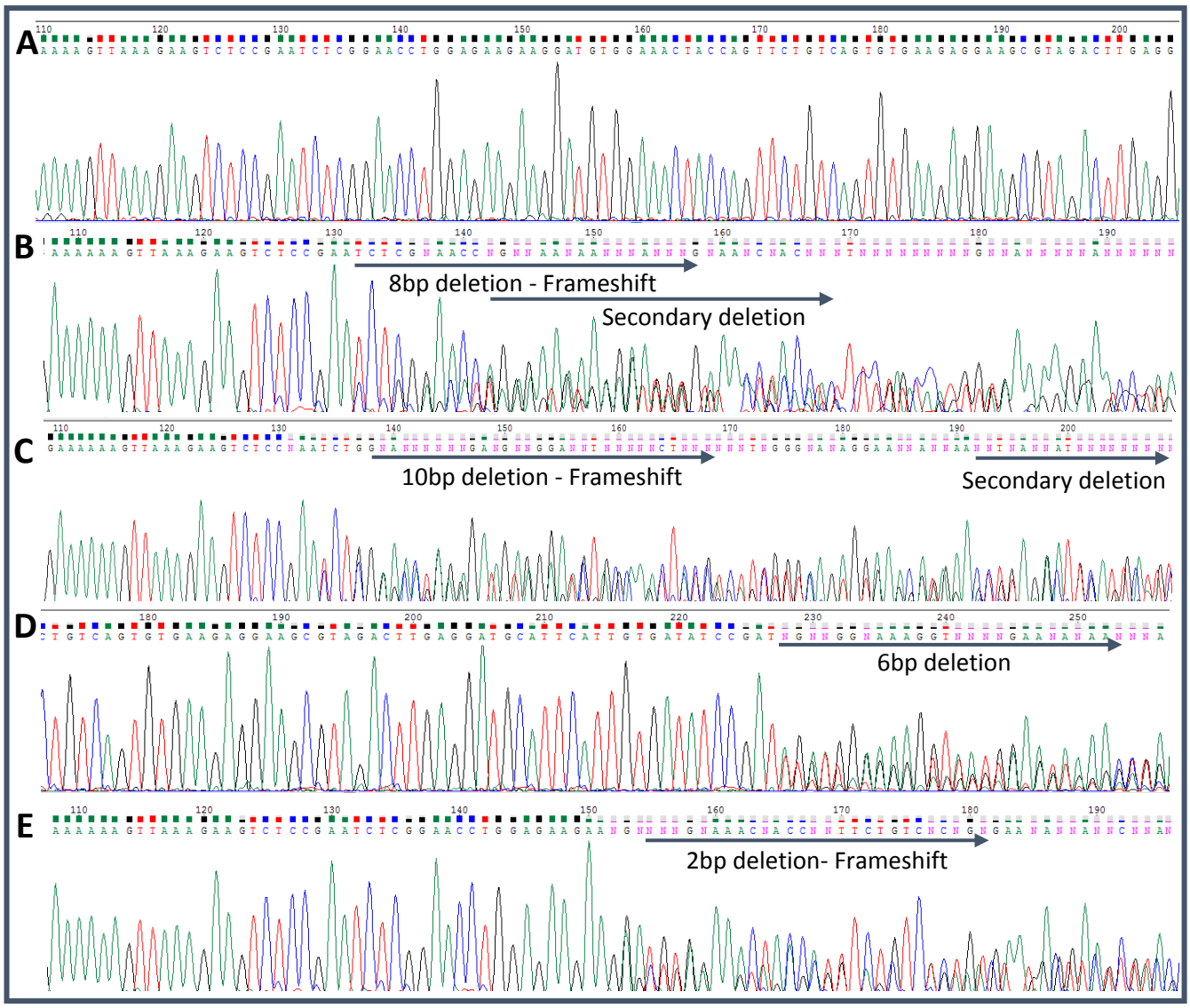


Figure 6.30: Chromatogram traces of exon 2 *UIMC1*-mutated monoclonal cell lines. All traces shown in the reverse direction, with identical results obtained for the forward sequence (data not shown). Point of mutation indicated by arrow with type of mutation shown. **A.** HEK293 (wildtype) **B.** e2-B1.14 indicates an 8bp deletion resulting in a frameshift from base 132 as indicated on the chromatogram, with a secondary deletion identified on the other allele at base 142. **C.** e2-B1.15 indicates a 10bp frameshift deletion from base 138 on the chromatogram, with a secondary deletion identified on the second allele from base 193. **D.** e2-B1.16 indicates a 6bp deletion on one allele from base 228 of the chromatogram trace. **E.** e2-B3.1 indicates a 2bp deletion resulting in a frameshift from base 152 of the chromatogram trace.

Deconvolution of the sequencing traces of the introduced mutations identified the generation of two homozygous knockout cell lines (panels **B** and **C**, **Figure 6.30**). Cell line e2-B1.14 showed the incorporation of two mutations within exon 2 of *UIMC1*. Within this cell line, there was the introduction of an 8bp deletion in one allele and a secondary deletion in the other allele. A second cell line containing mutations within both alleles was identified (e2-B1.15), with a 10bp deletion in one allele, and a secondary mutation within the remaining allele. Subsequent analysis of protein expression identified the introduced mutations in both cell lines resulted in a complete loss of

UIMC1 expression (**Figure 6.33**). Furthermore, this analysis also identified a 6bp deletion in one cell line (e2-B1.16) resulting in the in-frame deletion of two amino acids, threonine and isoleucine (panel **D, Figure 6.30**, confirmed by western analysis in **Figure 6.33**). Additionally, a heterozygous cell line was generated, with a 2bp deletion within one allele identified in e2-B3.1 (panel **E**, **Figure 6.30**). Results of Sanger sequencing analysis are summarised in **Table 6.3**.

Table 6.2: Summary table of sequence changes identified in screened monoclonal cell lines generated through CRISPR/Cas9 modification of exon 2 of UIMC1.

Cell Line	Sequence changes observed	Result on produced protein
HEK293	Nil	Nil
PX330- (CRISPR Sham)	Nil	Nil
e2-B1.14	8bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation
e2-B1.15	10bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation
e2-B1.16	6bp deletion in allele one	Deletion of Threonine and Isoleucine residues
e2-B3.1	2bp deletion in allele one	Frameshift, suspected premature truncation of one allele

6.3.11.3 Successful incorporation of knockout mutations in exon 13 of *UIMC1*.

Seventy-six monoclonal cell lines were developed and maintained for the premature truncation of *UIMC1* in exon 13. Thirty-six cell lines were screened for modification within exon 13 of *UIMC1* and cells that indicated the presence of an edit were subjected to further analysis via Sanger sequencing. Panel B indicates a large proportion of generated cell lines screened within this run of the assay contained indels within exon 13, with Lane 10 indicating a large insertion (approximately 150bp; **Figure 6.32**, panel **D**) and differing sizes of the digested products indicating the various nature of the introduced deletions within the generated cell lines. The majority of the screened cells displayed 3 bands, indicating a strong product at 300bp (full size product) and two smaller products at 200bp and 100bp (Lanes 3, 5-8, 11-18; **Figure 6.31**) .

The chromatograms illustrated in **Figure 6.30** show successful edits in exon 13 of UIMC1 within the sequenced cell lines. Panel **A** shows the sequence of the region of interest in the PX330- cell line, indicating wildtype sequence with no modification. Comparison of this region in panels **B, C** and **D** of **Figure 6.32** illustrate the generation of mutant cell lines, with the location of the introduced mutations varying slightly. For panel **B** of **Figure 6.32**, the introduced 3bp deletion results in the deletion of a cysteine residue and does not alter the reading frame of the protein. The effect of the

introduced sequence variants on protein expression was determined via western blot analysis prior to any further functional analysis (**Figure 6.33**).

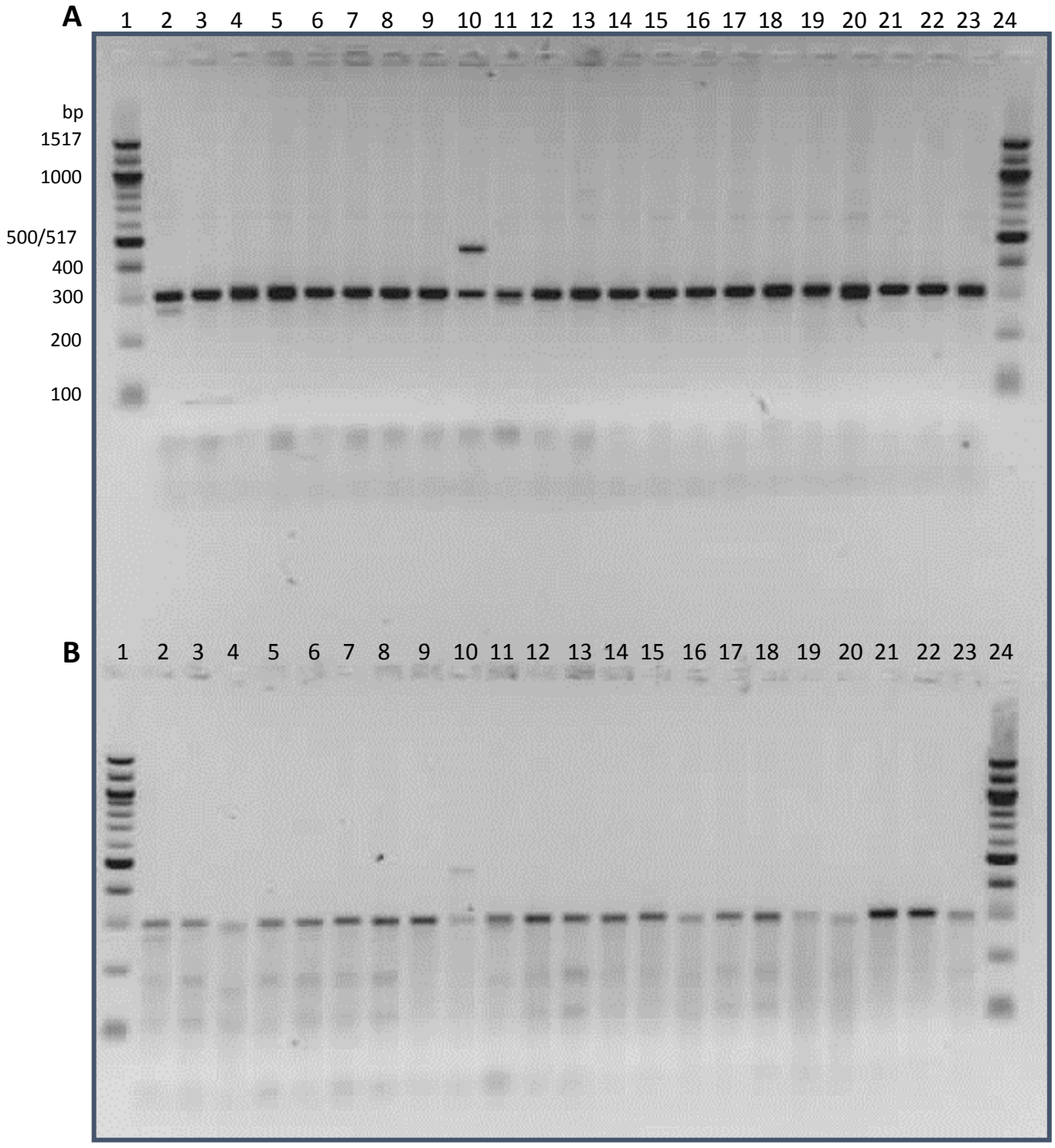


Figure 6.31: SURVEYOR assay carried out on monoclonal cell lines generated through nucleofection with PX330 exon 13 knockout plasmids. Gel electrophoresis run on 2.5 % agarose gel containing GelRed. **A.** SURVEYOR assay run with no SURVEYOR nuclease or enhancer. **B.** SURVEYOR assay run with the inclusion of SURVEYOR nuclease and enhancer. Lanes 1 + 24; NEB 100 bp ladder. Lane 2; polyclonal cell population, Lanes 2-13, 17,18, indicate three distinct bands in panel **B**

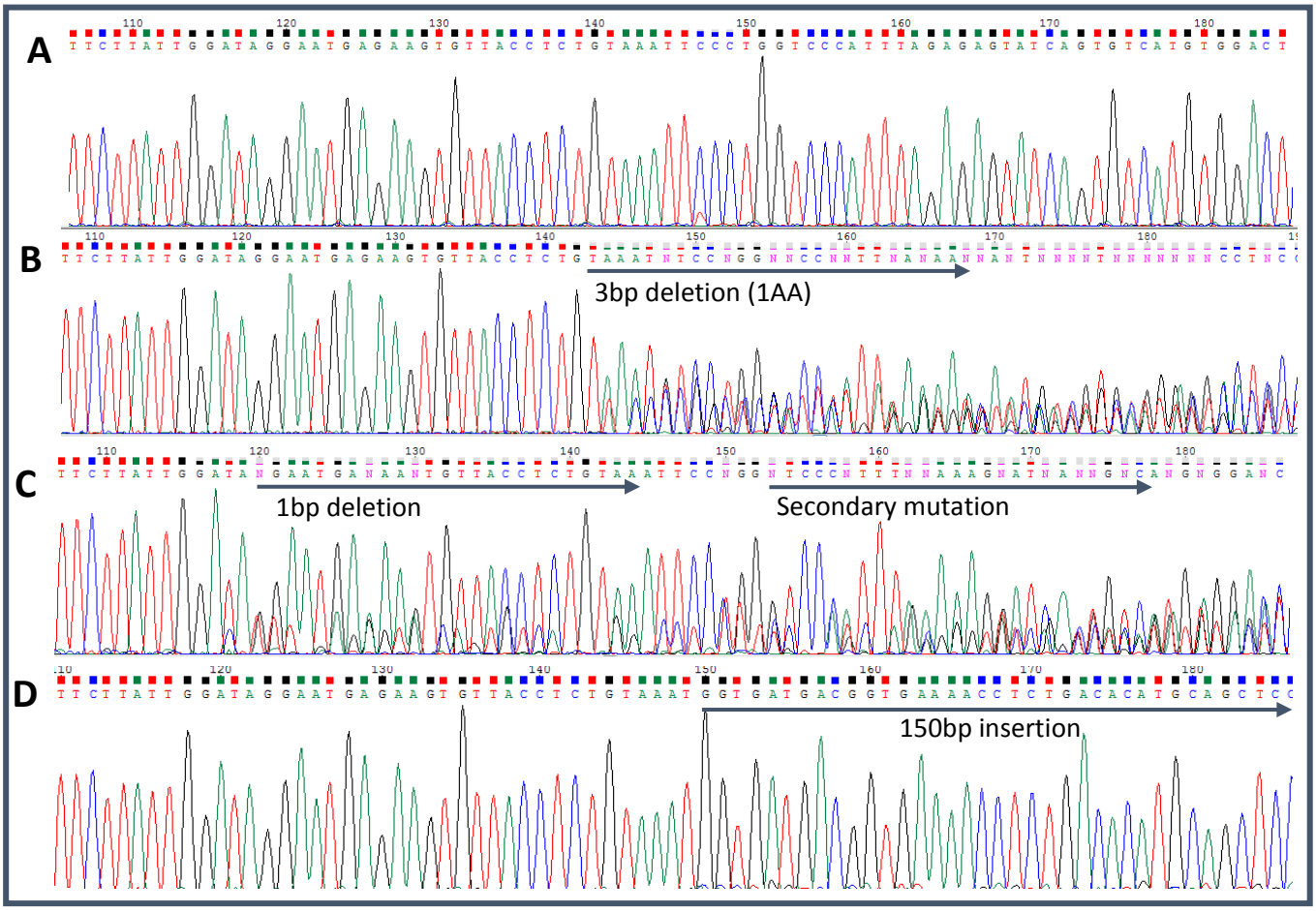


Figure 6.32: Chromatogram traces of exon 13 *UIMC1*-mutated monoclonal cell lines. All traces shown in the forward direction, with identical results obtained for the reverse sequence (data not shown). Point of mutation indicated by arrow with type of mutation shown. **A.** HEK293 (wildtype) **B.** e13-KO1 indicates a 3bp deletion from base 144 as indicated on the chromatogram. **C.** e13-KO2 indicates a 1bp deletion, resulting in a frameshift deletion from base 120 on the chromatogram, with a secondary mutation on the other allele introduced from base 154. **D.** e13-KO10 indicates a 150bp homozygous insertion from base 150 of the chromatogram trace.

Deconvolution of the Sanger sequencing results indicated that mutations were successfully introduced into the cells. A heterozygous cell line with a 3bp deletion of a single amino acid in one allele (panel **B**, **Figure 6.32**) of exon 13 was generated. Western blot analysis demonstrated an intense band at the expected size for this cell line, supporting the deletion of a single amino acid, rather than altering the reading frame (panel **B**, **Figure 6.33**). A cell line containing mutations within both alleles was identified (panel **C**, **Figure 6.32**) with a 1bp deletion in one allele, resulting in a frame shift, with a secondary deletion identified downstream in the second allele. Analysis of protein expression of UIMC1 through western blot analysis illustrated a reduction in expression, rather than a complete loss (panel **B**, **Figure 6.33**). Inexplicably, a 150 bp insertion was identified in both alleles of the e13-KO10 cell line (panel **D**, **Figure 6.32**). Due to the homozygous appearance of the insertion, it is unlikely that both alleles were modified in the same way via CRISPR/Cas9 editing.

It is possible that one allele was modified to include the 150bp insertion and the other allele was deleted. This would result in the homozygous appearance that was observed in Panel **D**, **Figure 6.32**. Summary of the sequencing analysis of the exon 13 modified cells is included in **Table 6.3**.

Table 6.3: Summary table of sequence changes identified in screened monoclonal cell lines generated through CRISPR/Cas9 modification of exon 13 of UIMC1.

Cell Line	Sequence changes observed	Result on produced protein
HEK293	Nil	Nil
PX330- (CRISPR Sham)	Nil	Nil
e13-KO1	3bp deletion in allele one	Deletion of cysteine residue (566)
e13-KO2	1bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation
e13-KO10	150bp insertion in allele one Unknown mutation in allele two	Unknown, located within intronic region of UIMC1

6.3.12 Reduction of UIMC1 protein in CRISPR edited monoclonal cell lines was verified through western blot analysis.

Prior to any further functional analysis, expression of UIMC1 protein in the modified cell lines was analysed via western blot analysis. UIMC1 is localised within the nucleus of cells, therefore whole protein extracts were carried out on 4 monoclonal cell lines with mutations within exon 2 and 3 monoclonal cell lines with mutations within exon 13. Protein levels were analysed via western blot quantification of protein expression in modified and wildtype cells (panel **B**, **Figure 6.33**). Comparison of the band intensities between the wildtype HEK293, negative control (PX330-) and modified cell lines indicated that the incorporation of some UIMC1 mutations resulted in a reduction of UIMC1 protein expression.

The polyacrylamide gel image in panel **A**, **Figure 6.33** demonstrates that the majority of lanes show equal loading, with the exception of lanes 2 and 4 which appear to have less protein. This under-loading was taken into consideration in the analysis of the western blot. As this was not a quantitative analysis, the presence or absence of the protein was still able to be determined despite this skew in protein loading. From the band intensities for UIMC1 observed in panel **B**, **Figure 6.33**, there are two bands observed for e2-B1.16, e2-B3.1 and e13-KO10. UIMC1 is expected to result in a band at 80kDa, which is visible in most samples (to varying intensities). However, there are both higher and lower molecular weight bands observed within the samples analysed.

Both e2-B1.14 and e2-B1.15 lack the presence of the 80kDa protein, but a faint lower molecular weight product is observed. This corresponds with the Sanger sequencing data, which indicated the presence of frameshift mutations within both alleles of *UIMC1* (**Table 6.4**). The cell lines e2-B3.1 and e13-KO2 both indicated a reduction in the 80kDa UIMC1 product through a significant difference in band intensities when compared to both the wildtype and negative control. This corresponds to the Sanger sequencing data for these cell lines, with a reduction in protein expression observed in comparison to the wildtype cells. The secondary mutation within the second allele of e13-KO2 was unable to be deconvoluted but could be a mutation that does not result in a frameshift mutation, as a reduction in expression was observed, rather than a complete knockout. Additionally, a lower molecular weight product was also observed in the e2-B3.1 cell line. This lower weight product could be attributed to various splice variants of the UIMC1 transcript (See **Section 6.4**). Sample e2-B1.16 illustrated the presence of a strong band at 80kDa, which was of similar intensity to the wildtype and negative control, indicating minimal reduction in UIMC1 protein levels following CRISPR/Cas9 modification. Therefore, this cell line was not used for further functional analysis.

Analysis of sequencing indicates deletion of a cysteine residue in one allele in e13-KO1, therefore you would not expect to see any loss or change in protein expression through western analysis (panel **B**, **Figure 6.32**). This is the closest we were able to get to the introduction of the variant within the ZFN of UIMC1. Cell lines e13-KO1 and e13-KO10 also displayed a strong band at 80kDa, however this was not unexpected as this antibody is known to bind within the first 100 amino acids of the UIMC1 protein, and modification within these cell lines was carried out in exon 13 (residues 510 – 567). However, a truncation in the remaining UIMC1 sequence would be expected to result in a shift in the product, as the final 100 amino acids would be truncated, which should be observed as a loss of approximately 10kDa. As this shift was not observed, it is unlikely that either cell line contains a mutation resulting in premature truncation. All three exon 13 mutated cell lines were initially selected for analysis, however both e13-KO2 and e13-KO10 failed to proliferate and were unable to be analysed further (Refer to **Section 6.4**).

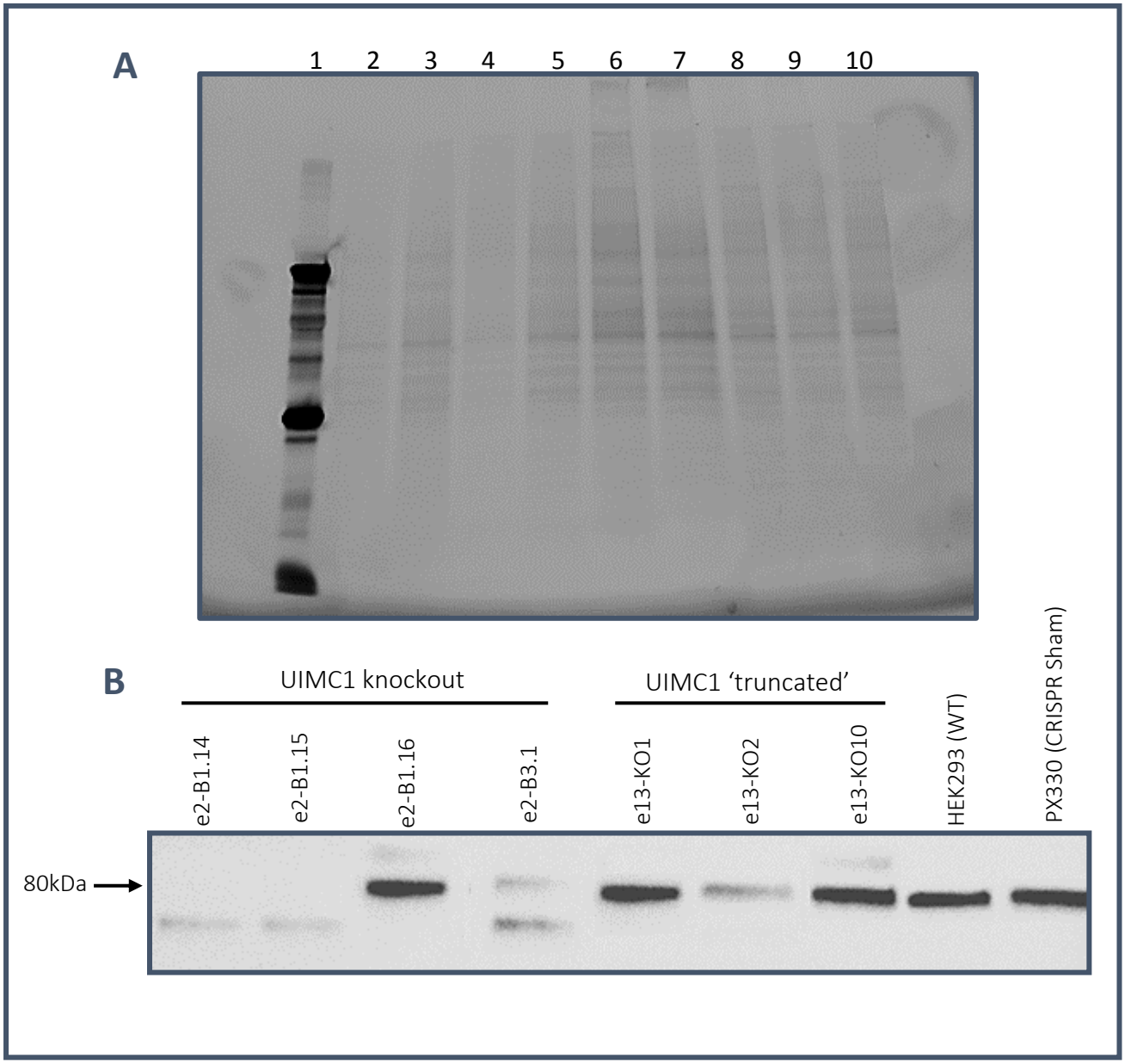


Figure 6.33: CRISPR/cas9 deletion in *UIMC1* results in a reduction in UIMC1 protein levels. A. Stain-free Polyacrylamide gel with 10 µg total protein lysate loaded for each sample. Lane 1**;** broad range (10-250 kDa) Precision Plus Protein standard, Lanes 2-5, total protein lysate from *UIMC1* exon 2 modified cell lines Lane 2; e2-B1.14, Lane 3; e2-B1.15, Lane 4; e2-B1.16, Lane 5; e2-B3.1 Lanes 6-8, total protein lysate from *UIMC1* exon 13 modified cell lines, Lane 6; e13-KO1, Lane 7; e13-KO2, Lane 8; e13-KO10, Lane 9; HEK293 (wildtype), Lane 10; PX330- (CRISPR Sham, negative control) **B.** Western blot of wildtype HEK293 cells, PX330- transfected cells (CRISPR sham) and monoclonal cell lines generated from transfection with PX330 CRISPR/Cas9 plasmids introducing indels in either exon 2 or exon 13 of *UIMC1.* Expected product size 80kDa.

Table 6.4: Summary of Sanger sequence and western blot analysis of UIMC1 modified cells generated through CRISPR/Cas9 modification. MW; Molecular weight, kDa; Kilo Daltons, WT; wildtype, NMD: Nonsense mediated decay

Cell Line	Sequence changes observed	Result on produced protein	Observed effect on protein via western blot
HEK293	Nil	Nil	Nil
PX330- (CRISPR Sham)	Nil	Nil	Nil
***UIMC1* exon 2 modified cells**			
e2-B1.14	8bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation	Loss of expression, homozygous knockout
e2-B1.15	10bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation	Loss of expression, homozygous knockout
e2-B1.16	6bp deletion in allele one	Deletion of Threonine and Isoleucine residues	Strong product at 80kDa observed, similar intensity to WT controls, in frame deletion on one allele
e2-B3.1	2bp deletion in allele one	Frameshift, suspected premature truncation of one allele	Reduction in expression of product at 80kDa, heterozygous knockout
***UIMC1* exon 13 modified cells**			
e13-KO1	3bp deletion in allele one	Deletion of cysteine residue (566)	Strong product at 80kDa observed, similar intensity to WT controls – In frame deletion on one allele
e13-KO2	1bp deletion in allele one Secondary deletion in allele two	Frameshift, suspected premature truncation	Reduction in expression of product at 80kDa, Potential heterozygous knockout, Candidate for NMD?
e13-KO10	150bp insertion in allele one Unknown mutation in allele two	Unknown, located within intronic region of *UIMC1*	Strong product at 80kDa observed, similar intensity to WT controls.

6.3.13 Mutation of *UIMC1* resulted in variable changes in rates of cell proliferation.

As uncontrolled cell growth is one of the hallmarks of cancer development, the proliferation rate of monoclonal cell lines was analysed. Cells were plated at a seeding density of 5000 cells and the number of cells was counted every 24 hours for a period of 7 days (**Figure 6.34**).

As illustrated in **Figure 6.34**, there was no statistical significance observed between the proliferation of the wildtype HEK293 (dark blue line) and PX330- (CRISPR sham, orange line). The monoclonal cell line e2-B1.15 proliferated rapidly, however, this was only statistically significant at 48 hours (p=0.037) and 168 hours (p=0.0002) when compared to the controls. The other exon 2 mutated cell

line (e2-B3.1) demonstrated the opposite trend, with cell proliferation significantly decreased at time points 120, 144 (where p<0.0001) and 168 hours (p=0.0001) when compared to both the WT and negative control. Cell proliferation of the exon 13 (e13-KO1) mutated cell line showed a similar trend to that of the control, with no statistically significant difference observed due to large variation between replicates.

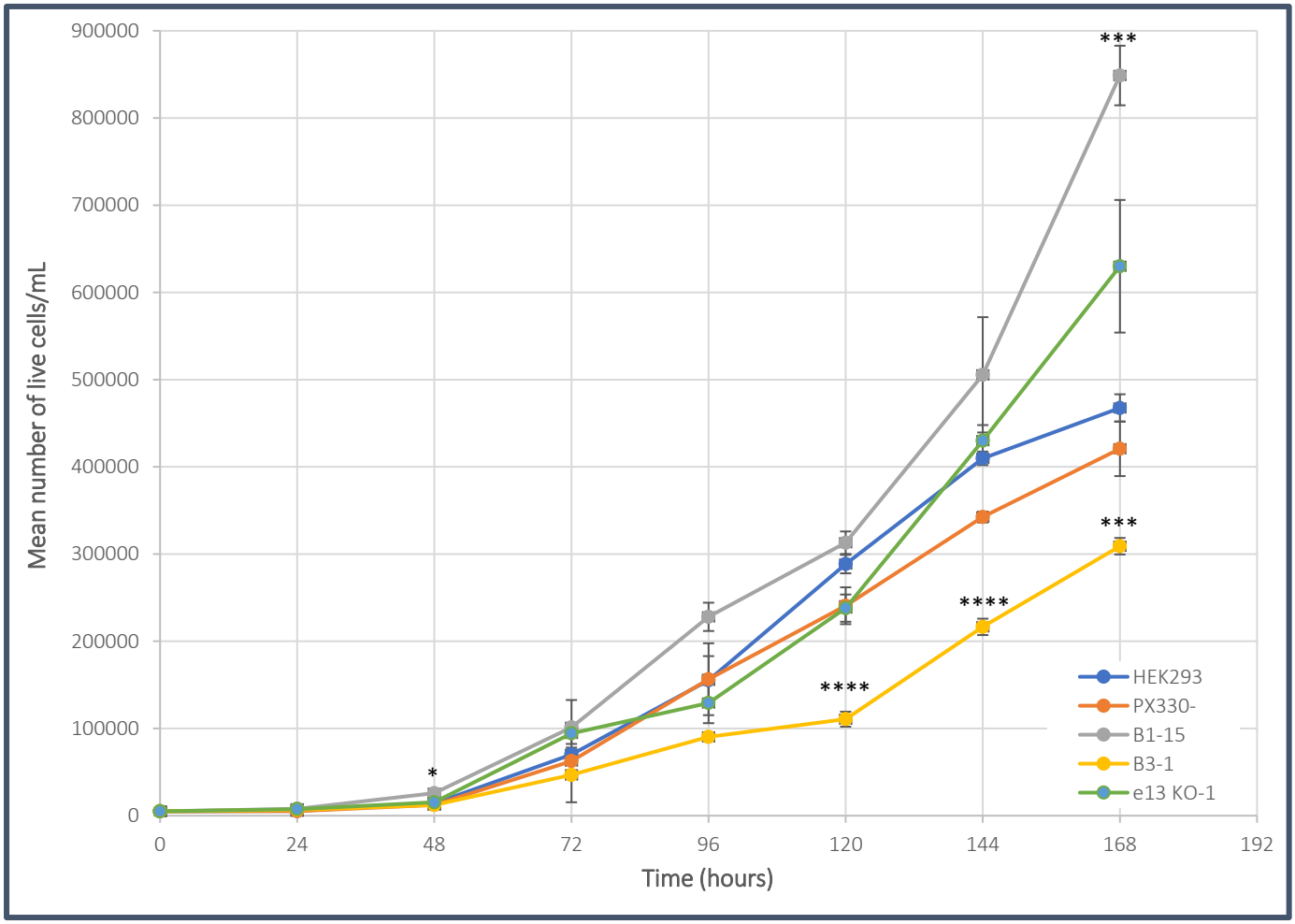


Figure 6.34: Growth curve of *UIMC1*-modified, wildtype and negative control HEK293 cells over 7 days. n=4. Cells seeded at starting density of 5000 cells. Mean cell growth determined by Trypan blue exclusion assay. HEK293; WT, PX330-; CRISPR Sham, e2-B1.15; UIMC1 homozygous knockout, e2-B3.1; UIMC1 heterozygous knockout, e13-KO1; UIMC1 with 1 AA deletion in ZFD. Mean +/- standard deviation, Statistical analysis carried out using two-way ANOVA, where *p<0.05, ***p<0.001 and **** p<0.0001.

6.3.14 *UIMC1*-mutated cells show a reduction in number of γH2AX foci and a delay to repair DNA double stranded breaks induced by irradiation.

DNA damage initiates a series of cellular events that function to control DNA repair, gene transcription, and cell cycle checkpoints. These DNA signalling responses function to maintain genomic integrity through the recruitment of signalling and repair proteins to the site of DNA damage. *UIMC1*-mutant cells, wildtype HEK293 and PX330- cells were exposed to two mechanisms that induce DNA double stranded breaks; 250µM doxorubicin hydrochloride (DOX) or 2 Gray ionising

radiation (IR). The ability of cells to repair DNA damage was assessed through the γH2AX induction. Phosphorylation of serine^{139} within the protein is integral to the formation of γH2AX foci. The number of foci corresponding to the amount of DNA damage is indicative of the cells ability to repair DNA damage. Using antibodies against the phosphorylated serine^{139} residue, it is possible to assess cellular differences in DNA DSB repair capacity over a 24-hour period.

The viability of UIMC1-modified cells was assessed following a 24-hour incubation with 250µM Doxorubicin hydrochloride (**Figure 6.35**). Viability of cells exposed to ionising and sham irradiation was assessed for all 5 cell lines at 1-, 4- and 24-hours post irradiation (**Figure 6.36**). Two-way ANOVA analysis was carried out on all cell lines and no statistical difference was observed between sham irradiated and 2Gy irradiated cell lines at any time point (**Figure 6.36**). The e2-B3.1 cell line (heterozygous knockout) displayed a decreased cell viability in comparison to all other cell lines, however this was not attributed to the exposure to radiation or doxorubicin as it was observed for all treatments, at every time point. There was approximately a 10% decrease in cell viability of WT and negative control cell lines when exposed to IR (orange and blue lines). However, when UIMC1 mutant cell lines were exposed to IR, this decrease in cell viability was slightly greater (approximately 15%), indicating cells lacking UIMC1 demonstrate an increased sensitivity to IR.

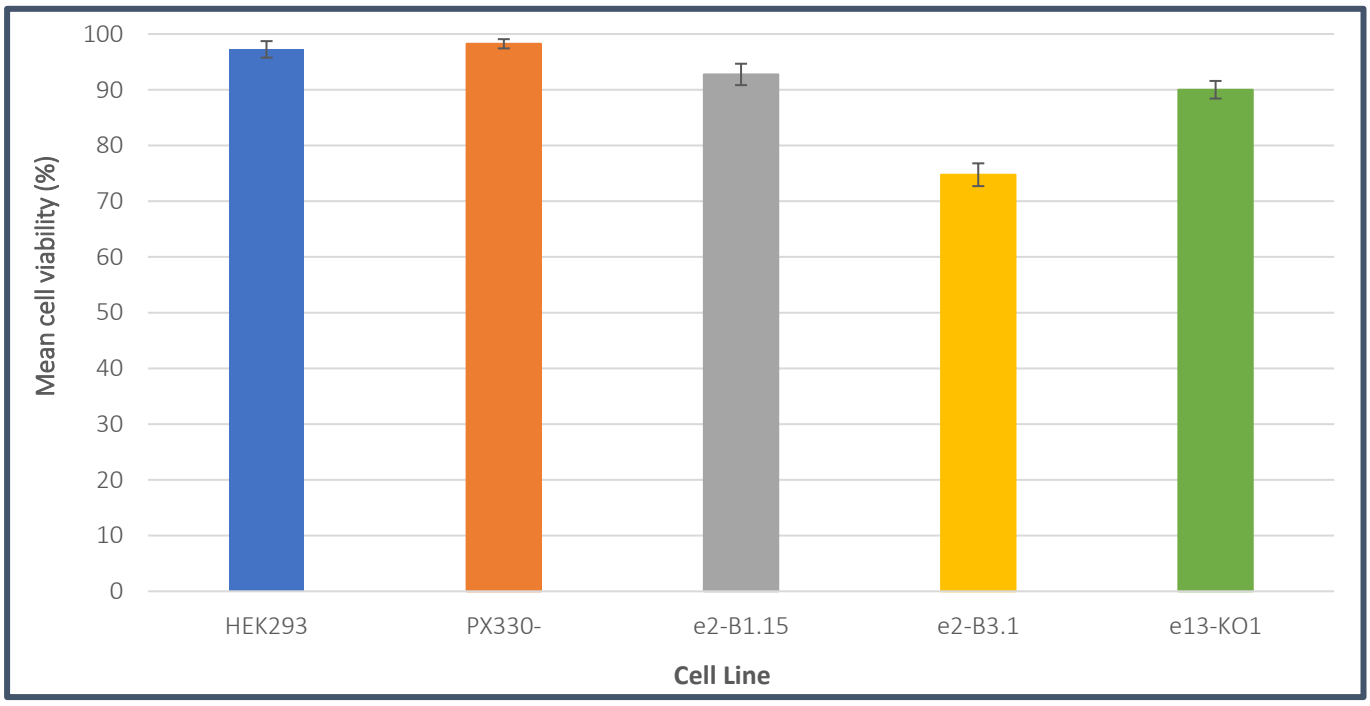


Figure 6.35: Mean cell viability following exposure to 250 µM Doxorubicin for 24 hours on UIMC1-modified, wildtype and negative control HEK293 cells. n=4. Mean cell viability determined by Trypan blue exclusion assay. Mean +/- standard deviation. HEK293; WT, PX330-; CRISPR Sham, e2-B1.15; UIMC1 homozygous knockout, e2-B3.1; UIMC1 heterozygous knockout, e13-KO1; UIMC1 with 1 AA deletion in ZFN

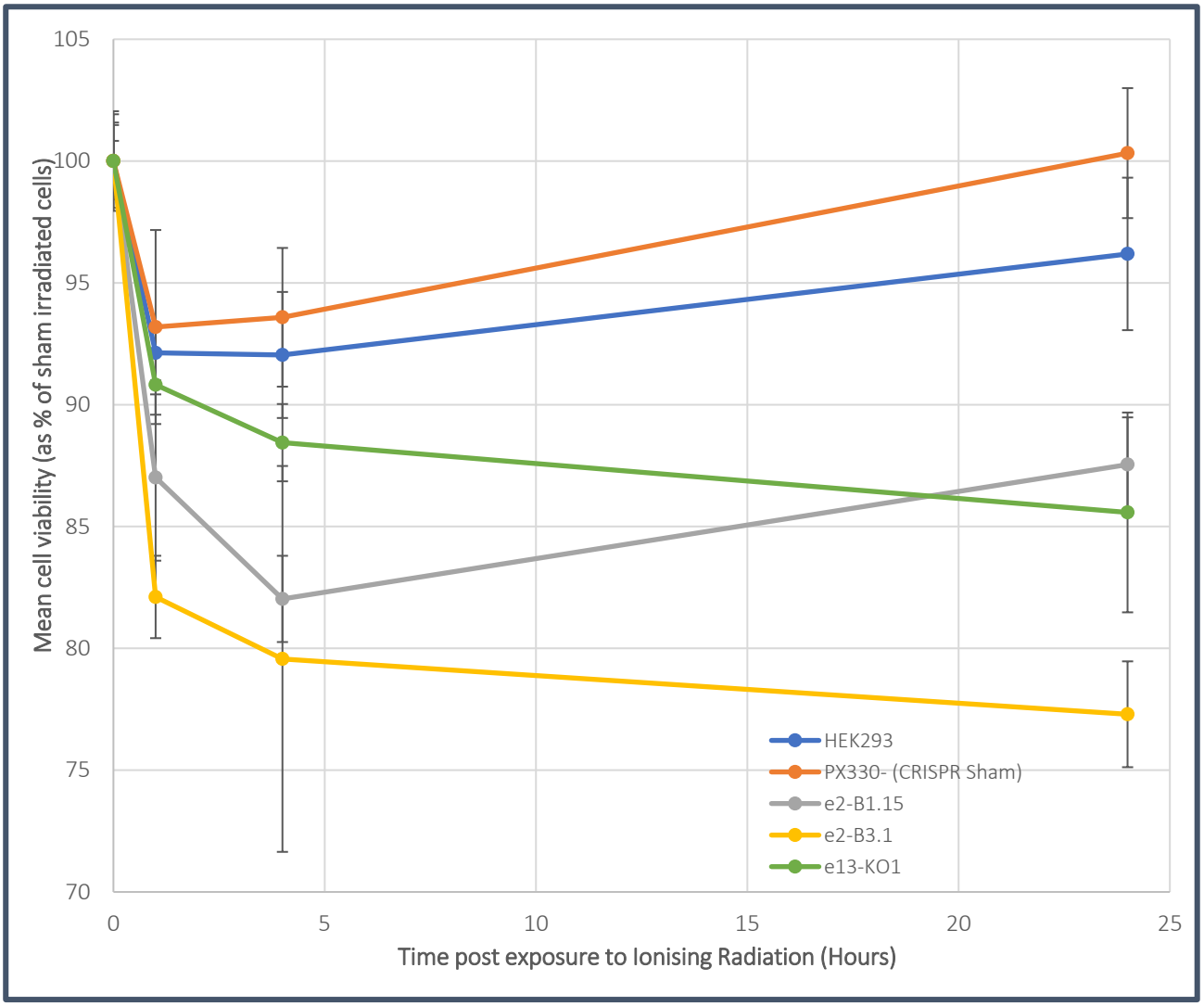


Figure 6.36: Mean cell viability following exposure to 2Gy/Sham irradiation on UIMC1-modified, wildtype and negative control HEK293 cells over 24 hours. n=4. Viability of 2Gy irradiated cells expressed as a percentage of viability of sham irradiated cells. Mean cell viability determined by Trypan blue exclusion assay. HEK293; WT, PX330-; CRISPR Sham, e2-B1.15; UIMC1 homozygous knockout, e2-B3.1; UIMC1 heterozygous knockout, e13-KO1; UIMC1 with 1 AA deletion in ZFN.

The mean number of γH2AX foci observed in the cell lines was quantified using the ISX imaging flow cytometer. The number of nuclear foci observed was determined for 3 time points (1, 4 and 24 hours) which were selected based on previous studies (Horn *et al.*, 2011, Mariotti *et al.*, 2013) and compared to Doxorubicin induced DNA DSBs (**Figure 6.38**). The wildtype HEK293 and negative control PX330- cell lines depicted a significant increase in the number of nuclear foci observed at 1-hour post irradiation when compared to the sham control. The number of nuclear foci is slightly elevated at the 4-hour time point in these cell lines, returning to baseline by 24 hours. In addition, in HEK293 and PX330- cell lines, less nuclear foci were induced by Ionising radiation than by Doxorubicin.

Analysis of the *UIMC1*-mutated cell lines indicates an overall trend of a significantly decreased number of nuclear foci in response to ionising radiation, with cell lines taking 24 hours to reach maximal H2AX foci. The *UIMC1*-mutated cell lines indicated a slight increase in H2AX foci number 1-hour post irradiation (when compared to the sham control), however, maximum foci were observed 24 hours post irradiation. Interestingly, nuclear foci observed in the *UIMC1*-mutated cell lines appeared larger, but fewer, in comparison to those observed in HEK293 and PX330- cells (**Figure 6.37**).

As shown in **Figure 6.38**, the standard deviation for assessing the number of nuclear foci is quite large. This is due to the considerable variation observed in the number of nuclear foci between replicates (refer to **Appendix J** for raw nuclear foci count plots). An increase in biological replicates may help to reduce this variation. Despite this, a statistically significant difference in DNA repair capacity (as quantified through γH2AX foci) was observed in wildtype and UIMC1 mutant cells exposed to DNA damaging agents. This was highly prominent in cells treated with 250µM Doxorubicin for 24 hours ($p=<0.0001$ for e2-B1.15 and e2-B3.1 and e13-KO1 in comparison to HEK293). Additionally, a significant increase in nuclear foci was observed in HEK293 cells 1-hour post 2 Gy irradiation exposure, which was not observed in UIMC1 modified cells ($p=<0.0001$ for e2-B1.15 and e13-KO1, $p<0.001$ for e2-B3.1). Conversely, a significant increase in γH2AX nuclear foci was observed in UIMC1 modified cells 24 hours post IR exposure ($p=0.009$ for e2-B1.15 and $p=<0.0001$ for e2-B3.1 and e13-KO1), suggesting cells lacking UIMC1 take longer to respond to DNA DSBs.

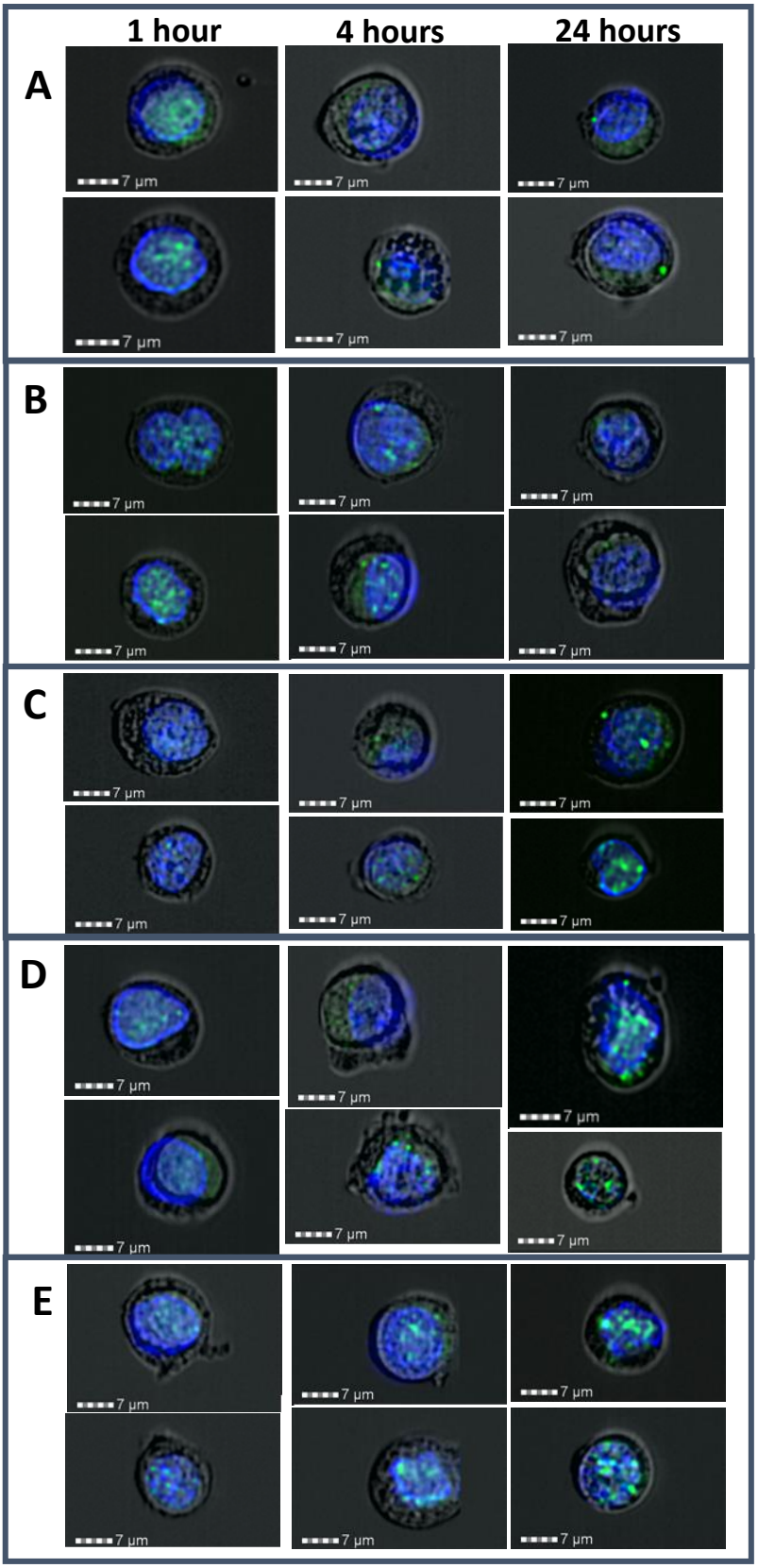


Figure 6.37: γH2A.X foci observed in cells exposed to 2 Gray ionising radiation at 1, 4 and 24 hours post irradiation. 2 cells for each cell line indicated each panel. **A.** WT HEK293 cells, with maximal foci observed 1-hour post irradiation **B**. PX330- (CRISPR sham, negative control) with maximal foci observed 1-hour post irradiation **C.** e2-B1.15 (UIMC1 homozygous KO) cells, maximal foci observed 24 hours post irradiation **D.** e2-B3.1 cells (UIMC1 heterozygous KO), maximal foci observed 24 hours post irradiation **E.** e13-KO1 (UIMC1 with 1AA deletion in ZFN) with maximal foci observed 24 hours post irradiation. Cells stained with FITC conjugated anti-H2AX antibody and DAPI. Cells visualised using the Image Stream X at 60X magnification using extended depth of field (EDF).

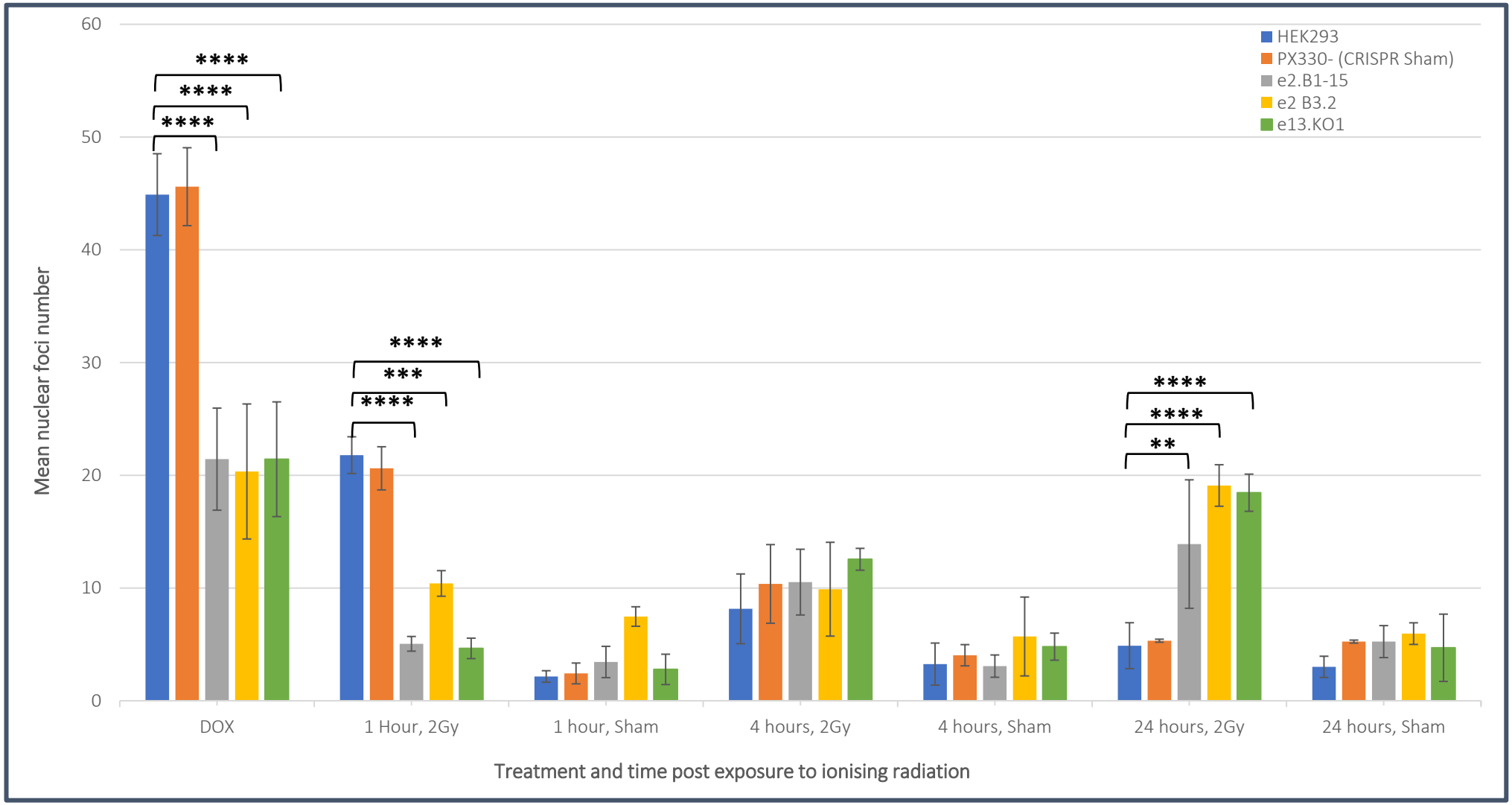


Figure 6.38: Mean number of nuclear foci following the induction of γH2AX of the nuclei of wildtype, UIMC1-modified and negative control HEK293 cells over a 24 hour time course following exposure to 2 Gy or Sham ionising radiation or 250 µM Doxorubicin hydrochloride (DOX) n=3. Data was derived from the capture of 500-1000 images of cells. Nuclear foci quantified using the IDEAS software. Mean +/- standard deviation. HEK293; WT, PX330-; CRISPR Sham, e2-B1.15; UIMC1 homozygous knockout, e2-B3.1; UIMC1 heterozygous knockout, e13-KO1; UIMC1 with 1 AA deletion in ZFN. Statistical analysis carried out using two-way ANOVA with Tukey's multiple comparisons, where **p<0.01, ***p<0.001 and **** p<0.0001.

6.4 Discussion

6.4.1 Linkage analysis of UIMC1 indicated individuals with the same mutation were most likely not related

As this study is a South Australian cohort study, it was necessary to determine if the UIMC1c.1690T>C variant segregated through a family or if it was detected within two unrelated individuals in this cohort. STS marker analysis was used on polymorphic microsatellites within the selected individuals for analysis. As highlighted in **Table 6.1**, it is unlikely that the individuals were related based on the STS markers selected. The differences in the repeat lengths between two markers (D5S2034 and D5S2006) in the *UIMC1*-mutated individuals provided further evidence that the two individuals were not related.

Unfortunately, the STS markers selected for analysis were not very informative as they were often of the same size for most individuals analysed. In order to overcome this limitation, it would be beneficial to analyse a greater number of markers, in addition to looking at microsatellites repeats surrounding *UIMC1*. A further limitation was that it was not possible to determine the phase of the marker alleles with the *UIMC1* variant. It is necessary to establish which chromosome carries the same size repeat fragments in order to determine relatedness between these individuals. In future, it may be beneficial to conduct a phasing analysis which identifies which chromosome is wildtype and which chromosome carries the variant. This would allow for the identification of haplotypes along the chromosome, providing additional information to determine relatedness in these individuals. This could be achieved through single molecule, long read sequencing approaches such as PacBio and Oxford Nanopore technologies.

6.4.2 UIMC1 functional analysis

6.4.2.1 Inability to generate single products of *UIMC1* open reading frame due to multiple splice variants.

The original experimental plan included functionally analysing the UIMC1:c.1690T>C variant by generation of a mammalian expression vector to produce a FLAG-tagged UIMC1 mutant protein. Due to the presence of internal splicing events it was not possible to design primers to amplify specific full length *UIMC1* ORFs. As a result, all variants were amplified and then excised from the gel. Countless attempts were made to optimise the second round PCR for the addition of restriction enzyme sites and a FLAG tag for cloning into pcDNA3.1 (data not shown). Despite these attempts, a

single product was unable to be generated. This generation of this construct would have been beneficial as it would have allowed for identification of interacting proteins through immunoprecipitation with the wildtype and mutated FLAG-tagged protein. Through the implementation of the tag, it would be possible to carry out analysis of the proteins interacting with UIMC1 using mass spectrometry, in order to determine if there were any changes observed with the mutated version.

Due to the aforementioned difficulties, the decision was made to proceed solely with CRISPR/Cas9 editing as a method of functional validation of the effect of the identified *UIMC1* mutation.

6.4.2.2 sgRNA design for CRISPR/Cas9 editing

At the commencement of this functional validation, CRISPR/Cas9 genome editing was well recognised, however the best methodology to utilise this powerful approach was not yet well established. There were many different approaches available for genome editing, and the literature advised that multiple sgRNAs were designed for each target region to maximise the chance of success (Ran *et al.*, 2013b). The CRISPR sgRNA design tool developed by Professor Feng Zhang's laboratory (which has since been deactivated) was utilised to design sgRNAs for the knockout of UIMC1 function within exon 2 (effectively rendering the protein non-functional) and also within exon 13 (the location of the identified variant). Whilst the identified mutation in UIMC1 did not result in premature protein truncation, this was carried out as it is significantly more difficult to insert the point mutation. The generation of a mutation within this zinc finger domain in UIMC1 is predicted to result in loss of function of this region. In addition, the generation of a truncation mutation within this final functional domain of UIMC1 may recapitulate the effect or be somewhat representative of the effect of the identified UIMC1:c.1690T>C mutation, allowing for some functional analysis of the effect of a mutation within this region of the gene.

The aforementioned sgRNA design program was used to generate a pair of sgRNAs around the specific UIMC1:c.1690T>C polymorphism for the targeted introduction of the sequence change. Through this methodology, a 'nick' is introduced onto each strand of the DNA, resulting in overhangs. This results in greater control over precise genome editing, as both Cas9 enzymes must nick the target DNA to create a double-stranded break, a method which has been shown to reduce off-target effects by 50-1500 fold in cell lines (Ran *et al.*, 2013a).

In addition to the design of paired sgRNAs, a homology directed repair template was designed and used to introduce the specific T>C variant of interest in UIMC1. Following the recommendations from the Zhang laboratory, the repair template was designed with the cut sites being as close to each other as possible, with 50-80bp of homology sequence (termed homology arms) flanking the desired base change, with an advised length of 100-150bp overall for the ssODN. Whilst multiple sgRNA pairs are recommended for the selection of the most efficient guides, the cost associated with the synthesis of long oligonucleotides made it too expensive to attempt to work with multiple sgRNA pairs for the incorporation of the UIMC1:c.1690T>C variant at any one time. As a result, the pair with the least predicted off-target effects was selected for the design of the long oligonucleotide. Since this work has been carried out, the advancements in sgRNA design has evolved rapidly. Now there are an array of computational tools that can be used to design the most optimal CRISPR/Cas9 experiments. These tools aim to maximise on-target activity and minimise off-target effects. Due to the rapid evolution of these tools, it is most likely that the pairs selected for editing in this instance may no longer be the most suitable pair for modification of this region.

6.4.2.3 Issues associated with the successful incorporation of sgRNAs into CRISPR plasmids

As mentioned in **Section 6.3.3**, the incorporation of the sgRNAs into the CRISPR plasmids required extensive optimisation. After multiple rounds of troubleshooting all aspects of the digestion and ligation, all oligonucleotides for sgRNA complexes were re-synthesised. sgRNA complexes were formed using Sigma annealing buffer, and a significant increase in the number of colonies was observed. PCR analysis and subsequent Sanger sequencing confirmed successful incorporation of the sgRNA complex into the CRISPR plasmid. This indicates that there may have been an issue associated with the original synthesised oligonucleotides. As all the oligonucleotides ordered were from the same batch and no sgRNA complexes were successfully incorporated into the plasmid, this is the most likely reason. The oligonucleotides ordered initially did not undergo any form of purification, and although the literature does state that purification should not affect sgRNA complex formation success, the newly synthesised oligos were subjected to purification. Whilst synthesised oligos will have a minimal level of contamination, there are various methods of purification which could be utilised in future. These methods of purification include polyacrylamide gel electrophoresis (PAGE) and high-performance liquid chromatography (HPLC). However, these purification methods can result in a substantial increase in the cost associated with multiple sgRNAs, in addition to a substantial decrease in oligo yield. After extensive attempts to optimise the sgRNA incorporation process, HPLC purified oligos were re-ordered and annealed. This result indicates that

it may be worth the increased cost of purified oligonucleotides, in comparison to the cost, time and stress associated with the lengthy optimisation process. Furthermore, as these issues associated with sgRNA complex formation are relatively common, various companies now offer a duplexed DNA service, where oligonucleotides are combined into the duplex formation and provided lyophilised for a nominal fee.

The optimisation of the sgRNA incorporation was a lengthy process which consumed a significant proportion of time within this research component. Whilst it is possible to order plasmids with the sgRNAs already incorporated, this was decided against. Though this option would significantly decrease the time associated with the CRISPR/Cas9 experimental procedures, it was expensive. Additionally, as a researcher, generation of the plasmids represented is a desired and technical skill that I wanted to learn and as such was determined to persevere and succeed. Through this process, the need to understand and troubleshoot all aspects of the experimental process became evident. This ordeal highlighted the need for controls in all stages of the sgRNA incorporation process, and the need to verify and even re-order oligos before proceeding redesigning new sgRNAs and starting over.

6.4.3 Cell lines selected for genome editing

The original research plan involved modifying *UIMC1* in two cell lines. MCF10A cells were selected for analysis as they are a non-malignant breast epithelium cell line, derived from benign proliferative breast tissue with fibrocystic disease which spontaneously immortalised (Soule *et al.*, 1990). MCF10As have a karyotype of 47 chromosomes, containing several chromosomal translocations and trisomy of chromosome 8 (Marella *et al.*, 2009). This cell line harbours a deletion of the locus containing *p16* and *p14ARF* genes, both of which play critical roles in regulating senescence (Soule *et al.*, 1990). This cell line is the most widely used *in vitro* model for studying normal breast function and understanding the genetic aberrations that play a role in the development of this disease phenotype. MCF10A was selected for functional validation as the wildtype model does not show any characteristics of tumour formation or invasiveness. Importantly, this cell line does not have any identified translocations or aneuploidies involving chromosome 5 (Marella *et al.*, 2009), on which *UIMC1* is located. However, these cells are known to have a low transfection efficiency (10-20%) and no published data could be found with respect to the success rate of CRISPR/Cas9 modification in this cell line. Therefore, an additional cell line was also utilised in order to maximise the possibility of successful modification of *UIMC1*.

The HEK293 cell line was selected as a positive control cell line, as these cells are often used in CRISPR experiments due to their high transfection efficiency. HEK293 cells are the second most commonly used cells in cell biology experiments and are known to proliferate well (Stepanenko and Dmitrenko, 2015). The HEK293 cell line utilised for this study was recently obtained from ATCC and had undergone minimal passages at the time of gene editing. Despite this, it is known that the HEK293 cell line has a hypotriploid karyotype, with a mean chromosome number between 54 – 63 chromosomes (Stepanenko and Dmitrenko, 2015). Initially, this cell line was only to be utilised as a positive control for transfection experiments, however due to the technical difficulties associated with working with the MCF10A cell lines, all functional analysis of *UIMC1* function was carried out in this cell line only. As these cells originate from immortalisation of embryonic kidney cells, the results must be analysed and interpreted within this context.

6.4.4 Lipofectamine Transfection Efficiencies

Prior to editing, the optimal conditions for transfection with Lipofectamine 2000 were determined for both HEK293 and MCF10A cells. Transfection efficiencies were initially determined for a control GFP-expressing plasmid (pmaxGFP) in order to optimise plasmid concentration, lipofectamine concentration and incubation times for each cell line. This was necessary for each cell line as the use of transfection reagents in low quantities results in low transfection efficiencies, however an increase in reagent concentration, particularly Lipofectamine, has been shown to induce cell toxicity (Avci-Adali *et al.*, 2014). As expected, transfections on HEK293 cells were fairly straightforward, with minimal cell death at recommended Lipofectamine 2000 and plasmid DNA concentrations (**Figure 6.14** and **Figure 6.15**). However, MCF10A cells were difficult to transfect. Manipulation of the Lipofectamine concentration and plasmid concentration did not result in a significant increase in GFP positive cells (**Figure 6.17**), with high levels of cell death due to lipofectamine-associated toxicity (**Figure 6.18**). Overall, only a maximum of approximately 20% GFP-expressing cells was observed for the MCF10A cell line, however this was observed in conjunction with a high level of cell death, with only 30% of cells remaining viable following transfection.

6.4.5 MCF10A cells demonstrated high levels of cell death and difficulty to transfect

To date, minimal studies have carried out CRISPR/Cas9 editing in MCF10As cells. This may be due to their low transfection efficiency, slow proliferation, and sensitivity to chemical and physical manipulation, as demonstrated by the work carried out within this thesis (**Figure 6.17**- **Figure 6.20**).

From the optimisation experiments carried out with this cell line, it was difficult to accurately analyse the transfection efficiencies of these cells, as the removal of cell culture media often resulted in cells losing their adherent quality. This issue was not associated with over-confluence, as cells were transfected at 85-90% confluence. Analysis of cell death via trypan blue indicated that even after 3 hours transfection, cell viability had decreased to 60-80% (**Figure 6.18**) with a maximum of 10% of cells expressing GFP (**Figure 6.17**). This indicates that these cells are particularly sensitive to chemical manipulation. Furthermore, the slow growing rates of the MCF10A cell line make them difficult to work with. This is a contrasting point within the literature, with some studies reporting that MFC10A cells grow quickly, having a doubling time of less than 24 hours (Thompson *et al.*, 2014, Bessette *et al.*, 2015). This is in contradiction with the cells that were obtained for this study, as these cells grew very slowly, even in the presence of cholera toxin and epidermal growth factor (EGF) which are used to stimulate the proliferation of epithelial cells. This may be due to batch-to-batch variation.

In addition, for a period of 9 months Lonza was unable to fulfil any orders for the MEGM bullet kit, which contains the media and growth components for MCF10A cell culture (as advised by ATCC). Alterative mammary epithelial cell culture medias were trialled during this period , including Cell Applications human mammary epithelial cell media and Gibco Media 171 with the addition of mammary epithelial growth supplements. These approaches maintained cell viability, however all cells failed to proliferate. Due to these issues and in the interest of time, it was decided to proceed solely with the HEK293 cells for CRISPR editing and UIMC1 functional analysis. Whilst this does impose a limitation on the inferences that can be drawn from this functional analysis, RNA seq analyses (as determined from the GTEx database) have shown that *UIMC1* is expressed in most major tissues within the body, with similar levels of expression observed in both kidney and breast tissues.

In future, it would be beneficial to attempt to use cell lines of the correct tissue of origin for further understanding of this UIMC1 mutation. Ideally, a different 'normal' breast cell line could be manipulated for functional analysis (i.e. MCF12A, 184A1, HBL-100 or even primary human mammary epithelial cells). This would allow the analysis of UIMC1 function to be carried out in a more physiologically relevant cell line. Failing this, a murine model for the analysis of UIMC1 function would also be beneficial.

6.4.6 Minimal success of *UIMC1* gene editing using Lipofectamine 2000

Despite published protocols advising the use of Lipofectamine for the transfection of CRISPR plasmids into cell lines, low transfection efficiency was achieved. In order to introduce the desired polymorphism, it was necessary to transfect all 3 required components (both plasmids containing guide A and guide B and the HDR template) into the nucleus of each cell. Furthermore, this method of genome modification relies on exploiting the less favoured homology directed repair (HDR) pathway for defined genome modification (Ran *et al.*, 2013a). It has previously been demonstrated that cells prefer to repair DNA breaks through the NHEJ pathway, rather than HDR (Kass and Jasin, 2010, Lieber, 2010). NHEJ is known to be significantly more efficient than HDR, which provides additional challenges in precise gene editing. Further to this, studies have shown a maximum of 10% editing efficiency when utilising the HDR pathway (Ran *et al.*, 2013a). Due to the limitations and results discussed above, alternative strategies to increase the likelihood of modifying UIMC1 using CRISPR/Cas9 in these cells were implemented.

Multiple mechanisms have been utilised to increase HDR efficiency. In order to increase the likelihood of getting all necessary components into the cell, it is possible to tether the HDR template to the Cas9 enzyme to ensure it is close by at the initiation of the DNA cuts (Aird *et al.*, 2018). Furthermore, it is possible to block the NHEJ repair pathway through blocking key NHEJ molecules and DNA ligase IV, ensuring DNA break repair is facilitated through the HDR pathway (Chu *et al.*, 2015). This has shown a 4 – 5-fold increase in HDR efficiency. Additionally, the use of compounds to synchronise cell cycling (arresting cells at G2/M phase and at the G1/S border) results in an increase from 10% HDR in unsynchronised cells to 35% in synchronised cells (Lin *et al.*, 2014b). Whilst one approach would have been to utilise these approaches in order to increase the success rate of HDR editing within the HEK293 cells in this study, these manipulations are often difficult to employ and can have detrimental effects to cellular function. Furthermore, it has been shown that there is a maximum capacity at which the HDR pathway can function within any one cell type, with HEK293 cells shown to have a maximum capacity of approximately 30% (Lin *et al.*, 2014b).

6.4.7 *UIMC1*-edited cell populations generated via transfection with Nucleofection

Review of recent literature illustrated that transfection using the Nucleofection method developed by Amaxa demonstrated greater transfection efficiencies (Liang *et al.*, 2015, Jacobi *et al.*, 2017), with a greater proportion of cell populations showing signs of modification. Protocols were optimised for HEK293 cells with both the control pmaxGFP and CRISPR PX461 GFP expressing plasmids for

accurate optimisation. Protocols documented a transfection efficiency of approximately 85% for HEK293 cells, however the observed transfection rate in this study was significantly lower. Furthermore, the pulse protocols with the highest transfection efficiency also resulted in the greatest cell death. This may have been attributed to a toxicity associated with the increase in plasmid concentrations required for this protocol.

When cells were transfected by nucleofection, a greater proportion of the cell population showed signs of modification when screened with the SURVEYOR assay. It would have been beneficial to determine the efficiency of transfection via nucleofection on the MCF10A cell line; however, each cell line requires a different Nucleofection kit, which is relatively costly compared to Lipofectamine ($1000 for 24 nucleofection reactions compared to $450 for 375 Lipofectamine reactions). Each cell line requires significant optimisation (8/24 reactions) and the MCF10A cell line had already shown to be difficult to work with.

6.4.8 Inability to incorporate the *UIMC1* point mutation into HEK293 cells

Despite increased success in transfection efficiency with nucleofection, the UIMC1:c.1690T<C polymorphism was unable to be introduced into the HEK293 cell line. Recently, studies have shown that there are several factors that can impact the efficiency of cooperative nicking (Bothmer *et al.*, 2017). Assessment of how sgRNA design affect HDR mediated editing has found that robust genome editing is only observed with the PAM sites faced outside the target region (termed a PAM-out orientation, **Figure 6.39**, panel **A**). The sgRNA pair used for this study had inward facing PAM sites (PAM-in orientation) which may be the reason why the point mutation was not able to be introduced with the designed guides. Furthermore, it has since been shown that Cas9 D10A nickase-mediated genome editing is more robust when the cleavage sites are 40-70bp apart, which is contrary to what was initially advised by Ran *et al.* (2013a). These features may explain why the point mutation was unable to be introduced with this particular pair of sgRNAs. Based on this new evidence, it would be beneficial to redesign the sgRNA pair with these additional features.

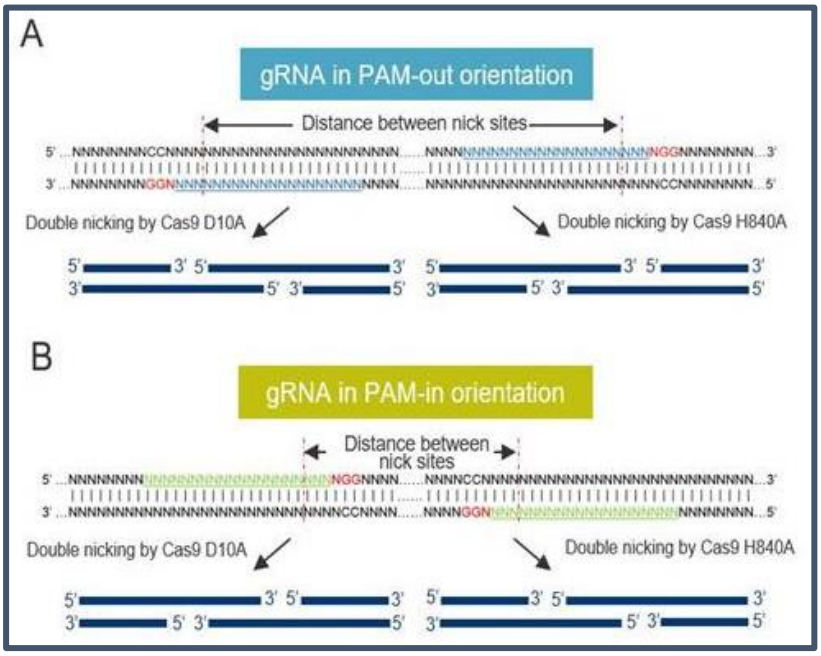


Figure 6.39: Paired sgRNAs with varied configurations of PAM sites. NGG PAM sites indicated in red, with corresponding cut sites are illustrated by red dashed lines. Modification with both Cas9 nickases D10A and H840A shown in panels A and B. Cooperative nicking with sgRNA pairs results in targeted nicking on opposite strands. **A.** PAM-out sgRNA orientation illustrates robust genome editing, with D10A nickases producing a 5' overhang, and H840A nickases generating a 3' overhang, with sgRNAs illustrated in blue. **B.** PAM-in sgRNA orientation has been shown to be inefficient for genome modification, resulting in 3' overhang with D10A modification and 5' overhang with H840A modification. sgRNA locations illustrated in green. Modified from Bothmer *et al.* (2017)

Furthermore, it has recently been found that cells can preferentially utilise the homologous chromosome rather than the provided HDR template for DNA break repair (Gorter de Vries *et al.*, 2018). The former repair mechanism has been shown to be significantly more efficient, and as a result, the use of the HDR template for targeted base pair changes is less likely to be successful. Therefore, this further complicates the application of CRISPR/Cas9 in introducing specific base changes.

Recent advances in CRISPR technology have greatly increased the ability and efficiency at which genomes can be functionally manipulated. Firstly, it is now possible to target the required modification to the RNA, through the utilisation of Cas13d nuclease. The advantage of this approach is the ability to modify the transcript without making permanent and potentially dangerous changes to the gene itself (Zhang *et al.*, 2018a). Additionally, it is now possible to utilise other mechanisms to deliver Cas9 and the sgRNA rather than through co-transformation of plasmids and HDR template. A study carried out by Gundry *et al.* (2016) utilised a plasmid- and virus-free system for genome modification through delivery of a Cas9/sgRNA ribonucleoprotein complex into wild-type

cell lines. This method is significantly faster, with the entire process being completed in 3 days, and an observed efficiency of 75% for knockout models and 20% for HDR based approaches. For any future validation work, it may be beneficial to utilise a more time efficient model for genome modification. However, this mechanism did not result in a significant increase in efficiency of genome modification via the HDR pathway. As it is this mechanism that is required for the introduction of point mutations, this method may not be beneficial for this type of genome editing and as such other mechanisms may need to be explored.

Interestingly, a recently developed technique named base editing is revolutionising the way point mutations can be introduced into cells for functional analysis. Base editing involves site-specific modification of DNA (Komor *et al.*, 2016). Base editors are chimeric proteins composed of a catalytically inactive Cas9 variant (dead Cas9; dCas9) which contains a catalytic domain capable of deaminating a cytidine or adenine base (Komor *et al.*, 2016, Gaudelli *et al.*, 2017). This mechanism does not require the generation of DSBs but instead utilises the dCas9 to target deaminase domains to specific loci for modification. Furthermore, this mechanism manipulates the DNA repair machinery to avoid unwanted repair of the modified base, hence limiting the generation of indels at target and off target sites (Hess *et al.*, 2017). More recently, it has been shown that Cas9 nickases can be used in place of dCas9, resulting in even higher frequencies of base editing (Eid *et al.*, 2018). Of particular interest for the UIMC1:c.1690T>C polymorphism is the adenine base editing (ABE) system. This mechanism results in the deamination of adenosine to inosine, which pairs with cytidine, and subsequently is corrected to guanine by polymerase enzymes. As a result, this mechanism is able to convert A/T→ G/C, with an activity window of 3-9bp from the protospacer (Gaudelli *et al.*, 2017).

Due to the rate at which the field of genome manipulation is expanding, it was not feasible to carry out the different methods that have been recently published in this area. However, it stands to reason that there are significantly more efficient methods of introducing the UIMC1:c.1690T>C polymorphism than those utilised within this chapter. For future work, it would be beneficial to use the ABE system as described above. However, as these base editing systems are still only relatively new, more work is required in both the delivery systems of these large molecules into cells and the analysis of any off-target effects.

6.4.9 Knockout of *UIMC1* in HEK293 cells

In addition to trying to incorporate the UIMC1:c.1690T>C variant into HEK293 cells, knockout cell lines were also generated. To date, there have been several studies that have used siRNA to knockout *UIMC1* function to assess its role in DNA damage repair (Jin et al., 2019, Wang et al., 2007, Sobhian et al., 2007, Yan et al., 2007a). These studies illustrated that UIMC1 is required for the localisation of BRCA1 to the sites of DNA damage and plays a vital role in the formation of the BRCA1-A complex in conjunction with BRCA1, CCDC98, BRCC36, BRCC45/BRE, and MERIT40/NBA1/HSPC142. The aim of this study was to add to the current knowledge of UIMC1 function through interrogation of the effect of the specific mutation identified in this breast cancer cohort.

Due to the nature of CRISPR/Cas9 editing, the sizes and locations of the introduced insertions or deletions in *UMIC1* varied for each monoclonal cell line produced. Comparison to the wildtype sequence was used to deconvolute the acquired mutation(s). This was relatively easy when only one allele was edited, however became more challenging once a second mutation was present. The increased complexity of the downstream sequencing proved difficult to identify the precise secondary mutation. Therefore, the functional effect of mutations in each cell line was determined by western blot analysis. It would have been beneficial to carry out whole genome sequencing of the *UIMC1* modified cell lines, as this approach would not only allow accurate characterisation of the exact mutations that had been introduced into each cell line, but also to determine any off-target effects of the CRISPR/Cas9-mediated editing.

Various studies have been published highlighting the off-target effects of CRISPR/Cas9 editing, with studies identifying high levels (>50%) of off-target RNA-guided endonuclease induced mutations (Fu *et al.*, 2013, Hsu *et al.*, 2013, Mali *et al.*, 2013a, Cho *et al.*, 2014, Zhang *et al.*, 2015b). As the protospacer is only 20bp long, it may be possible for the sgRNA and the PAM sequence to bind elsewhere in the genome (Fu *et al.*, 2013, Hsu *et al.*, 2013). Furthermore, it has been shown that the PAM and seed sequence (first 10-12bp) of the sgRNA determines Cas9 specificity, and as such, the lack of specificity between the distal region of the sgRNA sequence and the targeting strand does not prevent Cas9 cleavage, often resulting in off-target activity (Jinek *et al.*, 2012, Cong *et al.*, 2013, Zhang *et al.*, 2015b). Different strategies have been developed to improve Cas9 specificity derived from *S.pyogenes* (Sp), which include optimisation of sgRNA design (Hsu *et al.*, 2013, Wiles *et al.*, 2015), the use of paired nickases (Cong *et al.*, 2013, Mali *et al.*, 2013a, Ran *et al.*, 2013a),

modifications to Sp Cas9 for increased specificity (Slaymaker *et al.*, 2016), the use of shorter sgRNAs (Ran *et al.*, 2013a, Fu *et al.*, 2014), sgRNAs with two unpaired Gs on the 5' end that are more sensitive to mismatches (Kim *et al.*, 2015) and also through decreasing the Cas9/sgRNA complex concentration or the length of time active within the cell (Hsu *et al.*, 2013, Davis *et al.*, 2015). Although these approaches have shown greatly improved specificity of CRISPR mediated genome editing, they often result in a reduction in editing efficiency.

6.4.10 Confirmation of UIMC1 knockout via western blot analysis

In order to confirm the introduced mutations were indeed knockouts, western blot analysis was carried out for evaluation of protein expression. This analysis was non-quantitative and was carried out to determine presence or absence of UIMC1 within the generated cell lines and look for any size changes in protein transcript. From analysis of **Figure 6.33** it is evident that there are several generated monoclonal cell lines which showed a complete reduction in UIMC1 expression (e2-B1.14 and e2-B1.15) with a faint band observed at a lower molecular weight than expected. In the cell line e2-B3.1, a faint band could be observed at 80kDa, indicating a partial loss of expression. This could be attributed to the heterozygous nature of the introduced mutation within this cell line (**Figure 6.30**, Panel **E**), with one chromosome appearing to be unmutated. The e13-KO2 cell line also demonstrated a faint band at 80kDa. This was unexpected, as the introduced mutation was not expected to result in complete loss of UIMC1 expression. This could be attributed to off target effects of the CRISPR/Cas9 modification, however off-target effects within the same gene is unlikely unless it contains a repetitive sequence, which is avoided with sgRNA design. Furthermore, it may be possible that the introduced mutation resulted in the mRNA transcript undergoing nonsense mediated decay.

The additional two cell lines with exon 13 mutations (e13-KO1 and e13-KO10) displayed similar expression to the wildtype and negative control. This was expected due to the nature of the sequence changes introduced. With the exon 13 mutant cell lines, we were trying to detect a C-terminal premature truncation of approximately 100 amino acids. To analyse this, an antibody that binds to the N-terminus was utilised and a size shift would be expected with a premature truncation. In addition to this analysis, it could also be beneficial to utilise a secondary UIMC1 antibody for subsequent analysis of the UIMC1 exon 13 mutants, which binds to the end of the UIMC1 sequence (for example anti-UIMC1 antibody #ab70822 which binds to AA 600-650). By carrying out analysis with antibodies located near both termini of the protein, it may be possible to accurately determine

if the UIMC1 protein contains a premature truncation in the final exons of the transcript. Unfortunately, due to time restraints this was not carried out, and as a result all cell lines modified in exon 13 were utilised for functional analysis. Unfortunately, both e13-KO2 and e13-KO10 failed to proliferate and the number of cells required for functional analysis could not be obtained. This failure to proliferate could indicate that the introduced mutations were significantly affecting vital cellular processes, possibly resulting in stalled replication and ultimately resulting in cell death. In future, it would be beneficial to have multiple cell lines with mutations within both exon 2 and exon 13, with partial loss and complete loss of UIMC1 expression for functional validation. Through this, it would be possible to elucidate if haploinsufficiency of UIMC1 plays a role in normal cellular function and DNA damage repair capabilities.

From the western blot analysis, there was a cell line produced that did not show any reduction in UIMC1 expression (e2-B1.16). This is because this cell line carried a 6 bp in-frame deletion, resulting in the loss of 2 amino acids. This change was not expected to be observed on a western blot analysis. Additionally, some of the produced cell lines carried heterozygous edits of the DNA sequence. Within these cells, a 50% reduction in expression of UIMC1 would be expected, however it is not clear if this loss of expression would affect normal cellular function.

Multiple samples showed UIMC1 protein at higher and lower molecular weights than expected. The wildtype UIMC1 protein is approximately 80kDa in size. Larger molecular weight products can be attributed to post-translational modifications, as UIMC1 has been shown to undergo mono- or multi-sumoylation, which plays a role in regulation in UIMC1 function (Yan *et al.*, 2007b). These key post-translational modification events are carried out through the SUMO (Small Ubiquitin-like Modifier) proteins that serve to modulate cell function, such as transcriptional regulation, apoptosis, protein stability , stress response, nuclear-cytosolic transport and cell cycle progression (Hay, 2005, Yan *et al.*, 2007b). It is known that the ubiquinated form of UIMC1 migrates at 97kDa, and higher molecular weight products have been shown to correlate to UIMC1 phosphorylation products (Wang *et al.*, 2007). As previously discussed, there are multiple splice variants of UIMC1, and as such the smaller molecular weight products may be attributed to this.

6.4.11 Understanding the role of UIMC1 in cell function

Several studies have investigated the role of UIMC1 in DNA DSB repair (Kim *et al.*, 2007, Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). These studies have shown that cells lacking UIMC1

have defective G2/M checkpoint control and DNA damage repair capabilities. This suggests that cells with defective DNA damage repair capabilities would result in an accumulation of DSBs and mutations with the ability to drastically alter genome stability, one feature of which is an unregulated rate of cell proliferation. Furthermore, previous studies have previously shown that UIMC1 deficient cells show a reduced capacity to repair DNA DSBs as measured by γH2AX (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). Additionally, it has been demonstrated that a reduction in expression of vital DNA damage response proteins, such as BRCA1 (Zhang *et al.*, 2004), MDC1 (Stewart *et al.*, 2003), ATF2 (Bhoumik *et al.*, 2005) and BRIT1 (Lin *et al.*, 2005) can affect the sensitivity cells to ionising radiation (IR). Cells lacking UIMC1 also display hypersensitivity to IR, in both cells with a complete or partial loss (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007b). As a result, cell proliferation, sensitivity to IR and DNA DSB repair was assessed in cells lacking or containing a partial loss of UIMC1 to determine its potential aetiology in cancer.

Previous studies have shown several limitations which our study has attempted to address. Firstly, previous studies utilized siRNAs for inducible knockout of UIMC1. As these were inducible, the long-term effects were not assessed, whereas in our study complete knockout was used. These previous studies have also assessed only the effect of complete knockout of UIMC1 on DNA repair, while this study has also included analysis of partial loss. Additionally, previous studies also analysed DNA DSB repair in cells lacking UIMC1 at 1-2 hours post exposure to IR. While this timeframe has been established as when DBA DSB repair and γH2AX foci formation is greatest, this approach assumes normal cell function. This study has therefore included a time-course analysis of γH2AX analysis in order to assess whether mutant cells, which may rely on alternative pathways for DNA repair and foci formation, differ from normal cells. Additionally, two mechanisms of inducing DNA DSBs have been utilised for analysis.

6.4.11.1 Analysis of cell proliferation in UIMC1-deficient cell lines

It is known that cells lacking BRCA1 or BRCA2 show dysregulated cell growth due to lack of G2/M checkpoint control (Kim *et al.*, 2007, Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). As UIMC1 is known to function in the same pathway as BRCA1, it was hypothesised that cells lacking UIMC1 may also demonstrate perturbed cell growth. As UIMC1 is one of the key components of the BRCA1-A complex required for the repair of DNA double stranded breaks, it is feasible that a lack of UIMC1 will result in a decreased ability for DNA damage repair and accumulated mutagenesis within cells lacking UIMC1.

To assess whether cell proliferation capabilities were affected in UIMC1-mutant cell lines, wildtype and modified cells were plated (n=4), and cell growth was assessed over a 7-day period. The results were varied, with the UIMC1-mutant cell lines exhibiting both an increase and decrease in rates of cell proliferation dependant on mutation type (**Figure 6.34**). Cell lines e2-B1.15 (homozygous knockout) and e13-KO1 (UIMC1 with 1 AA deletion in ZFD) both demonstrated increased in cell proliferation in comparison to the wildtype and negative control, however only one time point (168 hours) was statistically significant due to the large variation between replicates. The e2-B3.1 (heterozygous knockout) cell line demonstrated a statistically significantly decreased rate of proliferation in comparison with controls, with minimal variation between replicates. This has been a contradicting point within the literature, with Jin *et al.* (2019) identifying that a reduction in UIMC1 can significantly inhibit cell growth in MCF7 cells and promote cell apoptosis, however Hu *et al.* (2011) identified that UIMC1 depleted cells did not illustrate a change in cell cycle distribution, proliferation or apoptosis.

Additional cell lines were selected for functional analysis of introduced variants within exon 13 of UIMC1, however these cells failed to proliferate. This could be attributed to the dysregulation of cell cycling and lack of genome stability within these cells, which may have accumulated pathogenic mutations within vital regions. However, as these cells failed to thrive, it was not possible to further elucidate the effects of UIMC1 loss within these cells. Due to this, only one cell line with a UIMC1 mutation within exon 13 was analysed further. Unfortunately, this limits the conclusions that can be drawn from this data, as the purpose of this study was to understand the role of the UIMC1:c.1690T>C variant in breast cancer. Moreover, the two exon 2 knockout cell lines displayed very different rates of cell proliferation, with e2-B1.15 (homozygous knockout) indicating large variation between replicates. This could have been attributed to the method used to determine cell proliferation rates, which has a lower precision than other approaches. In future, it may be beneficial to use methods with improved accuracy, such as the xCELLigence real time cell analysis assay, MTT assay, or even the IncuCyte to monitor cell proliferation rates. Furthermore, it would have been advantageous to select 3 cell lines with each type of modification (homozygous knockout, heterozygous knockout, exon 13 mutation) in order to have a more robust analysis of the effect of complete loss of UIMC1 expression in comparison to a premature truncation.

6.4.11.2 Effect of UIIMC1 mutations on cell viability and sensitivity to ionising radiation

Cells lacking UIMC1 have been shown to display hypersensitivity to IR in both cells with a complete or partial loss (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007b). Therefore, the viability of *UIMC1*-mutant cells exposed to therapeutic doses of IR (2 Gy) was examined over a 24-hour time course (**Figure 6.36**). Specifically introduced mutations within *UIMC1* can result in aberrant control of cell cycle checkpoints and DNA damage repair, subsequently reducing cell survival post-irradiation exposure. The data obtained within these cell lines indicates a slight decrease in cell viability (approximately 15%) following irradiation, however this was not statistically significant. It is worth noting that there was one mutant cell line, e2-B3.1 (carrying a heterozygous knockout mutation) which demonstrated a significantly lower baseline viability than all other cell lines analysed. This could be attributed to cumulative mutations within this particular cell line, or general genome instability.

Further analysis of previous studies identified that hypersensitivity to IR was observed in UIMC1-deficient cells at doses greater than 4 Gy, with all studies demonstrating a minor decrease in cell viability at doses lower than this (<10% decrease) (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007b). In order to determine if a similar hypersensitivity is observed within these knockout and partially mutated cell lines, it would have been beneficial to carry out a dose response analysis on the *UIMC1*-mutated cell lines, using 4 Gy, 6 Gy and 8 Gy IR. Through this, it would be possible to determine if the cells with truncated UIMC1 display the same hypersensitivity as knockout cells. Additionally, it has been demonstrated that cells deficient in UIMC1 exhibit defects in G2/M cell cycle control and homologous recombination capabilities in response to DNA DSBs (Wang *et al.*, 2007). Therefore, in order to ascertain the effect of a mutation within exon 13 it would have been beneficial to analyse these mechanisms within the developed cell lines. However, due to time constraints and issues with cell proliferation, this was not feasible within this study.

Inhibition of UIMC1 expression within breast cancer cells has been shown to induce apoptosis (Jin *et al.*, 2019). Initially, apoptosis analysis was to be carried out on the mutant cell lines post IR exposure to assess both DNA double-stranded break repair and cell death and apoptosis in UIMC1 modified cell lines. Unfortunately, analysis of phosphatidylserine was only available through FITC conjugated-Annexin V, which could not be carried out in parallel with the γH2AX analysis as the fluorophores overlap. Furthermore, due to the need to permeabilise the cells, it was not possible to use the Annexin V or PI stain on the irradiated cells for a multi-channel analysis, with cells needing

to be removed from one experiment in order to conduct the other. Whilst it would have been beneficial to analyse the additional mechanisms of DNA damage repair and cell cycling control within the UIMC1, there was a slight decrease in viability observed in UIMC1 deficient cells. This further supports the idea that UIMC1 functions as a facilitator of efficient DSB repair.

6.4.11.3 UIMC1-deficient cells showed a delayed DNA double stranded break repair capacity

Nuclear localisation of H2AX foci is a useful method to indirectly quantify DNA DSB repair capabilities. The induction of DSBs leads to phosphorylation of the minor histone protein H2AX at Serine^{139} within the protein to form γH2AX (Ivashkevich *et al.*, 2012). Thousands of phosphorylation events at the site of DNA DSBs results in the formation of a focus, with the number of foci observed correlating with the number of DNA DSBs and the cells ability to repair them. From the analysis carried out, it was evident that there was a trend towards UIMC1-deficient cells displaying a reduced capability to repair DNA DSBs (**Figure 6.38**). Although a trend was observed, a large variation in nuclear foci numbers was observed between technical replicates. This variation is attributed to technical replicate 3, where a significant decrease in H2AX foci was observed in all cell lines in comparison to the previous two replicates (Refer to **Appendix J** for raw data). These experiments were conducted on cell passages 4, 6 and 10, with a vast reduction in the number of nuclear foci observed in passage 10. This increase in cell passage coupled with knockout of important DNA repair proteins could result in a dysregulation of key components of cellular mechanisms. Ideally, it would have been beneficial to complete additional biological replicates for increased power, in addition to conducting biological replicates on sequential cell passages.

Studies have previously shown that UIMC1 deficient cells show a reduced capacity to repair DNA DSBs as measured by γH2AX (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). However, these studies have only analysed repair capabilities 1 to 2 hours post irradiation. From the time course analysis carried out within this study, cells expressing WT UIMC1 maximal H2AX foci are observed at 1-hour post irradiation. However, analysis in UIMC1-deficient or UIMC1-truncated cells illustrates peak H2AX at 24 hours post irradiation. There is a significant reduction in the overall number of H2AX foci observed (when compared to WT) in all UIMC1 mutant cells. This further supports that cells which are UIMC1-deficient or contain a truncated version of UIMC1 have impaired DNA DSB repair capabilities. Unsurprisingly, DNA damage repair was poorest in cells with a complete knockout of UIMC1 (e2-B1.15) as determined by the lowest number of nuclear foci,

whilst cells with a truncated UIMC1 or reduced expression were slightly better, but still markedly less than WT cells.

Analysis of nuclear H2AX foci observed in the UIMC1-mutated cell lines were larger in size than those observed in WT cells, and localised to several distinct regions within the nucleus as compared to a wider spread in WT cells (**Figure 6.37**). Maximal H2AX foci were observed at different time points and as such, could suggest that the UIMC1-deficient cells are using alternative means to facilitate DNA DSB repair. This slow increase in the appearance of maximal H2AX foci within UIMC1-deficient cells indicates that they may be using an alternative pathway or recruiting different enzymes for the repair of DNA DSBs which take a longer time to respond to the induction of DSBs by IR. There are several DNA repair complexes formed in association with BRCA1 for the facilitation of DNA repair, those being BRCA1-B and BRCA1-C, which are all known to function within different aspects of DNA DBS repair (Savage and Harkin, 2015). Further work for the investigation of this pathway could include some of the other key components of not only the BRCA1-A pathway, but also these other BRCA1-B and BRCA1-C pathways to determine if DNA damage repair is being carried out through alternative means.

A further complication to this analysis was the use of the ISX for imaging flow cytometry analysis. This machine required a minimum of 10^7 cells per treatment, with a high number of cells routinely lost in the various washes and stains of the experimental process. Whilst this machine simplified the assessment of γH2AX foci, the high numbers of cells required for all samples and issues with cell clumping were problematic. In order to overcome this, it may be beneficial to assess nuclear foci through conventional inversion microscope and quantify the γH2AX intensity through flow cytometry. Furthermore, UIMC1 has been shown to form discrete nuclear foci, which is often seen to overlap with the localisation of γH2AX foci (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). For further assessment of DNA DSB repair, it would be interesting to use an additional antibody with an alternative fluorophore to assess UIMC1 foci within the nuclear compartment. This would allow an understanding of global UIMC1 expression within WT SHAM-irradiated cells, but also further analysis of nuclear foci formation at the sites of DNA DSBS, in both UIMC1- mutant and WT cell lines over time.

In order to further understand the role of UIMC1 in DNA damage repair and cell proliferation, it may have been possible to attempt to rescue the cells through a complementation assay. Through the

use of a plasmid expression system, it could have been possible to re-express UIMC1 within the knockout cell lines (e2-B.15), and even introduce the specific UIMC1:c.1690T>C variant in the protein through site directed mutagenesis. Here, the same assays could be carried out and compared to assess their ability to repair DNA damage, proliferate and their sensitivity to ionising radiation in the presence of UIMC1. However, this assay relies on the overexpression of UIMC1 through the expression system and is not representative of the protein levels that would be expressed within the cell.

In addition to the work carried out within this thesis, a recently published study by Jin *et al.* (2019) analysed *UIMC1* expression in invasive breast cancer in comparison to paired normal tissue and also analysed the role of UIMC1 in cellular function. They identified that cells lacking UIMC1 demonstrate a higher level of apoptosis than wildtype MCF7 cells. They also showed that *UIMC1* expression was significantly lower in breast cancer cells, and that expression was related to tumour size and lymph node metastasis. Additionally, they demonstrated that UIMC1 is involved in the carcinogenesis of ER-negative breast cancer and that *UIMC1* mRNA expression was lowest in triple negative breast cancers. The work carried out by Jin *et al.* (2019), in addition to the work that has been carried out in this thesis, provides further evidence that UIMC1 may play a role in the predisposition to inherited breast cancers. From this, it is feasible that further work should be carried out to elucidate the role of this gene in cancer predisposition, in addition to carrying out this analysis on a larger cohort of individuals and controls.

Overall, the results from this study importantly support the previous literature in demonstrating that UIMC1 is a crucial protein in the DNA damage repair pathway and is required for the initiation of key processes in the recruitment of BRCA1 to the sites of DNA damage for repair (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). Significantly, this work demonstrates that cells lacking UIMC1 still demonstrate an ability to repair DNA DSB. However, it is thought that they may rely on an alternative pathway which takes longer to activate.

Chapter 7:
Final discussion and summary

This project aimed to identify the cause of inherited breast cancer in individuals that were found to be *BRCA1/2* mutation-negative. By utilisation of a custom gene panel and Ion Torrent MPS, 51 genes were sequenced and analysed within the selected cohort. Using the previous Sanger sequencing data, a bespoke bioinformatics pipeline was generated based on the previously sequenced *BRCA1* and *BRCA2* data. From the analysis of 131 individuals, 166 potentially pathogenic variants of interest were identified. These variants were subjected to detailed *in silico* analyses, including frequency and clinical significance databases, in addition to prediction of pathogenicity of amino acid changes and sequence conservation. Through this thorough analysis, 81 potentially pathogenic variants were identified in 82 individuals. A pathogenic nonsense PALB2 mutation was also identified in two individuals. From these identified variants, CRISPR/Cas9 was used to functionally validate the role of *UIMC1* in DNA DSB repair and cell proliferation. This analysis demonstrated that cells lacking UIMC1 showed delayed DNA damage repair capacity when exposed to ionising radiation and perturbed cell proliferation rates. Whilst further analysis is required, this preliminary data, in conjunction with recent publications illustrates that UIMC1 may play a role in breast cancer predisposition in *BRCA1/2* mutation-negative individuals.

7.1 Genetic risk prediction

The literature recommends *BRCA1/2* genetic screening be carried out only in individuals with Manchester scores >20 (Evans *et al.*, 2004). In this study over half of the included individuals were selected based on a score of <20, as this research included screening both known moderate- to low-susceptibility genes and potentially novel susceptibility genes with an unknown level of effect. Interestingly, 7/11 *BRCA1/2* mutation-positive individuals included in this study as controls had Manchester scores <20, suggesting that the cut-off of scores >20 are in fact too stringent to identify all individuals carrying a *BRCA1/2* mutation. In recent years there has been multiple updates to the MSS to improve accuracy. These updates now include pathological assessments of tumours, including taking into account the increased prevalence of triple negative tumours in hereditary breast cancer (Evans *et al.*, 2017). Although these recent updates have not been utilised on this patient cohort, it is likely that these changes would improve the predictive power of the MSS within this cohort. Given that of the large number of individuals being offerred genetic screening, with no clearly pathogenic mutations being identified within *BRCA1/2*, it may also be beneficial to utilise one of the more complex and detailed risk prediction models (ie. BOADICEA) in the future.

The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) is a risk prediction model that is used to determine the probability of identifying loss of function mutations within a multitude of breast and ovarian cancer genes. This model determines susceptibility not only based on mutations within *BRCA1/2* but also employs a polygenic risk model, illustrating the joint multiplicative effect of low penetrance susceptibility genes (Antoniou *et al.*, 2004). This model is uniquely flexible, allowing for the analysis of families of all disease inheritance patterns, taking into account effects of breast (male and female), ovarian, pancreatic and prostate cancer (Antoniou *et al.*, 2004). Recently, BOADICEA has been updated to include analysis of specific mutations in *PALB2*, *CHEK2* and *ATM,* in addition to polygenic risk factors associated with SNPs known to account for approximately 20% of polygenic variance in inherited breast cancer cases, and as such would have been beneficial in this study. This model considers mutation-testing sensitivity in addition to lifestyle choices, hormonal status and reproductive risk factors (Lee *et al.*, 2019). Analyses of this model have indicated it out-performs the BRCAPRO, Manchester scoring system and the Myriad tables in predicting the number of mutation carriers and the number of mutations within families (Antoniou *et al.*, 2006, Antoniou *et al.*, 2008). Despite this, the MSS is still utilised for the prediction of mutation carriers within families. The reason for this is unclear, however alternatives should be considered in the future.

7.2 Gene panel screening methodology

Despite the identification of *BRCA1* and *BRCA2* as cancer susceptibility genes over 20 years ago, the explanation for a majority of inherited breast cancers still remains unknown. A vast number of studies have been carried out in order to identify other highly penetrant genes involved in breast cancer susceptibility, with limited success. Whilst a range of mid to low penetrance genes have been identified, the cause of 70% of inherited breast cancers remains unknown. This is attributed to several factors, including the highly heterogenous nature of some of these families, in addition to the polygenic model of susceptibility whereby multiple mutations in genes play a role in cancer predisposition.

Overall, the results of this study have demonstrated that there is a clinical utility in expanding sequencing studies beyond *BRCA1* and *BRCA2* to include previously identified moderate- and low-penetrance breast cancer susceptibility genes. However, due to the number of variants of uncertain significance identified in the diagnostic genes, the clinical utility in screening beyond the known breast cancer susceptibility genes has not been established. The results obtained from this thesis

have demonstrated that the use of multi-gene panels as a screening method on *BRCA1/2* mutation-negative individuals has a benefit. As previously discussed, a pathogenic *PALB2* variant was identified in 2 individuals (1.67%) within the patient cohort. In addition, this panel also has the potential to identify novel genes which may be involved in the development of breast cancer. This has been demonstrated through the genetic screening, detailed bioinformatics analysis and functional analysis that has been conducted through this research project. Through this analysis, we have identified *UIMC1* as a potential candidate for breast cancer susceptibility, however further analysis is required to more comprehensively understand its role in carcinogenesis. This emphasises that there is no clinical utility in screening the discovery genes included on the gene panel, at least for the time being.

Despite these successful results, there were several issues associated with sequencing a large number of genes and the use of the custom gene panel which became apparent throughout this project.

7.2.1 Limitations of AmpliSeq gene panel screening

This study utilised a custom gene panel which consisted of 51 genes that were either already implicated in the development of hereditary breast cancer or were hypothesised to play a role in the development of breast cancer. This panel was designed in 2013 using the Ion AmpliSeq Designer, which resulted in the generation of a highly multiplexed PCR across 2 pools for the sequencing of the exonic and flanking intronic regions of the selected genes. As outlined in **Appendix B,** 100% coverage was not achieved for the majority of the selected genes, but most importantly 100% coverage of *BRCA1* or *BRCA2* was not achieved. This would need to be rectified for this panel to be used in a diagnostic setting. With MPS technologies becoming standard practice in the majority of sequencing labs, the algorithms for multiplexed PCR design are significantly better, with 100% coverage of *BRCA1* and *BRCA2* achieved through the commercially available gene panels. Due to the significant roles *BRCA1/2* play in cancer predisposition, these genes are routinely screened on a wide range of hereditary breast cancer panels and multi-cancer panels.

Due to the highly multiplexed nature of the custom PCR, it is not possible to modify the designed gene panel for the addition or removal of genes as new susceptibility loci are identified. Furthermore, the cost of sequencing the whole exome and even the whole genome has decreased significantly since the conception of this study, and as such may be a more viable option for

identifying the genetic cause of inherited conditions such as breast cancer in the future. It is likely that WES and WGS will supersede the use of gene panels, not only in research, but in the diagnostic setting in the future. Currently, exome sequencing is carried out for the diagnosis of paediatric onset disorders, where the cause of disease is unknown in a variety of hospitals (Need *et al.*, 2012, Sawyer *et al.*, 2016, Clark *et al.*, 2018, Mak *et al.*, 2018). As the cost falls, the main limitation of both WES and WGS remains the difficulties in interpreting the overwhelming amounts of data generated through these sequencing approaches, in addition to determining any clinically reportable findings (Dewey *et al.*, 2014). Whilst the usefulness of WES and WGS in the identification of novel disease genes has been demonstrated (Saitsu *et al.*, 2013, Reid *et al.*, 2016), findings are not clinically actionable without in-depth functional characterisation of the identified genes and specific mutations.

Recently, multiple studies have been carried out to compare the use of gene panels, WES and WGS for diagnosis of various Mendelian disorders (Sun *et al.*, 2015b, Cirino *et al.*, 2017, Hamblin *et al.*, 2017). Comparisons carried out in these studies have shown that WGS is able to detect nearly all variants identified through targeted gene panel approaches. However, this increase in sequencing data comes at the cost of read depth and with a significant increase in the number of VUS identified. Furthermore, this increase often results in an increase in incidental genetic findings. WES also demonstrated similar results, with all informative variants being identified. These studies demonstrated that whilst panel testing, WES and WGS provided a similar diagnostic yield, the main advantage associated with WES and WGS approaches are the ability for reanalysis of all coding or all genomic regions over time, allowing for incorporation of new knowledge.

In comparison to WES and WGS, gene panels are targeted to the regions of interest and result in a significantly smaller number of VUS within each patient. The drawback to this is that they are often unable to be modified and may be missing pathogenic, disease causing mutations within genes that have been overlooked. Despite the increase in data associated with both WES and WGS, there is an ever-expanding list of susceptibility loci, which are covered by these approaches. This becomes difficult territory to navigate; which is the best approach for sequencing analysis? Either the entire exome can be sequenced in these individuals, with the benefit of potential mutation identification outweighed by the myriad of VUS that will be present within each sample; or analysis of a targeted panel of functionally significant genes with a known role in cancer predisposition. Studies have previously identified that increasing the number of genes assayed has minimal clinical impact, which

was supported by the results of this study (Lincoln *et al.*, 2015, Tung *et al.*, 2015, Maxwell *et al.*, 2016, Prapa *et al.*, 2017). Therefore, whilst it may be beneficial to utilise WES for the analysis of these individuals, the prospect of the significant increase in VUS is detrimental. To overcome this, it may be possible to determine a list of genes for analysis post-exome sequencing, focussing solely on these regions. This would allow for a reduction in VUS and incidental findings within regions lacking relevance, but also provides the potential to expand this analysis to a wider range of genes at a later time if desired.

Another limitation associated with the custom AmpliSeq panel is the inability to detect copy number variations. Germline copy number variations (CNVs) are structural variations, resulting in the loss or gain of regions of genomic material. CNVs have been shown to play a role in breast cancer susceptibility and development (Krepischi *et al.*, 2012, Kuusisto *et al.*, 2013, Masson *et al.*, 2014, Kumaran *et al.*, 2017, Walker *et al.*, 2017) and have a range of effects based on their location, including gene dosage effects and *cis*-regulatory functions (Kumaran *et al.*, 2017). However, the distribution of CNVs throughout the genome is disproportionate, with the vast majority of CNV break points occurring within intergenic or deeply intronic regions (Ellingford *et al.*, 2018). This often results in ambiguity in the functional impact of these CNVs and determining their role in the observed phenotype is difficult. CNVs that overlap with protein coding genes often offer insights into disease biology and the observed phenotype, with nearly 80% of cancer-causing genes harbouring CNVs (Pang *et al.*, 2010). Therefore, it is beneficial to not only analyse sequence changes, but also structural changes when looking for pathogenic mechanisms of cancer within these familial cases. However, the location of CNVs makes it difficult to capture the vast majority of CNVs by gene panel MPS approaches, which primarily focus on exonic regions and verified pathogenic intronic regions. This creates a limitation in the types of variant detection that can be carried out on data generated through the use of gene panels (Ellingford *et al.*, 2018). Analysis of read depth approaches have been utilised recently in conjunction with gene panel generated data (Schmidt *et al.*, 2017, Germani *et al.*, 2018), however this requires a much greater read depth (minimum recommended 100 X coverage) than what has been obtained for the individuals included in this study (Yao *et al.*, 2017). Other mechanisms such as MLPA and microarrays have been utilised in conjunction with MPS panels for CNV detection within screened individuals (Alkan *et al.*, 2011, Schmidt *et al.*, 2017, Germani *et al.*, 2018). In future, analysis of CNVs within the selected genes and surrounding intergenic regions would be beneficial for these mutation-negative individuals included in this study.

7.2.2 Variants of uncertain significance

With the cost of genetic screening falling, in addition to healthcare providers increasingly integrating genetic testing into their medical management, there is a significant increase in the number of individuals undergoing genetic analysis. A major bottleneck not only associated with this study but most sequencing studies, in particular WES and WGS, is the challenge of identifying causative variants amongst the large number of sequence variants identified within each patient sample.

Interpretation guidelines have been developed for the analysis and classification of variants into one of five categories based on relevant databases and computational analysis programs. These are pathogenic, likely pathogenic, variant of uncertain significance (VUS), likely benign, and benign (Plon *et al.*, 2008, Richards *et al.*, 2015). Whilst those that are pathogenic variants are often clinically actionable and can be reported back to healthcare providers, the VUS, which have an unknown impact on cellular function and health are of interest within the patient cohort contained in this thesis. The majority of commonly identified VUS are missense and splice site mutations, but may also be synonymous mutations and in-frame insertions and deletions (Apostolou and Fostira, 2013). These variants usually have conflicting interpretations of pathogenicity within the literature or lack enough evidence to determine if the identified variant is disease causative. The identified VUS within the patient cohort pose significant problems, as it is often unclear if they predispose these individuals to cancer, or do not play a role in cancer development. Additionally, this lack of clarity within the public repositories is often problematic, with around 85% of sequence variants reported in ClinVar only being reported by single submitters, with minimal functional validation being carried out (Kobayashi *et al.*, 2017). In an era where the detection of sequence variants outpaces the ability to generate functional analyses of the phenotypic consequence of identified variants, how best to proceed and accurately analyse the significance of the myriad of VUS identified in each individual screened is an issue.

From the analysis carried out in this study, on average 125 sequence variants were identified within each patient sample, including a variety of common polymorphisms, silent and non-synonymous sequence variants, and a range of putatively functional and potentially pathogenic variants. The array of variants identified within each sample required detailed analysis for the elimination of benign sequence variants, and the identification of any clearly pathogenic mutations. However, for most individuals in this patient cohort, no clearly pathogenic mutations were identified, but instead a significant number of VUS were found, with 81 VUS identified in 82 mutation-negative individuals.

However, VUS are unable to be reported back to these individuals and their families unless they are clinically actionable, rendering these families unable to appropriately manage their risk and undergo any appropriate therapeutic measures, often resulting in additional stress for patients and families.

Variant interpretation guidelines have resulted in a more stringent analysis for the determination of causative, clinically actionable variants, however they have also resulted in a larger proportion of variants being deemed as VUS (Richards *et al.*, 2015). As recommended by the American College of Medical Genetics (ACMG) guidelines, VUS should not be used in any clinical decision making until more detailed functional analysis has been carried out (Richards *et al.*, 2015). However, unlike other ambiguous medical test results, the understanding and classifications of VUS often change over time and are regularly resolved as more data is gathered. Laboratories routinely reclassify VUS, with their classification either being upgraded to pathogenic, or more commonly downgraded to benign (Macklin *et al.*, 2018). These results are routinely relayed to healthcare professionals, which can in turn be used to alter treatment or surveillance in affected individuals. According to the ACMG guidelines, the onus falls onto the identifying laboratories to carry out further functional analysis of the identified VUS to gather enough supporting evidence to determine the pathogenicity of the identified sequence variants (Richards *et al.*, 2015). However, due to the large number of identified VUS within sequencing studies, this is impossible, and can end up being quite costly and time demanding. Once the role of pathogenesis has been established by the laboratories, it is recommended that this information is routinely disseminated to health care professionals and public repositories to assist in future diagnoses. However, determining the pathogenesis of variants is quite subjective, which is illustrated by the number of variants with conflicting interpretations of pathogenicity listed within public repositories. Of the 81 potentially pathogenic variants identified within this study, 10 (13%) were listed in ClinVar with conflicting interpretations of pathogenicity. This emphasises the need for the development of more systematic strategies for the determination of variant pathogenicity, however these are still far from being established.

Regarding the VUS identified within this study, the use of various computational analysis programs and databases only allowed for the elimination of so many variants. From the 131 individuals screened within this study, there were 81 VUS identified which were predicted to be pathogenic but had not been functionally validated. Whilst some sequence variants had stronger evidence supporting their role in cancer development than others, all were predicted to be pathogenic which made them potential candidates for cancer development within the screened individuals. However,

as previously emphasised, the lack of functional evidence demonstrating their role in cancer development meant that the identified variants were not clinically actionable until additional work had been carried out to fully elucidate the effect of the sequence variants. With the large number of variants identified from the screening of only 51 genes within a relatively small cohort of 131 individuals, even this number of variants is prohibitively time consuming and costly to attempt functional validation on. This demonstrates the need for a more robust, fool-proof method for the analysis of the functional effect of identified variants and their effect on normal cellular function and in disease pathogenesis. This could be carried out through the use of high throughput screening methods such as the CRISPR screens or through a cell screening facility.

7.2.3 Retrospective analysis of genes included on the custom AmpliSeq Panel

From the interrogation of sequence variants identified in all patient samples (as included in **Appendix H**), it was possible to visually determine genes in which either zero or very few sequence variants were identified. Furthermore, genes with a high level of sequence variation were also observed.

7.2.3.1 Diagnostic genes

As the only robustly identified and clinically actionable additional breast cancer predisposition genes are *ATM, BRIP1, BARD1*, *CHEK2* and *PALB2* the inclusion of all of these genes on this panel is justified (O'Leary *et al.*, 2017). From the analysis of the diagnostic component of the panel, there were several genes in which sequence variants were not routinely identified. These were *CHEK2* (15 rare variants in 22 individuals), *MRE11A* (10 rare variants in 11 individuals) and *XRCC2* (3 rare variants in 17 individuals).

It was surprising that no pathogenic mutations were identified within *CHEK2* within this cohort, as this is the most commonly reported gene with susceptibility mutations after *BRCA1* and *BRCA2* (Southey *et al.*, 2016, O'Leary *et al.*, 2017). Mutations within *CHEK2* are observed within 2-5% of inherited breast cancer cases (Walsh *et al.*, 2006, Desrichard *et al.*, 2011, Jalilvand *et al.*, 2017), indicating that we would expect to identify mutations within 2 – 6 individuals within this patient cohort. Analysis of the AmpliSeq primer design indicated that only 88.1% coverage was achieved for *CHEK2.* There are several key mutations within *CHEK2* that are known to be associated with breast cancer development (c.1100delC, p.R117G, p.delE161) (Sodha *et al.*, 2006). Analysis of sequencing results indicated no difference in coverage between CHEK2 and the other sequenced genes and no

issues with mapping were observed. The CHEK2:p.I157T missense variant was identified in this patient cohort, which has documented conflicting interpretations of pathogenicity (uncertain significance/likely pathogenic). This variant has been the subject of multiple functional assays and meta analyses to further understand its role in cancer predisposition (Nevanlinna and Bartek, 2006, Liu *et al.*, 2012). The p.I157T mutation occurs within the forkhead associated and kinase domains of CHEK2, resulting in homodimerization, autophosphorylation and impaired binding of BRCA1 (Cai *et al.*, 2009). However, this variant has been observed at a high frequency in European control populations (MAF 4.8 – 7.4% (Cybulski *et al.*, 2004, Kilpivaara *et al.*, 2004)), raising concerns about the significance of this variant as a risk factor (Apostolou and Papasotiriou, 2017). Despite the demonstrated effects of this variant on normal CHEK2 function, studies have identified that p.I157T is not associated with an increased risk of early death, cancer-associated death, metastasis or relapse (Muranen *et al.*, 2016). The lack of clarity surrounding the penetrance and cancer development risk associated with the inheritance of the p.I157T mutation clearly demonstrates the genetic purgatory associated with the reporting of these ambiguous variants to patients, as the clinical actionability of this variant is unclear.

Mutations within genes such as *MRE11A, RAD51D* and *XRCC2* are not often identified, nor are there management guidelines developed for a range of these genes (Apostolou and Fostira, 2013). Therefore, whilst it may be beneficial to screen these genes, unless the identified sequence variants are previously documented and validated pathogenic mutations, the variants are determined to be of unknown significance and as such are not clinically actionable.

7.2.3.2 Discovery genes

The functional implication of variants identified within the discovery portion of the gene panel presents additional challenges. As the role of these genes in hereditary breast cancer has not yet been demonstrated, the consequence of variants within these genes in disease pathogenesis is unclear. Whilst the potentially pathogenic variants identified within these genes have been subjected to detailed *in silico* analysis, the role of these genes, particularly in inherited cancer and disease pathogenesis is not well characterised. Whilst the analysis of putative cancer susceptibility genes within a targeted pathway results in a directed analysis of these individuals, the benefit is often outweighed by the identification of multiple VUS.

There were 7 discovery genes included in this analysis that returned a high number of sequence variants within this patient cohort (*EP300, HLTF, KAT2B, PRKDC, RPA1, RPS6KA1 and SLC19A1*). The increase in variant detection in the majority of these genes is attributed to the size of the gene itself (*EP300, PRKDC*). However, an increased level of variability is observed within the remaining genes. This could be attributed to general heterogeneity within the population selected or could indicate that these variants lie within a poorly conserved region of the gene. Conversely, these sequence changes may modify protein function and result in cancer predisposition with the affected individuals. This supports the premise of the work carried out within this thesis, as it is hypothesised that additional breast cancer predisposition genes must exist. Whilst the majority of the rare sequencing variants identified within these genes had an unknown functional effect, it is these sequence variants that may benefit from further functional analysis to determine the role of these particular genes in cancer predisposition.

Further analysis of the 32 discovery genes included in the panel illustrated there were 10 genes in which minimal variants were detected (*BRCC3*, *E2F1, E2F3, E2F4, E2F6, RFC2, RFC3, RPRM, SFN,* and *SMARCD2*). A majority of these genes have been shown to play a role in various aspects of cell cycle control and DNA damage repair, with a pathogenic variant within any of these genes predicted to result in an analogous phenotype to *BRCA1/2* mutated cancers. While these genes are known to function in similar pathways to *BRCA1/2*, their role in tumourigenesis has not been investigated. These genes have a multitude of roles within the cell, and whilst some of these genes do have just cause for being included on the gene panel, further work is required to understand their potential role in the development of breast cancer. However, their inclusion on the panel and analysis within this cohort resulted in the addition of variants of uncertain significance within genes that have no clearly pathogenic role in breast cancer predisposition. In order to combat this reoccurring issue of multiple VUS identified within patient samples, it may be beneficial to analyse a smaller number of genes, and therefore not include these genes in future analyses without compelling evidence. This analysis, particularly in regard to the diagnostic genes, has proven to be complex. There is minimal clinical and translational impact as the majority of variants identified within this study are not diagnostically significant. This emphasises the need for a greater effort to resolve the ambiguity surrounding the vast amount of VUS through functional analyses.

7.2.3.3 Genes not included on the AmpliSeq panel

Furthermore, since the development of this panel, there are more genes that have been implicated in the development of hereditary breast cancer that would be beneficial for screening purposes. As more studies identify new susceptibility loci involved in cancer predisposition, it would be beneficial to add these to the panel for screening. However, due to the highly multiplexed nature of these panels and the large quantities provided when ordered, it is not possible to modify them, nor financially feasible to redesign the panel. This indicates that whilst gene panel approaches are useful for targeted analysis, there will always be regions that are overlooked through this methodology. Moreover, *BRCA1* and *BRCA2* are vital DNA damage repair proteins, however their function is not exclusively limited to the pathways analysed within this study. Due to this, there may be mutations within genes that function in a similar manner, within other significant pathways that have not been included in this analysis.

7.3 Decline and limitations of Ion Torrent Sequencing

At the conception of this study, the Ion Torrent PGM was the only sequencing platform at the Flinders Genomics Facility in 2013. Since then, Illumina platforms have dominated the market, and are now the predominant sequencing platform offered by sequencing facilities worldwide. Furthermore, most sequencing facilities have since decommissioned their Ion Torrent platforms, as is the situation with the sequencing facility at Flinders University. This domination by Illumina attributed in part to the short shelf life of Ion Torrent sequencing reagents, and the large number of reactions provided in each kit. Illumina sell their reagents in single use kits, with minimal reagents required, whilst Ion Torrent sequencing kits are provided in multiples of 4, with multiple kits required for each different component of the sequencing process. This is a costly initial outlay, with reagents often expiring before all reactions have been utilised. Whilst the chemistry of the Ion Torrent platforms was comparable to Illumina, the poor marketing choices have led to the demise of the Ion Torrent sequencing, with Illumina now having a monopoly on the MPS market.

If this panel were to be run on a larger cohort of individuals, all samples would have to be sent to the sequencing facility at the University of Western Australia as they are one of the only facilities worldwide to still offer Ion Torrent Sequencing. As this panel is custom designed to be run on the Ion Torrent, it is unclear if it could be manipulated to be utilised for sequencing on other platforms. It may be possible to carry out a pilot study to determine is this panel could be utilised for sequencing on the Illumina platforms, which has not previously been carried out. This could be

achieved through the preparation of the Ion Torrent AmpliSeq barcoded libraries, with the addition of Illumina adapters which are required for the reversible terminator sequencing chemistry associated with Illumina sequencing. Alternatively, AmpliSeq custom panels have recently become available for Illumina sequencing, so the panel could be redesigned and synthesised for use on the Illumina MiSeq or HiSeq. This may result in a higher level of coverage for the selected genes on the panel. However, this then becomes an issue of gene panel versus whole exome sequencing as previously discussed.

Furthermore, there were several issues associated with Ion Torrent sequencing which became apparent throughout this project. Not only were there issues encountered with the sequencing chemistry (errors due to incorrect base calls within homopolymer stretches or at the 3' or 5' ends of sequencing reads), but also there were issues with the sequencing itself. Throughout the multiple rounds of sequencing carried out, a constant issue observed was associated with polyclonality (refer to **Appendix D** for all individual sequencing run data). Polyclonality occurs when more than one strand of template DNA is attached to an ISP. This type of error was a significantly recurring issue, with some sequencing runs exceeding the maximum level of 40% (as advised by Life Technologies). In order to try and overcome this, libraries were quantified using both the BioAnalyser and qPCR-based approaches prior to library pooling for sequencing, in addition to running half the recommended library concentration on the sequencing chips as recommended by the Life Technologies field application specialists. However, these changes were not able to significantly reduce the number of ISPs being lost to polyclonality. In contrast, Illumina sequencing is carried out through bridge amplification, and is not prone to this issue. Furthermore, the paired-end sequencing approach used by Illumina is more accurate. These clear advantages over Ion Torrent sequencing chemistries may have also contributed to the downfall of Ion Torrent sequencing.

7.4 Other biological mechanisms responsible for familial cancer

Whilst this study has focussed on the identification of DNA sequence variants, it is essential to understand that mutations within coding DNA are not the only mutational mechanisms that can affect gene function, leading to an increased susceptibility and the development of breast cancer. Various other mechanisms have the potential to result in an increased susceptibility to inherited cancers and should be considered to improve the success rate of the identification of pathogenic causes of breast cancer.

Mutations within the promotor regions, 5' and 3'UTRs, or deep intronic mutations that influence splicing will be missed by the targeted gene panel. Mutations within the UTRs can affect mRNA stability and ribosomal loading capabilities, which may result in transcripts becoming candidates for nonsense mediated decay. Furthermore, mutations within the UTRs can affect miRNA binding, thereby having a silencing effect or resulting in over-expression of genes (Shen *et al.*, 2008, Chang and Sharan, 2012, Li *et al.*, 2012). Mutations within these regions have the potential to act as the mutational driver or enhance the mutability of the developed tumour. Furthermore, miRNA downregulation caused by hypermethylation of miRNA promoters is often observed in cancers (Portela and Esteller, 2010). For example, it has been demonstrated that miR-124a is epigenetically silenced in a variety of tumours, including breast cancer (Lv *et al.*, 2011, Wong *et al.*, 2011a, Shi *et al.*, 2012). This down regulation results in an up regulation of its target, CDK6, leading to the phosphorylation of the retinoblastoma protein (Rb), which then contributes to the abnormal proliferation of tumour cells involved in the invasion-metastasis cascade observed in breast cancer (Lv *et al.*, 2011). Studies have also shown that *BRCA1/2* mutation-positive cancers, in addition to sporadic breast cancers, are associated with hyper-expression of specific miRNA profiles. Conversely, hereditary cancers that are not attributed to *BRCA1/2* mutations demonstrated hypo-expression of these same miRNAs (Murria Estal *et al.*, 2013). The expression of specific subclasses of miRNAs have been associated with HER2 over-expression, *CDKN2A* and *CDH1* mutations and methylation status, implying that these miRNAs contribute to the driving role of these genetic aberrations in breast cancer. Moreover, miRNAs located within regions with recurrent genomic mutations, may play a role in driving breast cancer development (Riaz *et al.*, 2013). This demonstrates that it may be beneficial to analyse miRNA expression within a subset of the selected patient cohort. miR-expression arrays are a rapid approach that could be utilised within this patient cohort to facilitate the identification sporadic cancers, and those with generic or epigenetic changes that are responsible of the BRCA-like phenotype observed with inherited cancers.

Furthermore, analysis of not only but DNA, but also RNA has demonstrated that a change in gene expression plays a role in disease pathogenesis. Alterations in gene expression and analysis of gene expression profiles, through RNA-seq and transcriptome-wide association studies have recently been utilised in breast cancer studies and have resulted in the identification of novel genes, antisense transcripts and long non-coding RNAs (lncRNAs) that play a role in breast cancer development (Michailidou *et al.*, 2017, Wu *et al.*, 2018). Through the analysis of expression changes of respective genes, particularly those of interest included on the custom AmpliSeq panel, it may be

possible to identify mechanisms of gene regulation and expression that are resulting in cancer development (Winkler and Wiemann, 2016).

It has been shown that altered expression of long non-coding RNAs (lncRNAs) can be associated with tumorigenesis, metastasis and tumour progression (Richard and Eichhorn, 2018). Therefore, it is feasible that these individuals have mutations within the non-coding regions of their genome, affecting lncRNA expression and function. As this is a relatively new field of research, there are a large proportion of identified lncRNAs that have not been functionally characterised, and as such their biological functions in critical cellular processes are not fully understood. However, analysis of transcriptome profiles has identified that thousands of lncRNAs are aberrantly expressed or mutated in various cancers (Bhan *et al.*, 2017). Studies have identified that lncRNAs have been linked to breast cancer initiation, progression and metastasis, in addition to limiting sensitivity to specific targeted therapeutics. Furthermore, deregulated expression of lncRNAs has been observed in a multitude of cancers, including breast cancer (Ding *et al.*, 2014). Often, it is estimated that that approximately 85% of the mutational load resulting in mendelian diseases come from the exome (Lifton, 2010, Braun *et al.*, 2016). However, this is mainly due to ascertainment bias, as the exome is the most well understood and is predominantly studied for the identification of disease-causing mutations. As studies move beyond the coding regions of the genome (2%), it is now recognised that more than 75% of the genome is functional, encoding a large number of non-coding RNAs with a plethora of functions (Bhan *et al.*, 2017). As such, it may be necessary to look beyond the coding regions within a proportion of the individuals included in this study.

Furthermore, despite the large number of GWAS and familial linkage studies that have been carried out, another highly penetrant BRCA-like gene is unlikely. Therefore, there may be a multiplicative effect of pathogenic mutations within low- and mid-penetrance susceptibility genes that are resulting in cancer within these mutation-negative individuals.

7.5 Role of UIMC1 in breast cancer

As discussed in detail in **Chapter 6**, UIMC1 has been shown to be a crucial protein in the DNA damage repair pathway and is required for the initiation of key processes in the recruitment of BRCA1 to the sites of DNA damage for repair (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). It is also shown that cells lacking UIMC1 displayed increased sensitivity to ionising radiation and a reduced capability to repairing DNA DSBs induced by ionising radiation 1 to 2 hours post-irradiation exposure

(Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007b). Recently, work carried out by Jin *et al.* (2019) analysed the expression and biological function of UIMC1 in breast cancer cell lines and matched normal breast tissue and tumour samples. This study demonstrated that *UIMC1* expression was decreased in tumour biopsies, with the lowest levels of *UIMC1* expression associated with a more severe prognosis. These studies these results strongly support a role for UIMC1 in breast cancer.

Importantly, the results obtained from this study were in support of this previous work, but also found that whilst cells lacking *UIMC1* showed a reduced capability to repair DNA DSBs, they took longer to form nuclear foci and show signs of repairing DNA DSBs as determined by the γH2AX analysis carried out. These results may suggest that an alternative pathway of DNA DSB repair has been activated within UIMC1-mutated and UIMC1-deficient cells. Furthermore, the results obtained from this analysis indicated that cells with mutations within the zinc finger domain of UIMC1 also show a reduction in DNA damage repair capabilities, illustrating that this region may have an important role in the function of UIMC1. This is significant and should be investigated further as the function of this region is still poorly understood. Whilst these are only preliminary findings, and further functional validation is required, this data demonstrated that cells lacking UIMC1 or containing mutations within the highly conserved zinc finger domains of UIMC1 show dysregulation of vital cellular processes such as cell cycling and DNA damage repair capabilities, which could indeed play a role in cancer predisposition.

7.6 Future directions

In order to further characterise and understand the role of *UIMC1* and its potential involvement in breast cancer development, but also the identified UIMC1:c.1690T>C variant and its potential pathogenicity, it is necessary to carry out further functional work. In addition to an increased number of replicates for the work already carried out, it would be beneficial to generate cell lines containing the identified mutation of interest. This would provide not only more power but may also increase the statistical significance of the results observed. Furthermore, cell lines and any other models utilised for functional validation of UIMC1 should be subjected to whole genome sequencing, in order to ascertain that no off-target mutations have been introduced due to the CRISPR/Cas9 editing technique employed. In addition to this, RNAseq would be informative for the analysis of changes in gene expression based on manipulation via CRISPR in addition to helping understand the changes observed with an increase in cell passage number.

It is important to note that the number of individuals screened within this study is small by current standards, with only 132 affected individuals analysed. Future work would involve not only increasing the number of individuals screened, but also analysis of matched controls. In order to fully examine the role of these putative cancer susceptibility genes, it is necessary to carry out a much larger screening study and include control individuals. Alternatively, whole exome sequencing could be utilised on the individuals already screened within this cohort to exclude mutations within other known cancer susceptibility genes not included on this panel.

Furthermore, it is necessary to robustly illustrate that a gene is indeed associated with breast cancer predisposition and pathogenesis. This is not only achieved through sequencing analysis, but in-depth functional analysis. Interrogation of the designed amplicon panel indicated that only 95.5% coverage of *UIMC1* was achieved (**Appendix B**). Due to this, it would be beneficial to sequence the missed regions of this gene to ensure no other potentially pathogenic variants have been missed in the patient cohort. Further to this, a more robust characterisation of the functional validation of the role of the identified gene in cancer predisposition must be established. In 2015, *RECQL* was reported as a breast cancer susceptibility gene with initial studies illustrating an association between loss of function variants in *RECQL* and an increased risk of breast cancer (Cybulski *et al.*, 2015, Sun *et al.*, 2015a). However, further studies have failed to support this association and have emphasised that there in insufficient evidence to categorise *RECQL* as a cancer predisposition gene (Kwong *et al.*, 2016, Li *et al.*, 2018, Nguyen-Dumont *et al.*, 2018, Bowden and Tischkowitz, 2019). This emphasises the need for a robust characterisation of the functional role of the identified predisposition genes in cancer in addition to analysis in a high-powered genetic analysis to fully elucidate the role of the identified genes in cancer predisposition.

Unfortunately, due to the technical difficulties and limitations discussed in **Chapter 6**, it was not possible to evaluate the role of the UIMC1:c.1690T>C variant identified within the patient cohort. In order to characterise the effect of the identified sequence variant and its role on DNA damage repair abilities and cell proliferation, it is necessary to successfully introduce this SNP into selected cell lines. As previously discussed, the recently established and more successful base editing mechanism may be a more effective system to utilise for the introduction of this base change within cell lines, rather than the low efficiency, error prone CRISPR/Cas9 HDR mechanism of gene editing that was utilised within this study. As mentioned in **Section 6.4.8**, base editing enables the

irreversible conversion of a desired DNA base to another without the generation of DNA DSBs through the use of programmable nucleotide deaminases (Komor *et al.*, 2016, Gaudelli *et al.*, 2017).This mechanism can be utilised for the introduction of C>T mutations (through C>U deamination) and A>G mutations (through A>I deamination) (Gaudelli *et al.*, 2017) with significantly greater success rates (35-75% success) than that the HDR mechanism (0.5-20% success) utilised within this study. As the variant identified within UIMC1 was a C>T change, this mechanism of editing is the most promising method for the successful introduction of the variant into the desired cells.

Furthermore, the functional work carried out within this thesis was only conducted within HEK293 cells. For future work, it would be beneficial to carry out this work in a more physiologically relevant cell line (i.e. a 'normal' breast cell line), or even within a murine model. Functional studies within a mouse model would be beneficial as they would allow for a more complete analysis of the function of UIMC1 within an entire organism. The generation of mice lacking UIMC1 or containing desired mutations would allow for a more thorough analysis of DNA damage repair pathways and in-depth analysis of any tumour development. Mice models have been used to further the understanding of *BRCA1* and *BRCA2* and cancer predisposion, however homozygous knockout models often display embryonic lethality (Gowen *et al.*, 1996, Hakem *et al.*, 1996, Moynahan *et al.*, 1999, Evers and Jonkers, 2006). This suggests that it would be beneficial to generate models with a partial knockout, or containing the variant of interest, rather than a complete knockout. Mouse models would also allow for a comparison between normal and perturbed cellular function, focussing on cell proliferation, apoptosis and cell cycle regulation in WT and *UIMC1*-mutated models. This would gain an increased understanding on the role of UIMC1 in cancer predisposition and tumour development in a more representative model.

In order to more completely understand the function of UIMC1 within the DNA damage repair pathway, it would be beneficial to analyse the effect of UIMC1 on apoptosis. Recently, cells lacking UIMC1 have shown an increased level of apoptosis in comparison to WT cells (Jin *et al.*, 2019). It would have been beneficial to analyse expression of UIMC1 in the knockout cells generated for this study to determine if there was a change in expression observed within the generated cell lines. Moreover, additional work needs to be carried out to analyse the presence and formation of UIMC1 localisation with respect to γH2AX foci. As previously discussed, UIMC1 is known to form discrete foci which overlap with the γH2AX foci formed at the site of DNA damage repair (Sobhian *et al.*,

2007, Wang *et al.*, 2007, Yan *et al.*, 2007a). Therefore, using additional fluorophores, it would be possible to analyse UIMC1 foci formation in both mutant and WT cells over time. This would not only allow for further understanding of the localisation of UIMC1 to sites of DNA DSBs and the activation of these pathways, but also the other proteins involved in DNA damage repair in cells lacking functional UIMC1. Through this, it may be possible to gain further insight into the alternative pathways utilised for key cellular processes such as DNA DSB repair in the absence of UIMC1.

As indicated by this study, it may be beneficial to carry out the analysis of vital cellular pathways, such as DNA damage repair, on cells with a lower passage number and in a greater number of replicates. Despite passage numbers of cells remaining low (<15 passages), a significant decrease was observed in experimental results obtained for the final replicate of γH2AX formation that was carried out on cells with an increased passage number. It may be beneficial to carry out these experiments on cells with sequential passage numbers, or frozen stocks of the same passage to better understand the repair capabilities of the cells. This is not surprising as an increase in passage number has been shown to affect cell morphology, response to stimuli, growth rate, transfection efficiency and even protein expression (Chang-Liu and Woloschak, 1997, Esquenet *et al.*, 1997, Yu *et al.*, 1997, Wenger *et al.*, 2004). Additionally, increased passaging of cells lacking this functional DNA damage repair pathway may accumulate mutations, significantly altering cellular function and the observed phenotype. Repetition of these experiments is required in order to gain further insight into the DNA damage repair capabilities of cells lacking UIMC1.

In addition to further work to functionally validate the role of UIMC1, an expanded sequencing study for the analysis of *UIMC1* within a breast cancer cohort would be beneficial. Further sequencing analysis of a greater number of *BRCA1/2* mutation-negative individuals that are available for analysis from SA Pathology may help determine if *UIMC1* is a potential breast cancer susceptibility gene. At the commencement of this study, a list of 698 *BRCA1/2* mutation-negative individuals that had been referred for genetic screening between June 2005 and June 2014 was provided. One hundred and twenty of these individuals have been screened within this study, leaving an additional 578 samples that require further analysis. From the 120 samples screened within this study, 5 potentially pathogenic mutations were identified in UIMC1 in 8 individuals (~7%). To further understand the role of UIMC1 in cancer development, it would be beneficial to functionally validate these variants as well. Furthermore, it may be beneficial to not only screen individuals with familial cancer, but also screen *UIMC1* within individuals with sporadic cancer to determine its role in cancer

susceptibility and development. In conjunction with this sequencing study, it may also be beneficial to analyse *UIMC1* expression within these individuals, as recent studies have demonstrated that *UIMC1* expression is significantly lower in cancer cells, with a reduction in mRNA expression related to tumour size and metastasis (Jin *et al.*, 2019). This study highlights the need to not only analyse UIMC1 for any sequence variants within an expanded cohort, but also to look at the expression of this important gene. Initially, a comparison between expression in the breast cancer tumour and within the blood would be beneficial to see if there is any correlation, as this has not yet been looked at, and analysing UIMC1 expression from the tumour tissue itself would not be a practical diagnostic test.

7.7 Conclusions

This study has demonstrated the utility of expanding mutational screening beyond *BRCA1* and *BRCA2.* The analysis of known cancer susceptibility genes within the patient cohort resulted in the identification of a known pathogenic mutation within two individuals within the patient cohort. This *PALB2* mutation was previously undetected within these individuals, allowing for additional family members to undergo cascade testing and providing them the opportunity for increased surveillance. Furthermore, this study has demonstrated the benefit in screening beyond even the known breast cancer susceptibility genes in mutation-negative individuals with familial breast cancer. Through the analysis of genes which function within the same pathways as BRCA1 and BRCA2, we have identified *UIMC1* as a gene which may be implicated in cancer development.

Using CRISPR/Cas9, we have shown that *UIMC1* is important for DNA damage repair, with *UIMC1*-mutant cells taking longer to repair DNA DSBs induced by ionising radiation. It has previously been shown that UIMC1 is required for the recruitment of BRCA1 to sites of DNA damage (Sobhian *et al.*, 2007, Wang *et al.*, 2007, Yan *et al.*, 2007a), and this in conjunction with our results suggests that cells lacking *UIMC1* may be relying on an alternative pathway for the recruitment of BRCA1 to the site of DNA damage to facilitate repair. Furthermore, these cells indicated dysregulated cell growth and increased sensitivity to ionising radiation. These observations support the idea that UIMC1 is required for cell cycle regulation and DNA damage repair; vital pathways which when are mutated often have the capability to result in mutagenesis. Whilst these are only preliminary findings, and further work is required, this study provides evidence that UIMC1 is a key protein within these DNA damage repair and cell cycle control pathways which, when mutated has the ability to play a role in the development of breast cancer.

Chapter 8:
References

Abdel-Latif, A. & Osman, G. 2017. Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize. *Plant Methods,* Vol. 13, No. 1**,** 1.

Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. 2013. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics,* Vol. 76, No. 1**,** 7.20-27.20.41.

Aguilera, A. & Garcia-Muse, T. 2012. R loops: from transcription byproducts to threats to genome stability. *Mol Cell,* Vol. 46, No. 2**,** 115-124.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. & Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology,* Vol. 12, No. 2**,** R18.

Aird, E. J., Lovendahl, K. N., St. Martin, A., Harris, R. S. & Gordon, W. R. 2018. Increasing Cas9-mediated homology-directed repair efficiency through covalent tethering of DNA repair template. *Communications Biology,* Vol. 1, No. 1**,** 54.

Akbari, M. R., Ghadirian, P., Robidoux, A., Foumani, M., Sun, Y., Royer, R., Zandvakili, I., Lynch, H. & Narod, S. A. 2009. Germline RAP80 mutations and susceptibility to breast cancer. *Breast Cancer Res Treat,* Vol. 113, No. 2**,** 377-381.

Akil, A., Ezzikouri, S., El Feydi, A. E., Benazzouz, M., Afifi, R., Diagne, A. G., Benjouad, A., Dejean, A., Pineau, P. & Benjelloun, S. 2012. Associations of genetic variants in the transcriptional coactivators EP300 and PCAF with hepatocellular carcinoma. *Cancer Epidemiol,* Vol. 36, No. 5**,** e300-305.

Akisik, E., Yazici, H. & Dalay, N. 2011. ARLTS1, MDM2 and RAD51 gene variations are associated with familial breast cancer. *Mol Biol Rep,* Vol. 38, No. 1**,** 343-348.

Al-Hebshi, N. N., Li, S., Nasher, A. T., El-Setouhy, M., Alsanosi, R., Blancato, J. & Loffredo, C. 2016. Exome sequencing of oral squamous cell carcinoma in users of Arabian snuff reveals novel candidates for driver genes. *Int J Cancer,* Vol. 139, No. 2**,** 363-372.

Ali, R., Rakha, E. A., Madhusudan, S. & Bryant, H. E. 2017. DNA damage repair in breast cancer and its therapeutic implications. *Pathology,* Vol. 49, No. 2**,** 156-165.

Alkan, C., Coe, B. P. & Eichler, E. E. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet,* Vol. 12, No. 5**,** 363-376.

Allinen, M., Huusko, P., Mantyniemi, S., Launonen, V. & Winqvist, R. 2001. Mutation analysis of the CHK2 gene in families with hereditary breast cancer. *Br J Cancer,* Vol. 85, No. 2**,** 209-212.

Altmann, A., Weber, P., Quast, C., Rex-Haffner, M., Binder, E. B. & Müller-Myhsok, B. 2011. vipR: variant identification in pooled DNA using R. *Bioinformatics,* Vol. 27, No. 13**,** 77-84.

Amir, E., Freedman, O. C., Seruga, B. & Evans, D. G. 2010. Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. *JNCI: Journal of the National Cancer Institute,* Vol. 102, No. 10**,** 680-691.

Amit, I., Citri, A., Shay, T., Lu, Y., Katz, M., Zhang, F., Tarcic, G., Siwak, D., Lahad, J., Jacob-Hirsch, J., Amariglio, N., Vaisman, N., Segal, E., Rechavi, G., Alon, U., Mills, G. B., Domany, E. & Yarden, Y. 2007. A module of negative feedback regulators defines growth factor signaling. *Nature Genetics,* Vol. 39, No. 4**,** 503-512.

Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'alfonso, S. & De Bellis, G. 2016. Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. *Scientific Reports,* Vol. 6, No. 1**,** 33735.

Anderson, C. W. & Lees-Miller, S. P. 1992. The nuclear serine/threonine protein kinase DNA-PK. *Critical Reviews in Eukaryotic Gene Expression,* Vol. 2, No. 4**,** 283-314.

Anensen, N., Skavland, J., Stapnes, C., Ryningen, A., Borresen-Dale, A. L., Gjertsen, B. T. & Bruserud, O. 2006. Acute myelogenous leukemia in a patient with Li-Fraumeni syndrome treated with valproic acid, theophyllamine and all-trans retinoic acid: a case report. *Leukemia,* Vol. 20, No. 4**,** 734-736.

Antoniou, A., Hardy, R., Walker, L., Evans, D., Shenton, A., Eeles, R., Shanley, S., Pichert, G., Izatt, L., Rose, S., Douglas, F., Eccles, D., Morrison, P. J., Scott, J., Zimmern, R., Easton, D. F. & Pharoah, P. D. P. 2008. Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *Journal of Medical Genetics,* Vol. 45, No. 7**,** 425-431.

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, A., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tulinius, H., Thorlacius, S., Eerola, H.,

Nevanlinna, H., Syrjakoski, K., Kallioniemi, O. P., Thompson, D., Evans, C., Peto, J., Lalloo, F., Evans, D. G. & Easton, D. F. 2003. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *American Journal of Human Genetics,* Vol. 72, No. 5**,** 1117-1130.

Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkas, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F., Tomiak, E., Neuhausen, S. L., Teo, Z. L., Khan, S., Aittomaki, K., Moilanen, J. S., Turnbull, C., Seal, S., Mannermaa, A., Kallioniemi, A., Lindeman, G. J., Buys, S. S., Andrulis, I. L., Radice, P., Tondini, C., Manoukian, S., Toland, A. E., Miron, P., Weitzel, J. N., Domchek, S. M., Poppe, B., Claes, K. B., Yannoukakos, D., Concannon, P., Bernstein, J. L., James, P. A., Easton, D. F., Goldgar, D. E., Hopper, J. L., Rahman, N., Peterlongo, P., Nevanlinna, H., King, M. C., Couch, F. J., Southey, M. C., Winqvist, R., Foulkes, W. D. & Tischkowitz, M. 2014. Breast-cancer risk in families with mutations in PALB2. *New England Journal of Medicine,* Vol. 371, No. 17**,** 497-506.

Antoniou, A. C., Durocher, F., Smith, P., Simard, J. & Easton, D. F. 2006. BRCA1 and BRCA2 mutation predictions using the BOADICEA and BRCAPRO models and penetrance estimation in high-risk French-Canadian families. *Breast Cancer Research,* Vol. 8, No. 1**,** R3.

Antoniou, A. C. & Easton, D. F. 2006. Models of genetic susceptibility to breast cancer. *Oncogene,* Vol. 25, No. 43**,** 5898-5905.

Antoniou, A. C., Pharoah, P. D., Mcmullan, G., Day, N. E., Stratton, M. R., Peto, J., Ponder, B. J. & Easton, D. F. 2002. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *British Journal of Cancer,* Vol. 86, No. 1**,** 76-83.

Antoniou, A. C., Pharoah, P. P., Smith, P. & Easton, D. F. 2004. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British Journal of Cancer,* Vol. 91, No. 8**,** 1580-1590.

Apostolou, P. & Fostira, F. 2013. Hereditary breast cancer: the era of new susceptibility genes. *BioMed Research International,* Vol. 2013, No. 747318**,** 1-11.

Apostolou, P. & Papasotiriou, I. 2017. Current perspectives on CHEK2 mutations in breast cancer. *Breast cancer,* Vol. 12, No. 9**,** 331-335.

Arany, Z., Newsome, D., Oldread, E., Livingston, D. M. & Eckner, R. 1995. A family of transcriptional adaptor proteins targeted by the E1A oncoprotein. *Nature,* Vol. 374, No. 6517**,** 81-84.

Arnold, N., Peper, H., Bandick, K., Kreikemeier, M., Karow, D., Teegen, B. & Jonat, W. 2002. Establishing a control population to screen for the occurrence of nineteen unclassified variants in the BRCA1 gene by denaturing high-performance liquid chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci,* Vol. 782, No. 1**,** 99-104.

Ascano, J. J., Frierson, H., Jr., Moskaluk, C. A., Harper, J. C., Roviello, F., Jackson, C. E., El-Rifai, W., Vindigni, C., Tosi, P. & Powell, S. M. 2001. Inactivation of the E-cadherin gene in sporadic diffuse-type gastric cancer. *Modern Pathology,* Vol. 14, No. 10**,** 942-949.

Australian Government. 2015. *Breast Screening* [Online]. Available: http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/breast-campaign-home [Accessed 7 October 2015].

Australian Institue of Health and Welfare. 2016. *Cancer data in Australia* [Online]. Canberra: Australian Government. Available: https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/summary [Accessed 21 March 2016].

Australian Institute of Health and Welfare. 2013. *Cancer in Australia: key facts* [Online]. Canberra: Australian Government. [Accessed September 1 2014].

Avci-Adali, M., Behring, A., Keller, T., Krajewski, S., Schlensak, C. & Wendel, H. P. 2014. Optimized conditions for successful transfection of human endothelial cells with in vitro synthesized and modified mRNA for induction of protein expression. *Journal of biological engineering,* Vol. 8, No. 1**,** 8-8.

Bambury, R. M., Bhatt, A. S., Riester, M., Pedamallu, C. S., Duke, F., Bellmunt, J., Stack, E. C., Werner, L., Park, R., Iyer, G., Loda, M., Kantoff, P. W., Michor, F., Meyerson, M. & Rosenberg, J. E. 2015. DNA copy number analysis of metastatic urothelial carcinoma with comparison to primary tumors. *BMC Cancer,* Vol. 15, No. 1**,** 242.

Banin, S., Moyal, L., Shieh, S., Taya, Y., Anderson, C. W., Chessa, L., Smorodinsky, N. I., Prives, C., Reiss, Y., Shiloh, Y. & Ziv, Y. 1998. Enhanced phosphorylation of p53 by ATM in response to DNA damage. *Science,* Vol. 281, No. 5383**,** 1674-1677.

Bansal, V. 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics,* Vol. 26, No. 12**,** 318-324.

Bansal, V., Tewhey, R., Leproust, E. M. & Schork, N. J. 2011. Efficient and Cost Effective Population Resequencing by Pooling and In-Solution Hybridization. *PLOS ONE,* Vol. 6, No. 3**,** e18353.

Bartkova, J., Tommiska, J., Oplustilova, L., Aaltonen, K., Tamminen, A., Heikkinen, T., Mistrik, M., Aittomaki, K., Blomqvist, C., Heikkila, P., Lukas, J., Nevanlinna, H. & Bartek, J. 2008. Aberrations of the MRE11-RAD50-NBS1 DNA damage sensor complex in human breast cancer: MRE11 as a candidate familial cancer-predisposing gene. *Molecular Oncology,* Vol. 2, No. 4**,** 296-316.

Bayraktar, S. & Gluck, S. 2012. Systemic therapy options in BRCA mutation-associated breast cancer. *Breast Cancer Res Treat,* Vol. 135, No. 2**,** 355-366.

Beasley, W. D., Beynon, J., Jenkins, G. J. & Parry, J. M. 2008. Reprimo 824 G>C and p53R2 4696 C>G single nucleotide polymorphisms and colorectal cancer: a case-control disease association study. *Int J Colorectal Dis,* Vol. 23, No. 4**,** 375-381.

Behrmann, L., Mccomb, S., Aguadé-Gorgorió, J., Huang, Y., Hermann, M., Pelczar, P., Aguzzi, A., Bourquin, J.-P. & Bornhauser, B. C. 2017. Efficient Generation of Multi-gene Knockout Cell Lines and Patient-derived Xenografts Using Multi-colored Lenti-CRISPR-Cas9. *Bio-protocol,* Vol. 7, No. 7**,** e2222.

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W. & Kingsmore, S. F. 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med,* Vol. 3, No. 65**,** 65ra64.

Bell, D. W., Kim, S. H., Godwin, A. K., Schiripo, T. A., Harris, P. L., Haserlat, S. M., Wahrer, D. C., Haiman, C. A., Daly, M. B., Niendorf, K. B., Smith, M. R., Sgroi, D. C., Garber, J. E., Olopade, O. I., Le Marchand, L., Henderson, B. E., Altshuler, D., Haber, D. A. & Freedman, M. L. 2007. Genetic and functional analysis of CHEK2 (CHK2) variants in multiethnic cohorts. *Int J Cancer,* Vol. 121, No. 12**,** 2661-2667.

Beltrame, L., Di Marino, M., Fruscio, R., Calura, E., Chapman, B., Clivio, L., Sina, F., Mele, C., Iatropoulos, P., Grassi, T., Fotia, V., Romualdi, C., Martini, P., Noris, M., Paracchini, L., Craparotta, I., Petrillo, M., Milani, R., Perego, P., Ravaggi, A., Zambelli, A., Ronchetti, E., D'incalci, M. & Marchini, S. 2015. Profiling cancer gene mutations in longitudinal epithelial ovarian cancer biopsies by targeted next-generation sequencing: a retrospective study. *Annals of Oncology,* Vol. 26, No. 7**,** 1363-1371.

Bendjennat, M., Boulaire, J., Jascur, T., Brickner, H., Barbier, V., Sarasin, A., Fotedar, A. & Fotedar, R. 2003. UV irradiation triggers ubiquitin-dependent degradation of p21(WAF1) to promote DNA repair. *Cell,* Vol. 114, No. 5**,** 599-610.

Bernal, C., Aguayo, F., Villarroel, C., Vargas, M., Diaz, I., Ossandon, F. J., Santibanez, E., Palma, M., Aravena, E., Barrientos, C. & Corvalan, A. H. 2008. Reprimo as a potential biomarker for early detection in gastric cancer. *Clin Cancer Res,* Vol. 14, No. 19**,** 6264-6269.

Bessette, D. C., Tilch, E., Seidens, T., Quinn, M. C. J., Wiegmans, A. P., Shi, W., Cocciardi, S., Mccart-Reed, A., Saunus, J. M., Simpson, P. T., Grimmond, S. M., Lakhani, S. R., Khanna, K. K., Waddell, N., Al-Ejeh, F. & Chenevix-Trench, G. 2015. Using the MCF10A/MCF10CA1a Breast Cancer Progression Cell Line Model to Investigate the Effect of Active, Mutant Forms of EGFR in Breast Cancer Development and Treatment Using Gefitinib. *PloS one,* Vol. 10, No. 5**,** e0125232-e0125232.

Bhan, A., Soleimani, M. & Mandal, S. S. 2017. Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Research,* Vol. 77, No. 15**,** 3965-3981.

Bhatia, V., Barroso, S. I., Garcia-Rubio, M. L., Tumini, E., Herrera-Moyano, E. & Aguilera, A. 2014. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature,* Vol. 511, No. 7509**,** 362-365.

Bhoumik, A., Takahashi, S., Breitweiser, W., Shiloh, Y., Jones, N. & Ronai, Z. E. 2005. ATM-dependent phosphorylation of ATF2 is required for the DNA damage response. *Molecular cell,* Vol. 18, No. 5**,** 577-587.

Bochar, D. A., Wang, L., Beniya, H., Kinev, A., Xue, Y., Lane, W. S., Wang, W., Kashanchi, F. & Shiekhattar, R. 2000. BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. *Cell,* Vol. 102, No. 2**,** 257-265.

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., Kieleczawa, J., Lyons, R. H., Neubert, T. A., Noll, A. C., Singh, S., Steen, R. & Zianni, M. 2013. Comparison of Commercially Available Target

Enrichment Methods for Next-Generation Sequencing. *Journal of Biomolecular Techniques : JBT,* Vol. 24, No. 2**,** 73-86.

Bogdanova, N., Helbig, S. & Dörk, T. 2013. Hereditary breast cancer: ever more pieces to the polygenic puzzle. *Hereditary Cancer in Clinical Practice,* Vol. 11, No. 1**,** 12-12.

Boland, J. F., Chung, C. C., Roberson, D., Mitchell, J., Zhang, X., Im, K. M., He, J., Chanock, S. J., Yeager, M. & Dean, M. 2013. The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Human genetics,* Vol. 132, No. 10**,** 1153-1163.

Bonadona, V., Bonaïti, B., Olschwang, S. & Et Al. 2011. Cancer risks associated with germline mutations in mlh1, msh2, and msh6 genes in lynch syndrome. *JAMA,* Vol. 305, No. 22**,** 2304-2310.

Bonni, A., Brunet, A., West, A. E., Datta, S. R., Takasu, M. A. & Greenberg, M. E. 1999. Cell survival promoted by the Ras-MAPK signaling pathway by transcription-dependent and -independent mechanisms. *Science,* Vol. 286, No. 5443**,** 1358-1362.

Borg, Å., Sandberg, T., Nilsson, K., Johannsson, O., Klinker, M., Måsbäck, A., Westerdahl, J., Olsson, H. & Ingvar, C. 2000. High Frequency of Multiple Melanomas and Breast and Pancreas Carcinomas in CDKN2A Mutation-Positive Melanoma Families. *Journal of the National Cancer Institute,* Vol. 92, No. 15**,** 1260-1266.

Bothmer, A., Phadke, T., Barrera, L. A., Margulies, C. M., Lee, C. S., Buquicchio, F., Moss, S., Abdulkerim, H. S., Selleck, W., Jayaram, H., Myer, V. E. & Cotta-Ramusino, C. 2017. Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nature Communications,* Vol. 8, No. 1**,** 13905.

Bourdon, J. C. 2007. p53 and its isoforms in cancer. *British Journal of Cancer,* Vol. 97, No. 3**,** 277-282.

Bowden, A. R. & Tischkowitz, M. 2019. Clinical implications of germline mutations in breast cancer genes: RECQL. *Breast Cancer Res Treat,* Vol. 174, No. 3**,** 553-560.

Boyault, S., Drouet, Y., Navarro, C., Bachelot, T., Lasset, C., Treilleux, I., Tabone, E., Puisieux, A. & Wang, Q. 2012. Mutational characterization of individual breast tumors: TP53 and PI3K pathway genes are frequently and distinctively mutated in different subtypes. *Breast Cancer Res Treat,* Vol. 132, No. 1**,** 29-39.

Bozhanov, S. S., Angelova, S. G., Krasteva, M. E., Markov, T. L., Christova, S. L., Gavrilov, I. G. & Georgieva, E. I. 2010. Alterations in p53, BRCA1, ATM, PIK3CA, and HER2 genes and their effect in modifying clinicopathological characteristics and overall survival of Bulgarian patients with breast cancer. *J Cancer Res Clin Oncol,* Vol. 136, No. 11**,** 1657-1669.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P. & Tyson, G. W. 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol,* Vol. 9, No. 4**,** e1003031.

Braun, D. A., Schueler, M., Halbritter, J., Gee, H. Y., Porath, J. D., Lawson, J. A., Airik, R., Shril, S., Allen, S. J., Stein, D., Al Kindy, A., Beck, B. B., Cengiz, N., Moorani, K. N., Ozaltin, F., Hashmi, S., Sayer, J. A., Bockenhauer, D., Soliman, N. A., Otto, E. A., Lifton, R. P. & Hildebrandt, F. 2016. Whole exome sequencing identifies causative mutations in the majority of consanguineous or familial cases with childhood-onset increased renal echogenicity. *Kidney international,* Vol. 89, No. 2**,** 468-475.

Braun, S., Lower, K. & Grist, S. 2013. *Next Generation sequence analysis of BRCA1/2-related DNA damage response and checkpoint control pathways in familial breast cancer.* Bachelor of Science (Honours), Flinders Universtiy.

Brisco, M. J., Latham, S., Bartley, P. A. & Morley, A. A. 2010. Incorporation of measurement of DNA integrity into qPCR assays. *Biotechniques,* Vol. 49, No. 6**,** 893-897.

Brose, M. S., Rebbeck, T. R., Calzone, K. A., Stopfer, J. E., Nathanson, K. L. & Weber, B. L. 2002. Cancer Risk Estimates for BRCA1 Mutation Carriers Identified in a Risk Evalutation Program. *Journal of the National Cancer Institute,* Vol. 94, No. 18**,** 1365-1372.

Buzolin, A. L., Moreira, C. M., Sacramento, P. R., Oku, A. Y., Fornari, A. R. D. S., Antonio, D. S. M., Quaio, C. R. D. a. C., Baratela, W. R. & Mitne-Neto, M. 2017. Development and validation of a variant detection workflow for BRCA1 and BRCA2 genes and its clinical application based on the Ion Torrent technology. *Human Genomics,* Vol. 11, No. 1**,** 1-14.

Cai, Z., Chehab, N. H. & Pavletich, N. P. 2009. Structure and activation mechanism of the CHK2 DNA damage checkpoint kinase. *Mol Cell,* Vol. 35, No. 6**,** 818-829.

Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burtt, N. P., Rivas, M., Guiducci, C., Bruno, D. L., Goldberger, O. A., Redman, M. C., Wiltshire, E., Wilson, C. J., Altshuler, D., Gabriel, S. B., Daly,

M. J., Thorburn, D. R. & Mootha, V. K. 2010. High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. *Nature Genetics,* Vol. 42, No. 10**,** 851-858.

Cancer Australia. 2014. *Breast Cancer Statistics* [Online]. Canberra: Australian Government. Available: http://canceraustralia.gov.au/affected-cancer/cancer-types/breast-cancer/breast-cancer-statistics [Accessed 1 September 2014].

Carney, J. P., Maser, R. S., Olivares, H., Davis, E. M., Le Beau, M., Yates, J. R., 3rd, Hays, L., Morgan, W. F. & Petrini, J. H. 1998. The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. *Cell,* Vol. 93, No. 3**,** 477-486.

Carvalho, M. A., Marsillac, S. M., Karchin, R., Manoukian, S., Grist, S., Swaby, R. F., Urmenyi, T. P., Rondinelli, E., Silva, R., Gayol, L., Baumbach, L., Sutphen, R., Pickard-Brzosowicz, J. L., Nathanson, K. L., Sali, A., Goldgar, D., Couch, F. J., Radice, P. & Monteiro, A. N. 2007. Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. *Cancer Res,* Vol. 67, No. 4**,** 1494-1501.

Chan, M., Ji, S. M., Yeo, Z. X., Gan, L., Yap, E., Yap, Y. S., Ng, R., Tan, P. H., Ho, G. H., Ang, P. & Lee, A. S. 2012. Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. *J Mol Diagn,* Vol. 14, No. 6**,** 602-612.

Chan, T. A., Hermeking, H., Lengauer, C., Kinzler, K. W. & Vogelstein, B. 1999. 14-3-3Sigma is required to prevent mitotic catastrophe after DNA damage. *Nature,* Vol. 401, No. 6753**,** 616-620.

Chang-Liu, C. M. & Woloschak, G. E. 1997. Effect of passage number on cellular response to DNA-damaging agents: cell survival and gene expression. *Cancer Lett,* Vol. 113, No. 1**,** 77-86.

Chang, S. & Sharan, S. K. 2012. BRCA1 and microRNAs: emerging networks and potential therapeutic targets. *Molecules and cells,* Vol. 34, No. 5**,** 425-432.

Chang, Z., Zhou, H. & Liu, Y. 2014. Promoter methylation and polymorphism of E-cadherin gene may confer a risk to prostate cancer: a meta-analysis based on 22 studies. *Tumour Biol,* Vol. 35, No. 10**,** 10503-10513.

Chen, C. R., Kang, Y., Siegel, P. M. & Massague, J. 2002. E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression. *Cell,* Vol. 110, No. 1**,** 19-32.

Chi, X., Zhang, Y., Xue, Z., Feng, L., Liu, H., Wang, F. & Qi, X. 2014. Discovery of rare mutations in extensively pooled DNA samples using multiple target enrichment. *Plant biotechnology journal,* Vol. 12, No. 6**,** 709-717.

Cho, S. W., Kim, S., Kim, J. M. & Kim, J. S. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol,* Vol. 31, No. 3**,** 230-232.

Cho, S. W., Kim, S., Kim, Y., Kweon, J., Kim, H. S., Bae, S. & Kim, J.-S. 2014. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome research,* Vol. 24, No. 1**,** 132-141.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE,* Vol. 7, No. 10**,** e46688.

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K. & Kühn, R. 2015. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology,* Vol. 33, No. 1**,** 543-548.

Cirino, A. L., Lakdawala, N. K., Mcdonough, B., Conner, L., Adler, D., Weinfeld, M., O'gara, P., Rehm, H. L., Machini, K., Lebo, M., Blout, C., Green, R. C., Macrae, C. A., Seidman, C. E. & Ho, C. Y. 2017. A Comparison of Whole Genome Sequencing to Multigene Panel Testing in Hypertrophic Cardiomyopathy Patients. *Circ Cardiovasc Genet,* Vol. 10, No. 5**,** e001768.

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D. & Kingsmore, S. F. 2018. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Medicine,* Vol. 3, No. 1**,** e16.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A. & Zhang, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science,* Vol. 339, No. 6121**,** 819-823.

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,* Vol. 286, No. 5442**,** 1162-1166.

Costa, J. L., Sousa, S., Justino, A., Kay, T., Fernandes, S., Cirnes, L., Schmitt, F. & Machado, J. C. 2013. Nonoptical massive parallel DNA sequencing of BRCA1 and BRCA2 genes in a diagnostic setting. *Hum Mutat,* Vol. 34, No. 4**,** 629-635.

Coulet, F., Fajac, A., Colas, C., Eyries, M., Dion-Miniere, A., Rouzier, R., Uzan, S., Lefranc, J. P., Carbonnel, M., Cornelis, F., Cortez, A. & Soubrier, F. 2013. Germline RAD51C mutations in ovarian cancer susceptibility. *Clin Genet,* Vol. 83, No. 4**,** 332-336.

Cristofanilli, M., Gonzalez-Angulo, A., Sneige, N., Kau, S. W., Broglio, K., Theriault, R. L., Valero, V., Buzdar, A. U., Kuerer, H., Buchholz, T. A. & Hortobagyi, G. N. 2005. Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes. *J Clin Oncol,* Vol. 23, No. 1**,** 41-48.

Cybulski, C., Carrot-Zhang, J., Kluzniak, W., Rivera, B., Kashyap, A., Wokolorczyk, D., Giroux, S., Nadaf, J., Hamel, N., Zhang, S., Huzarski, T., Gronwald, J., Byrski, T., Szwiec, M., Jakubowska, A., Rudnicka, H., Lener, M., Masojc, B., Tonin, P. N., Rousseau, F., Gorski, B., Debniak, T., Majewski, J., Lubinski, J., Foulkes, W. D., Narod, S. A. & Akbari, M. R. 2015. Germline RECQL mutations are associated with breast cancer susceptibility. *Nat Genet,* Vol. 47, No. 6**,** 643-646.

Cybulski, C., Huzarski, T., Gorski, B., Masojc, B., Mierzejewski, M., Debniak, T., Gliniewicz, B., Matyjasik, J., Zlowocka, E., Kurzawski, G., Sikorski, A., Posmyk, M., Szwiec, M., Czajka, R., Narod, S. A. & Lubinski, J. 2004. A novel founder CHEK2 mutation is associated with increased prostate cancer risk. *Cancer Res,* Vol. 64, No. 8**,** 2677-2679.

Dal Molin, M., Zhang, M., De Wilde, R. F., Ottenhof, N. A., Rezaee, N., Wolfgang, C. L., Blackford, A., Vogelstein, B., Kinzler, K. W., Papadopoulos, N., Hruban, R. H., Maitra, A. & Wood, L. D. 2015. Very Long-term Survival Following Resection for Pancreatic Cancer Is Not Explained by Commonly Mutated Genes: Results of Whole-Exome Sequencing Analysis. *Clin Cancer Res,* Vol. 21, No. 8**,** 1944-1950.

Damiati, E., Borsani, G. & Giacopuzzi, E. 2016. Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. *Human Genetics,* Vol. 135, No. 5**,** 499-511.

Damiola, F., Pertesi, M., Oliver, J., Le Calvez-Kelm, F., Voegele, C., Young, E. L., Robinot, N., Forey, N., Durand, G., Vallée, M. P., Tao, K., Roane, T. C., Williams, G. J., Hopper, J. L., Southey, M. C., Andrulis, I. L., John, E. M., Goldgar, D. E., Lesueur, F. & Tavtigian, S. V. 2014. Rare key functional domain missense substitutions in MRE11A, RAD50, and NBNcontribute to breast cancer susceptibility: results from a Breast Cancer Family Registry case-control mutation-screening study. *Breast Cancer Research,* Vol. 16, No. 3**,** R58.

Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. 2015. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nat Chem Biol,* Vol. 11, No. 5**,** 316-318.

De Magalhaes, J. P., Finch, C. E. & Janssens, G. 2010. Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. *Ageing Res Rev,* Vol. 9, No. 3**,** 315-323.

Debniak, T., Gorski, B., Huzarski, T., Byrski, T., Cybulski, C., Mackiewicz, A., Gozdecka-Grodecka, S., Gronwald, J., Kowalska, E., Haus, O., Grzybowska, E., Stawicka, M., Swiec, M., Urbanski, K., Niepsuj, S., Wasko, B., Gozdz, S., Wandzel, P., Szczylik, C., Surdyka, D., Rozmiarek, A., Zambrano, O., Posmyk, M., Narod, S. A. & Lubinski, J. 2005a. A common variant of CDKN2A (p16) predisposes to breast cancer. *J Med Genet,* Vol. 42, No. 10**,** 763-765.

Debniak, T., Scott, R. J., Huzarski, T., Byrski, T., Rozmiarek, A., Debniak, B., Zaluga, E., Maleszka, R., Kladny, J., Gorski, B., Cybulski, C., Gronwald, J., Kurzawski, G. & Lubinski, J. 2005b. CDKN2A common variants and their association with melanoma risk: a population-based study. *Cancer Res,* Vol. 65, No. 3**,** 835-839.

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., Mckenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* Vol. 43, No. 5**,** 491-498.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M. & Béroud, C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research,* Vol. 37, No. 9**,** e67-e67.

Desrichard, A., Bidet, Y., Uhrhammer, N. & Bignon, Y.-J. 2011. CHEK2 contribution to hereditary breast cancer in non-BRCAfamilies. *Breast Cancer Research,* Vol. 13, No. 6**,** R119.

Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., Merker, J. D., Goldfeder, R. L., Enns, G. M., David, S. P., Pakdaman, N., Ormond, K. E., Caleshu, C., Kingham, K., Klein, T. E., Whirl-Carrillo, M., Sakamoto, K., Wheeler, M. T., Butte, A. J., Ford, J. M., Boxer, L., Ioannidis, J. P., Yeung, A.

C., Altman, R. B., Assimes, T. L., Snyder, M., Ashley, E. A. & Quertermous, T. 2014. Clinical interpretation and implications of whole-genome sequencing. *Jama,* Vol. 311, No. 10**,** 1035-1045.

Ding, X., Zhu, L., Ji, T., Zhang, X., Wang, F., Gan, S., Zhao, M. & Yang, H. 2014. Long intergenic non-coding RNAs (LincRNAs) identified by RNA-seq in breast cancer. *PLoS One,* Vol. 9, No. 8**,** e103270.

Diogo, D., Kurreeman, F., Stahl, Eli a., Liao, Katherine p., Gupta, N., Greenberg, Jeffrey d., Rivas, Manuel a., Hickey, B., Flannick, J., Thomson, B., Guiducci, C., Ripke, S., Adzhubey, I., Barton, A., Kremer, Joel m., Alfredsson, L., Sunyaev, S., Martin, J., Zhernakova, A., Bowes, J., Eyre, S., Siminovitch, Katherine a., Gregersen, Peter k., Worthington, J., Klareskog, L., Padyukov, L., Raychaudhuri, S. & Plenge, Robert m. 2013. Rare, Low-Frequency, and Common Variants in the Protein-Coding Sequence of Biological Candidate Genes from GWASs Contribute to Risk of Rheumatoid Arthritis. *The American Journal of Human Genetics,* Vol. 92, No. 1**,** 15-27.

Dong, Y., Hakimi, M. A., Chen, X., Kumaraswamy, E., Cooch, N. S., Godwin, A. K. & Shiekhattar, R. 2003. Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in DNA repair. *Mol Cell,* Vol. 12, No. 5**,** 1087-1099.

Dork, T., Bendix, R., Bremer, M., Rades, D., Klopper, K., Nicke, M., Skawran, B., Hector, A., Yamini, P., Steinmann, D., Weise, S., Stuhrmann, M. & Karstens, J. H. 2001. Spectrum of ATM gene mutations in a hospital-based series of unselected breast cancer patients. *Cancer Res,* Vol. 61, No. 20**,** 7608-7615.

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,* Vol. 26, No. 5**,** 1179-1184.

Doss, C. G. P., Chakraborty, C., Chen, L. & Zhu, H. 2014. Integrating in silico prediction methods, molecular docking, and molecular dynamics simulation to predict the impact of ALK missense mutations in structural perspective. *BioMed research international,* Vol. 2014, No. 1**,** e895831.

Doudna, J. A. & Charpentier, E. 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science,* Vol. 346, No. 6213**,** e1258096.

Duzkale, H., Shen, J., Mclaughlin, H., Alfares, A., Kelly, M. A., Pugh, T. J., Funke, B. H., Rehm, H. L. & Lebo, M. S. 2013. A systematic approach to assessing the clinical significance of genetic variants. *Clinical genetics,* Vol. 84, No. 5**,** 453-463.

Easton, D. F., Bishop, D. T., Ford, D. & Crockford, G. P. 1993. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet,* Vol. 52, No. 4**,** 678-701.

Easton, D. F., Deffenbaugh, A. M., Pruss, D., Frye, C., Wenstrup, R. J., Allen-Brady, K., Tavtigian, S. V., Monteiro, A. N., Iversen, E. S., Couch, F. J. & Goldgar, D. E. 2007. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet,* Vol. 81, No. 5**,** 873-883.

Easton, D. F., Pharoah, P. D. P., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., Devilee, P., Meindl, A., Couch, F. J., Southey, M., Goldgar, D. E., Evans, D. G. R., Chenevix-Trench, G., Rahman, N., Robson, M., Domchek, S. M. & Foulkes, W. D. 2015. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *New England Journal of Medicine,* Vol. 372, No. 23**,** 2243-2257.

Eid, A., Alshareef, S. & Mahfouz, M. M. 2018. CRISPR base editors: genome editing without double-stranded breaks. *The Biochemical journal,* Vol. 475, No. 11**,** 1955-1964.

El-Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W. & Vogelstein, B. 1993. WAF1, a potential mediator of p53 tumor suppression. *Cell,* Vol. 75, No. 4**,** 817-825.

Ellingford, J. M., Horn, B., Campbell, C., Arno, G., Barton, S., Tate, C., Bhaskar, S., Sergouniotis, P. I., Taylor, R. L., Carss, K. J., Raymond, L. F. L., Michaelides, M., Ramsden, S. C., Webster, A. R. & Black, G. C. M. 2018. Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases. *Journal of medical genetics,* Vol. 55, No. 2**,** 114-121.

Eng, C. 1998. Genetics of Cowden syndrome: through the looking glass of oncology. *Int J Oncol,* Vol. 12, No. 3**,** 701-710.

Eng, C. 2003. PTEN: one gene, many syndromes. *Hum Mutat,* Vol. 22, No. 3**,** 183-198.

Erkko, H., Xia, B., Nikkila, J., Schleutker, J., Syrjakoski, K., Mannermaa, A., Kallioniemi, A., Pylkas, K., Karppinen, S. M., Rapakko, K., Miron, A., Sheng, Q., Li, G., Mattila, H., Bell, D. W., Haber, D. A., Grip, M., Reiman, M., Jukkola-Vuorinen, A., Mustonen, A., Kere, J., Aaltonen, L. A., Kosma, V. M., Kataja, V., Soini, Y.,

Drapkin, R. I., Livingston, D. M. & Winqvist, R. 2007. A recurrent mutation in PALB2 in Finnish cancer families. *Nature,* Vol. 446, No. 7133**,** 316-319.

Esquenet, M., Swinnen, J. V., Heyns, W. & Verhoeven, G. 1997. LNCaP prostatic adenocarcinoma cells derived from low and high passage numbers display divergent responses not only to androgens but also to retinoids. *J Steroid Biochem Mol Biol,* Vol. 62, No. 5-6**,** 391-399.

Evans, D. G., Eccles, D. M., Rahman, N., Young, K., Bulman, M., Amir, E., Shenton, A., Howell, A. & Lalloo, F. 2004. A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. *J Med Genet,* Vol. 41, No. 6**,** 474-480.

Evans, D. G., Harkness, E. F., Plaskocinska, I., Wallace, A. J., Clancy, T., Woodward, E. R., Howell, T. A., Tischkowitz, M. & Lalloo, F. 2017. Pathology update to the Manchester Scoring System based on testing in over 4000 families. *J Med Genet,* Vol. 54, No. 10**,** 674-681.

Evans, D. G., Lalloo, F., Cramer, A., Jones, E. A., Knox, F., Amir, E. & Howell, A. 2009. Addition of pathology and biomarker information significantly improves the performance of the Manchester scoring system for BRCA1 and BRCA2 testing. *J Med Genet,* Vol. 46, No. 12**,** 811-817.

Evans, D. G. R., Lalloo, F., Wallace, A. & Rahman, N. 2005. Update on the Manchester Scoring System for BRCA1 and BRCA2 testing. *Journal of Medical Genetics,* Vol. 42, No. 7**,** e39.

Evers, B. & Jonkers, J. 2006. Mouse models of BRCA1 and BRCA2 deficiency: past lessons, current understanding and future prospects. *Oncogene,* Vol. 25, No. 43**,** 5885-5897.

Ewen, M. E., Xing, Y. G., Lawrence, J. B. & Livingston, D. M. 1991. Molecular cloning, chromosomal mapping, and expression of the cDNA for p107, a retinoblastoma gene product-related protein. *Cell,* Vol. 66, No. 6**,** 1155-1164.

Fackenthal, J. D. & Olopade, O. I. 2007. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat Rev Cancer,* Vol. 7, No. 12**,** 937-948.

Fang, N. Y., Greiner, T. C., Weisenburger, D. D., Chan, W. C., Vose, J. M., Smith, L. M., Armitage, J. O., Mayer, R. A., Pike, B. L., Collins, F. S. & Hacia, J. G. 2003. Oligonucleotide microarrays demonstrate the highest frequency of ATM mutations in the mantle cell subtype of lymphoma. *Proc Natl Acad Sci U S A,* Vol. 100, No. 9**,** 5372-5377.

Farra, C., Dagher, C., Badra, R., Hammoud, M. S., Alameddine, R., Awwad, J., Seoud, M., Abbas, J., Boulos, F., El Saghir, N. & Mukherji, D. 2019. BRCA mutation screening and patterns among high-risk Lebanese subjects. *Hereditary cancer in clinical practice,* Vol. 17, No**.,** 4-4.

Ferguson, A. T., Evron, E., Umbricht, C. B., Pandita, T. K., Chan, T. A., Hermeking, H., Marks, J. R., Lambers, A. R., Futreal, P. A., Stampfer, M. R. & Sukumar, S. 2000. High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer. *Proc Natl Acad Sci U S A,* Vol. 97, No. 11**,** 6049-6054.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. 2011. Bioinformatics challenges for personalized medicine. *Bioinformatics,* Vol. 27, No. 13**,** 1741-1748.

Fletcher, O., Johnson, N., Dos Santos Silva, I., Orr, N., Ashworth, A., Nevanlinna, H., Heikkinen, T., Aittomaki, K., Blomqvist, C., Burwinkel, B., Bartram, C. R., Meindl, A., Schmutzler, R. K., Cox, A., Brock, I., Elliott, G., Reed, M. W., Southey, M. C., Smith, L., Spurdle, A. B., Hopper, J. L., Couch, F. J., Olson, J. E., Wang, X., Fredericksen, Z., Schurmann, P., Waltes, R., Bremer, M., Dork, T., Devilee, P., Van Asperen, C. J., Tollenaar, R. A., Seynaeve, C., Hall, P., Czene, K., Humphreys, K., Liu, J., Ahmed, S., Dunning, A. M., Maranian, M., Pharoah, P. D., Chenevix-Trench, G., Beesley, J., Bogdanova, N. V., Antonenkova, N. N., Zalutsky, I. V., Anton-Culver, H., Ziogas, A., Brauch, H., Ko, Y. D., Hamann, U., Fasching, P. A., Strick, R., Ekici, A. B., Beckmann, M. W., Giles, G. G., Severi, G., Baglietto, L., English, D. R., Milne, R. L., Benitez, J., Arias, J. I., Pita, G., Nordestgaard, B. G., Bojesen, S. E., Flyger, H., Kang, D., Yoo, K. Y., Noh, D. Y., Mannermaa, A., Kataja, V., Kosma, V. M., Garcia-Closas, M., Chanock, S., Lissowska, J., Brinton, L. A., Chang-Claude, J., Wang-Gohrke, S., Broeks, A., Schmidt, M. K., Van Leeuwen, F. E., Van't Veer, L. J., Margolin, S., Lindblom, A., Humphreys, M. K., Morrison, J., Platte, R., Easton, D. F. & Peto, J. 2010. Missense variants in ATM in 26,101 breast cancer cases and 29,842 controls. *Cancer Epidemiol Biomarkers Prev,* Vol. 19, No. 9**,** 2143-2151.

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A. & Stratton, M. R. 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics,* Vol. 10, No. 11**,** Unit-10.11.

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D. T., Weber, B., Lenoir, G., Chang-Claude, J., Sobol, H., Teare, M. D., Struewing, J., Arason, A., Scherneck, S., Peto, J., Rebbeck, T. R., Tonin, P., Neuhausen, S., Barkardottir, R., Eyfjord, J., Lynch, H., Ponder, B. A., Gayther, S. A., Zelada-Hedman, M. & Et Al. 1998. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet,* Vol. 62, No. 3**,** 676-689.

Fortuno, C., James, P. A., Young, E. L., Feng, B., Olivier, M., Pesaran, T., Tavtigian, S. V. & Spurdle, A. B. 2018. Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants. *Human Mutation,* Vol. 39, No. 8**,** 1061-1069.

Friedenson, B. 2007. The BRCA1/2 pathway prevents hematologic cancers in additon to breast and ovairan cancers. *BioMed Central Cancer,* Vol. 7, No. 152.

Fu, Y., Foden, J. A., Khayter, C., Maeder, M. L., Reyon, D., Joung, J. K. & Sander, J. D. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature biotechnology,* Vol. 31, No. 9**,** 822-826.

Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol,* Vol. 32, No. 3**,** 279-284.

Fujita, S., Masago, K., Okuda, C., Hata, A., Kaji, R., Katakami, N. & Hirata, Y. 2017. Single nucleotide variant sequencing errors in whole exome sequencing using the Ion Proton System. *Biomedical Reports,* Vol. 7, No. 1**,** 17-20.

Fuller, M. S., Lee, C. I. & Elmore, J. G. 2015. Breast cancer screening: an evidence-based update. *The Medical clinics of North America,* Vol. 99, No. 3**,** 451-468.

Gadd, S., Huff, V., Walz, A. L., Ooms, A., Armstrong, A. E., Gerhard, D. S., Smith, M. A., Auvil, J. M. G., Meerzaman, D., Chen, Q. R., Hsu, C. H., Yan, C., Nguyen, C., Hu, Y., Hermida, L. C., Davidsen, T., Gesuwan, P., Ma, Y., Zong, Z., Mungall, A. J., Moore, R. A., Marra, M. A., Dome, J. S., Mullighan, C. G., Ma, J., Wheeler, D. A., Hampton, O. A., Ross, N., Gastier-Foster, J. M., Arold, S. T. & Perlman, E. J. 2017. A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat Genet,* Vol. 49, No. 10**,** 1487-1494.

Garber, J. E., Goldstein, A. M., Kantor, A. F., Dreyfus, M. G., Fraumeni, J. F., Jr. & Li, F. P. 1991. Follow-up study of twenty-four families with Li-Fraumeni syndrome. *Cancer Res,* Vol. 51, No. 22**,** 6094-6097.

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. & Moineau, S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature,* Vol. 468, No.**,** 67.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. & Liu, D. R. 2017. Programmable base editing of A toT to G to C in genomic DNA without DNA cleavage. *Nature,* Vol. 551, No. 7681**,** 464-471.

Gayther, S. A., Batley, S. J., Linger, L., Bannister, A., Thorpe, K., Chin, S. F., Daigo, Y., Russell, P., Wilson, A., Sowter, H. M., Delhanty, J. D., Ponder, B. A., Kouzarides, T. & Caldas, C. 2000. Mutations truncating the EP300 acetylase in human cancers. *Nat Genet,* Vol. 24, No. 3**,** 300-303.

Gayther, S. A., Mangion, J., Russell, P., Seal, S., Barfoot, R., Ponder, B. A., Stratton, M. R. & Easton, D. 1997. Variation of risks of breast and ovarian cancer associated with different germline mutations of the BRCA2 gene. *Nat Genet,* Vol. 15, No. 1**,** 103-105.

George, J., Lim, J. S., Jang, S. J., Cun, Y., Ozretic, L., Kong, G., Leenders, F., Lu, X., Fernandez-Cuesta, L., Bosco, G., Muller, C., Dahmen, I., Jahchan, N. S., Park, K. S., Yang, D., Karnezis, A. N., Vaka, D., Torres, A., Wang, M. S., Korbel, J. O., Menon, R., Chun, S. M., Kim, D., Wilkerson, M., Hayes, N., Engelmann, D., Putzer, B., Bos, M., Michels, S., Vlasic, I., Seidel, D., Pinther, B., Schaub, P., Becker, C., Altmuller, J., Yokota, J., Kohno, T., Iwakawa, R., Tsuta, K., Noguchi, M., Muley, T., Hoffmann, H., Schnabel, P. A., Petersen, I., Chen, Y., Soltermann, A., Tischler, V., Choi, C. M., Kim, Y. H., Massion, P. P., Zou, Y., Jovanovic, D., Kontic, M., Wright, G. M., Russell, P. A., Solomon, B., Koch, I., Lindner, M., Muscarella, L. A., La Torre, A., Field, J. K., Jakopovic, M., Knezevic, J., Castanos-Velez, E., Roz, L., Pastorino, U., Brustugun, O. T., Lund-Iversen, M., Thunnissen, E., Kohler, J., Schuler, M., Botling, J., Sandelin, M., Sanchez-Cespedes, M., Salvesen, H. B., Achter, V., Lang, U., Bogus, M., Schneider, P. M., Zander, T., Ansen, S., Hallek, M., Wolf, J., Vingron, M., Yatabe, Y., Travis, W. D., Nurnberg, P., Reinhardt, C., Perner, S., Heukamp, L., Buttner, R., Haas, S. A., Brambilla, E., Peifer, M., Sage, J. & Thomas, R. K. 2015. Comprehensive genomic profiles of small cell lung cancer. *Nature,* Vol. 524, No. 7563**,** 47-53.

Germani, A., Libi, F., Maggi, S., Stanzani, G., Lombardi, A., Pellegrini, P., Mattei, M., De Marchis, L., Amanti, C., Pizzuti, A., Torrisi, M. R. & Piane, M. 2018. Rapid detection of copy number variations and point mutations in BRCA1/2 genes using a single workflow by ion semiconductor sequencing pipeline. *Oncotarget,* Vol. 9, No. 72**,** 33648-33655.

Ghiasi, N., Habibagahi, M., Rosli, R., Ghaderi, A., Yusoff, K., Hosseini, A., Abdullah, S. & Jaberipour, M. 2014. Tumour suppressive effects of WEE1 gene silencing in breast cancer cells. *Asian Pac J Cancer Prev,* Vol. 14, No. 11**,** 6605-6611.

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., Dennis, J., Wang, Q., Humphreys, M. K., Luccarini, C., Baynes, C., Conroy, D., Maranian, M., Ahmed, S., Driver, K., Johnson, N., Orr, N., Dos Santos Silva, I., Waisfisz, Q., Meijers-Heijboer, H., Uitterlinden, A. G., Rivadeneira, F., Hall, P., Czene, K., Irwanto, A., Liu, J., Nevanlinna, H., Aittomaki, K., Blomqvist, C., Meindl, A., Schmutzler, R. K., Muller-Myhsok, B., Lichtner, P., Chang-Claude, J., Hein, R., Nickels, S., Flesch-Janys, D., Tsimiklis, H., Makalic, E., Schmidt, D., Bui, M., Hopper, J. L., Apicella, C., Park, D. J., Southey, M., Hunter, D. J., Chanock, S. J., Broeks, A., Verhoef, S., Hogervorst, F. B., Fasching, P. A., Lux, M. P., Beckmann, M. W., Ekici, A. B., Sawyer, E., Tomlinson, I., Kerin, M., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Guenel, P., Truong, T., Cordina-Duverger, E., Menegaux, F., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Milne, R. L., Alonso, M. R., Gonzalez-Neira, A., Benitez, J., Anton-Culver, H., Ziogas, A., Bernstein, L., Dur, C. C., Brenner, H., Muller, H., Arndt, V., Stegmaier, C., Justenhoven, C., Brauch, H., Bruning, T., Wang-Gohrke, S., Eilber, U., Dork, T., Schurmann, P., Bremer, M., Hillemanns, P., Bogdanova, N. V., Antonenkova, N. N., Rogov, Y. I., Karstens, J. H., Bermisheva, M., Prokofieva, D., Khusnutdinova, E., Lindblom, A., Margolin, S., Mannermaa, A*., et al.* 2012. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet,* Vol. 44, No. 3**,** 312-318.

Giannakis, M., Hodis, E., Jasmine Mu, X., Yamauchi, M., Rosenbluh, J., Cibulskis, K., Saksena, G., Lawrence, M. S., Qian, Z. R., Nishihara, R., Van Allen, E. M., Hahn, W. C., Gabriel, S. B., Lander, E. S., Getz, G., Ogino, S., Fuchs, C. S. & Garraway, L. A. 2014. RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat Genet,* Vol. 46, No. 12**,** 1264-1266.

Giannakis, M., Mu, X. J., Shukla, S. A., Qian, Z. R., Cohen, O., Nishihara, R., Bahl, S., Cao, Y., Amin-Mansour, A., Yamauchi, M., Sukawa, Y., Stewart, C., Rosenberg, M., Mima, K., Inamura, K., Nosho, K., Nowak, J. A., Lawrence, M. S., Giovannucci, E. L., Chan, A. T., Ng, K., Meyerhardt, J. A., Van Allen, E. M., Getz, G., Gabriel, S. B., Lander, E. S., Wu, C. J., Fuchs, C. S., Ogino, S. & Garraway, L. A. 2016. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep,* Vol. 15, No. 4**,** 857-865.

Gilad, S., Khosravi, R., Shkedy, D., Uziel, T., Ziv, Y., Savitsky, K., Rotman, G., Smith, S., Chessa, L., Jorgensen, T. J., Harnik, R., Frydman, M., Sanal, O., Portnoi, S., Goldwicz, Z., Jaspers, N. G., Gatti, R. A., Lenoir, G., Lavin, M. F., Tatsumi, K., Wegner, R. D., Shiloh, Y. & Bar-Shira, A. 1996. Predominance of null mutations in ataxia-telangiectasia. *Hum Mol Genet,* Vol. 5, No. 4**,** 433-439.

Ginsberg, D., Vairo, G., Chittenden, T., Xiao, Z. X., Xu, G., Wydner, K. L., Decaprio, J. A., Lawrence, J. B. & Livingston, D. M. 1994. E2F-4, a new member of the E2F transcription factor family, interacts with p107. *Genes Dev,* Vol. 8, No. 22**,** 2665-2679.

Goldgar, D. E., Easton, D. F., Deffenbaugh, A. M., Monteiro, A. N. A., Tavtigian, S. V. & Couch, F. J. 2004. Integrated Evaluation of DNA Sequence Variants of Unknown Clinical Significance: Application to BRCA1 and BRCA2. *The American Journal of Human Genetics,* Vol. 75, No. 4**,** 535-544.

Gorski, B., Debniak, T., Masojc, B., Mierzejewski, M., Medrek, K., Cybulski, C., Jakubowska, A., Kurzawski, G., Chosia, M., Scott, R. & Lubinski, J. 2003. Germline 657del5 mutation in the NBS1 gene in breast cancer patients. *Int J Cancer,* Vol. 106, No. 3**,** 379-381.

Gorter de vries, A. R., Pronk, J. T., Ter horst, J., Couwenberg, L. G. F., Van den broek, M., De la torre cortés, P. & Daran, J.-M. G. 2018. Allele-specific genome editing using CRISPR–Cas9 is associated with loss of heterozygosity in diploid yeast. *Nucleic Acids Research,* Vol. 47, No. 3**,** 1362-1372.

Gowen, L. C., Johnson, B. L., Latour, A. M., Sulik, K. K. & Koller, B. H. 1996. Brca1 deficiency results in early embryonic lethality characterized by neuroepithelial abnormalities. *Nat Genet,* Vol. 12, No. 2**,** 191-194.

Grossman, S. R., Deato, M. E., Brignone, C., Chan, H. M., Kung, A. L., Tagami, H., Nakatani, Y. & Livingston, D. M. 2003. Polyubiquitination of p53 by a ubiquitin ligase activity of p300. *Science,* Vol. 300, No. 5617**,** 342-344.

Guilford, P., Hopkins, J., Harraway, J., Mcleod, M., Mcleod, N., Harawira, P., Taite, H., Scoular, R., Miller, A. & Reeve, A. E. 1998. E-cadherin germline mutations in familial gastric cancer. *Nature,* Vol. 392, No. 6674**,** 402-405.

Gumy-Pause, F., Wacker, P., Maillet, P., Betts, D. R. & Sappino, A. P. 2006. ATM alterations in childhood non-Hodgkin lymphoma. *Cancer Genet Cytogenet,* Vol. 166, No. 2**,** 101-111.

Gundry, M. C., Brunetti, L., Lin, A., Mayle, A. E., Kitano, A., Wagner, D., Hsu, J. I., Hoegenauer, K. A., Rooney, C. M., Goodell, M. A. & Nakada, D. 2016. Highly Efficient Genome Editing of Murine and Human Hematopoietic Progenitor Cells by CRISPR/Cas9. *Cell Rep,* Vol. 17, No. 5**,** 1453-1461.

Gutierrez-Enriquez, S., Bonache, S., De Garibay, G. R., Osorio, A., Santamarina, M., Ramon Y Cajal, T., Esteban-Cardenosa, E., Tenes, A., Yanowsky, K., Barroso, A., Montalban, G., Blanco, A., Cornet, M., Gadea, N., Infante, M., Caldes, T., Diaz-Rubio, E., Balmana, J., Lasa, A., Vega, A., Benitez, J., De La Hoya, M. & Diez, O. 2014. About 1% of the breast and ovarian Spanish families testing negative for BRCA1 and BRCA2 are carriers of RAD51D pathogenic variants. *Int J Cancer,* Vol. 134, No. 9**,** 2088-2097.

Gutierrez-Enriquez, S., Fernet, M., Dork, T., Bremer, M., Lauge, A., Stoppa-Lyonnet, D., Moullan, N., Angele, S. & Hall, J. 2004. Functional consequences of ATM sequence variants for chromosomal radiosensitivity. *Genes Chromosomes Cancer,* Vol. 40, No. 2**,** 109-119.

Hakem, R., De La Pompa, J. L., Sirard, C., Mo, R., Woo, M., Hakem, A., Wakeham, A., Potter, J., Reitmair, A., Billia, F., Firpo, E., Hui, C. C., Roberts, J., Rossant, J. & Mak, T. W. 1996. The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. *Cell,* Vol. 85, No. 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,* Vol. 250, No. 4988**,** 1684-1689.

Hamblin, A., Wordsworth, S., Fermont, J. M., Page, S., Kaur, K., Camps, C., Kaisaki, P., Gupta, A., Talbot, D., Middleton, M., Henderson, S., Cutts, A., Vavoulis, D. V., Housby, N., Tomlinson, I., Taylor, J. C. & Schuh, A. 2017. Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: Retrospective validation and prospective audit in the UK National Health Service. *PLOS Medicine,* Vol. 14, No. 2**,** e1002230.

Hanahan, D. & Weinberg, Robert a. 2011. Hallmarks of Cancer: The Next Generation. *Cell,* Vol. 144, No. 5**,** 646-674.

Hao, J. J., Lin, D. C., Dinh, H. Q., Mayakonda, A., Jiang, Y. Y., Chang, C., Jiang, Y., Lu, C. C., Shi, Z. Z., Xu, X., Zhang, Y., Cai, Y., Wang, J. W., Zhan, Q. M., Wei, W. Q., Berman, B. P., Wang, M. R. & Koeffler, H. P. 2016. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet,* Vol. 48, No. 12**,** 1500-1507.

Harakalova, M., Nijman, I. J., Medic, J., Mokry, M., Renkens, I., Blankensteijn, J. D., Kloosterman, W., Baas, A. F. & Cuppen, E. 2011. Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm regions of interest-challenges and limitations. *Journal of cardiovascular translational research,* Vol. 4, No. 3**,** 271-280.

Hartley, K. O., Gell, D., Smith, G. C., Zhang, H., Divecha, N., Connelly, M. A., Admon, A., Lees-Miller, S. P., Anderson, C. W. & Jackson, S. P. 1995. DNA-dependent protein kinase catalytic subunit: a relative of phosphatidylinositol 3-kinase and the ataxia telangiectasia gene product. *Cell,* Vol. 82, No. 5**,** 849-856.

Hatchi, E., Skourti-Stathaki, K., Ventz, S., Pinello, L., Yen, A., Kamieniarz-Gdula, K., Dimitrov, S., Pathania, S., Mckinney, K. M., Eaton, M. L., Kellis, M., Hill, S. J., Parmigiani, G., Proudfoot, N. J. & Livingston, D. M. 2015. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Mol Cell,* Vol. 57, No. 4**,** 636-647.

Hay, R. T. 2005. SUMO: a history of modification. *Mol Cell,* Vol. 18, No. 1**,** 1-12.

Heald, R., Mcloughlin, M. & Mckeon, F. 1993. Human wee1 maintains mitotic timing by protecting the nucleus from cytoplasmically activated Cdc2 kinase. *Cell,* Vol. 74, No. 3**,** 463-474.

Heikkinen, K., Rapakko, K., Karppinen, S. M., Erkko, H., Nieminen, P. & Winqvist, R. 2005. Association of common ATM polymorphism with bilateral breast cancer. *Int J Cancer,* Vol. 116, No. 1**,** 69-72.

Hert, D. G., Fredlake, C. P. & Barron, A. E. 2008. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis,* Vol. 29, No. 23**,** 4618-4626.

Hess, G. T., Tycko, J., Yao, D. & Bassik, M. C. 2017. Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol Cell,* Vol. 68, No. 1**,** 26-43.

Hickson, I. D. 2003. RecQ helicases: caretakers of the genome. *Nat Rev Cancer,* Vol. 3, No. 3**,** 169-178.

Higasa, K., Miyake, N., Yoshimura, J., Okamura, K., Niihori, T., Saitsu, H., Doi, K., Shimizu, M., Nakabayashi, K., Aoki, Y., Tsurusaki, Y., Morishita, S., Kawaguchi, T., Migita, O., Nakayama, K., Nakashima, M., Mitsui, J., Narahara, M., Hayashi, K., Funayama, R., Yamaguchi, D., Ishiura, H., Ko, W. Y., Hata, K., Nagashima, T., Yamada, R., Matsubara, Y., Umezawa, A., Tsuji, S., Matsumoto, N. & Matsuda, F. 2016. Human genetic variation database, a reference database of genetic variations in the Japanese population. *J Hum Genet,* Vol. 61, No. 6**,** 547-553.

Hilbers, F. S., Wijnen, J. T., Hoogerbrugge, N., Oosterwijk, J. C., Collee, M. J., Peterlongo, P., Radice, P., Manoukian, S., Feroce, I., Capra, F., Couch, F. J., Wang, X., Guidugli, L., Offit, K., Shah, S., Campbell, I. G., Thompson, E. R., James, P. A., Trainer, A. H., Gracia, J., Benitez, J., Van Asperen, C. J. & Devilee, P. 2012. Rare variants in XRCC2 as breast cancer susceptibility alleles. *J Med Genet,* Vol. 49, No. 10**,** 618-620.

Hill, S. J., Rolland, T., Adelmant, G., Xia, X., Owen, M. S., Dricot, A., Zack, T. I., Sahni, N., Jacob, Y., Hao, T., Mckinney, K. M., Clark, A. P., Reyon, D., Tsai, S. Q., Joung, J. K., Beroukhim, R., Marto, J. A., Vidal, M., Gaudet, S., Hill, D. E. & Livingston, D. M. 2014. Systematic screening reveals a role for BRCA1 in the response to transcription-associated DNA damage. *Genes Dev,* Vol. 28, No. 17**,** 1957-1975.

Hinz, J. M., Tebbs, R. S., Wilson, P. F., Nham, P. B., Salazar, E. P., Nagasawa, H., Urbin, S. S., Bedford, J. S. & Thompson, L. H. 2006. Repression of mutagenesis by Rad51D-mediated homologous recombination. *Nucleic Acids Research,* Vol. 34, No. 5**,** 1358-1368.

Hiraguri, S., Godfrey, T., Nakamura, H., Graff, J., Collins, C., Shayesteh, L., Doggett, N., Johnson, K., Wheelock, M., Herman, J., Baylin, S., Pinkel, D. & Gray, J. 1998. Mechanisms of inactivation of E-cadherin in breast cancer cell lines. *Cancer Res,* Vol. 58, No. 9**,** 1972-1977.

Hirsch, P., Zhang, Y., Tang, R., Joulin, V., Boutroux, H., Pronier, E., Moatti, H., Flandrin, P., Marzac, C., Bories, D., Fava, F., Mokrani, H., Betems, A., Lorre, F., Favier, R., Féger, F., Mohty, M., Douay, L., Legrand, O., Bilhou-Nabera, C., Louache, F. & Delhommeau, F. 2016. Genetic hierarchy and temporal variegation in the clonal history of acute myeloid leukaemia. *Nature Communications,* Vol. 7, No.**,** 12475.

Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. 1991. p53 mutations in human cancers. *Science,* Vol. 253, No. 5015**,** 49-53.

Horn, S., Barnard, S. & Rothkamm, K. 2011. Gamma-H2AX-based dose estimation for whole and partial body radiation exposure. *PloS one,* Vol. 6, No. 9**,** e25113-e25113.

Houvras, Y., Benezra, M., Zhang, H., Manfredi, J. J., Weber, B. L. & Licht, J. D. 2000. BRCA1 physically and functionally interacts with ATF1. *J Biol Chem,* Vol. 275, No. 46**,** 36230-36237.

Hsia, T. C., Chang, W. S., Chen, W. C., Liang, S. J., Tu, C. Y., Chen, H. J., Liang, J. A., Tsai, C. W., Hsu, C. M., Tsai, C. H. & Bau, D. T. 2014. Genotype of DNA double-strand break repair gene XRCC7 is associated with lung cancer risk in Taiwan males and smokers. *Anticancer Res,* Vol. 34, No. 12**,** 7001-7005.

Hsu, Patrick d., Lander, Eric s. & Zhang, F. 2014. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell,* Vol. 157, No. 6**,** 1262-1278.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G. & Zhang, F. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol,* Vol. 31, No. 9**,** 827-832.

Hu, J. & Ng, P. C. 2013. SIFT Indel: Predictions for the Functional Effects of Amino Acid Insertions/Deletions in Proteins. *PLOS ONE,* Vol. 8, No. 10**,** e77940.

Hu, Y., Scully, R., Sobhian, B., Xie, A., Shestakova, E. & Livingston, D. M. 2011. RAP80-directed tuning of BRCA1 homologous recombination function at ionizing radiation-induced nuclear foci. *Genes & development,* Vol. 25, No. 7**,** 685-700.

Huang, Y., Wang, W., Chen, Y., Huang, Y., Zhang, J., He, S., Tan, Y., Qiang, F., Li, A., Roe, O. D., Wang, S., Zhou, Y. & Zhou, J. 2014. The opposite prognostic significance of nuclear and cytoplasmic p21 expression in resectable gastric cancer patients. *J Gastroenterol,* Vol. 49, No. 11**,** 1441-1452.

Huppert, J. L. & Balasubramanian, S. 2006. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research,* Vol. 35, No. 2**,** 406-413.

International Agency for Research on Cancer 2014. *World Cancer Report 2014*, World Health Organisation.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol,* Vol. 169, No. 12**,** 5429-5433.

Ivashkevich, A., Redon, C. E., Nakamura, A. J., Martin, R. F. & Martin, O. A. 2012. Use of the γ-H2AX assay to monitor DNA damage and repair in translational cancer research. *Cancer Letters,* Vol. 327, No. 1**,** 123-133.

Jacobi, A. M., Rettig, G. R., Turk, R., Collingwood, M. A., Zeiner, S. A., Quadros, R. M., Harms, D. W., Bonthuis, P. J., Gregg, C., Ohtsuka, M., Gurumurthy, C. B. & Behlke, M. A. 2017. Simplified CRISPR tools for efficient genome editing and streamlined protocols for their delivery into mammalian cells and mouse zygotes. *Methods,* Vol. 121-122, No. 1**,** 16-28.

Jalilvand, M., Oloomi, M., Najafipour, R., Alizadeh, S. A., Saki, N., Rad, F. S. & Shekari, M. 2017. An association study between CHEK2 gene mutations and susceptibility to breast cancer. *Comparative Clinical Pathology,* Vol. 26, No. 4**,** 837-845.

Jensen, R. B., Carreira, A. & Kowalczykowski, S. C. 2010. Purified human BRCA2 stimulates RAD51-mediated recombination. *Nature,* Vol. 467, No. 7316**,** 678-683.

Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol,* Vol. 31, No. 3**,** 233-239.

Jin, G., Mao, X., Qiao, Z., Chen, B. & Jin, F. 2019. RAP80 expression in breast cancer and its relationship with apoptosis in breast cancer cells. *OncoTargets and therapy,* Vol. 12, No.**,** 625-634.

Jin, S. C., Benitez, B. A., Deming, Y. & Cruchaga, C. 2016. Pooled-DNA Sequencing for Elucidating New Genomic Risk Factors, Rare Variants Underlying Alzheimer's Disease. *Methods Mol Biol,* Vol. 1303, No. 1**,** 299-314.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science,* Vol. 337, No. 6096**,** 816-821.

Jo, V. Y., Sholl, L. M. & Krane, J. F. 2016. Distinctive Patterns of CTNNB1 (beta-Catenin) Alterations in Salivary Gland Basal Cell Adenoma and Basal Cell Adenocarcinoma. *Am J Surg Pathol,* Vol. 40, No. 8**,** 1143-1150.

Johnson, R. D., Liu, N. & Jasin, M. 1999. Mammalian XRCC2 promotes the repair of DNA double-strand breaks by homologous recombination. *Nature,* Vol. 401, No. 6751**,** 397-399.

Judkins, T., Leclair, B., Bowles, K., Gutin, N., Trost, J., Mcculloch, J., Bhatnagar, S., Murray, A., Craft, J., Wardell, B., Bastian, M., Mitchell, J., Chen, J., Tran, T., Williams, D., Potter, J., Jammulapati, S., Perry, M., Morris, B., Roa, B. & Timms, K. 2015. Development and analytical validation of a 25-gene next generation sequencing panel that includes the BRCA1 and BRCA2 genes to assess hereditary cancer risk. *BMC Cancer,* Vol. 15, No.**,** 215.

Kalmyrzaev, B., Pharoah, P. D. P., Easton, D. F., Ponder, B. a. J. & Dunning, A. M. 2008. Hyaluronan-Mediated Motility Receptor Gene Single Nucleotide Polymorphisms and Risk of Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention,* Vol. 17, No. 12**,** 3618-3620.

Kamb, A., Gruis, N. A., Weaver-Feldhaus, J., Liu, Q., Harshman, K., Tavtigian, S. V., Stockert, E., Day, R. S., 3rd, Johnson, B. E. & Skolnick, M. H. 1994. A cell cycle regulator potentially involved in genesis of many tumor types. *Science,* Vol. 264, No. 5157**,** 436-440.

Kanagal-Shamanna, R., Portier, B. P., Singh, R. R., Routbort, M. J., Aldape, K. D., Handal, B. A., Rahimi, H., Reddy, N. G., Barkoh, B. A., Mishra, B. M., Paladugu, A. V., Manekia, J. H., Kalhor, N., Chowdhuri, S. R., Staerkel, G. A., Medeiros, L. J., Luthra, R. & Patel, K. P. 2014. Next-generation sequencing-based multi-gene mutation profiling of solid tumors using fine needle aspiration samples: promises and challenges for routine clinical diagnostics. *Mod Pathol,* Vol. 27, No. 2**,** 314-327.

Kanchi, K. L., Johnson, K. J., Lu, C., Mclellan, M. D., Leiserson, M. D., Wendl, M. C., Zhang, Q., Koboldt, D. C., Xie, M., Kandoth, C., Mcmichael, J. F., Wyczalkowski, M. A., Larson, D. E., Schmidt, H. K., Miller, C. A., Fulton, R. S., Spellman, P. T., Mardis, E. R., Druley, T. E., Graubert, T. A., Goodfellow, P. J., Raphael, B. J., Wilson, R. K. & Ding, L. 2014. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun,* Vol. 5, No.**,** 3156.

Kang, X., He, W., Huang, Y., Yu, Q., Chen, Y., Gao, X., Sun, X. & Fan, Y. 2016. Introducing precise genetic modifications into human 3PN embryos by CRISPR/Cas-mediated genome editing. *Journal of Assisted Reproduction and Genetics,* Vol. 33, No. 5**,** 581-588.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J. X., Samocha, K. E., Pierce-Hoffman, E., Zappala,

Z., O'donnell-Luria, A. H., Minikel, E. V., Weisburd, B., Lek, M., Ware, J. S., Vittal, C., Armean, I. M., Bergelson, L., Cibulskis, K., Connolly, K. M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M. E., Neale, B. M., Daly, M. J. & Macarthur, D. G. 2019. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv,* Vol. No**.,** 531210.

Karppinen, S. M., Heikkinen, K., Rapakko, K. & Winqvist, R. 2004. Mutation screening of the BARD1 gene: evidence for involvement of the Cys557Ser allele in hereditary susceptibility to breast cancer. *J Med Genet,* Vol. 41, No. 9**,** e114.

Kass, E. M. & Jasin, M. 2010. Collaboration and competition between DNA double-strand break repair pathways. *FEBS Lett,* Vol. 584, No. 17**,** 3703-3708.

Kast, K., Schmutzler, R. K., Rhiem, K., Kiechle, M., Fischer, C., Niederacher, D., Arnold, N., Grimm, T., Speiser, D., Schlegelberger, B., Varga, D., Horvath, J., Beer, M., Briest, S., Meindl, A. & Engel, C. 2014. Validation of the Manchester scoring system for predicting BRCA1/2 mutations in 9,390 families suspected of having hereditary breast and ovarian cancer. *Int J Cancer,* Vol. 135, No. 10**,** 2352-2361.

Kenemans, P., Verstraeten, R. A. & Verheijen, R. H. 2008. Oncogenic pathways in hereditary and sporadic breast cancer. *Maturitas,* Vol. 61, No. 1-2**,** 141-150.

Kerangueven, F., Essioux, L., Dib, A., Noguchi, T., Allione, F., Geneix, J., Longy, M., Lidereau, R., Eisinger, F., Pebusque, M. J. & Et Al. 1995. Loss of heterozygosity and linkage analysis in breast carcinoma: indication for a putative third susceptibility gene on the short arm of chromosome 8. *Oncogene,* Vol. 10, No. 5**,** 1023-1026.

Kilpivaara, O., Vahteristo, P., Falck, J., Syrjakoski, K., Eerola, H., Easton, D., Bartkova, J., Lukas, J., Heikkila, P., Aittomaki, K., Holli, K., Blomqvist, C., Kallioniemi, O. P., Bartek, J. & Nevanlinna, H. 2004. CHEK2 variant I157T may be associated with increased breast cancer risk. *Int J Cancer,* Vol. 111, No. 4**,** 543-547.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H. R., Hwang, J., Kim, J. I. & Kim, J. S. 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods,* Vol. 12, No. 3**,** 237-243, 231 p following 243.

Kim, H., Chen, J. & Yu, X. 2007. Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. *Science,* Vol. 316, No. 5828**,** 1202-1205.

Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J.-S. 2014. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome research,* Vol. 24, No. 6**,** 1012-1019.

King, M. C., Marks, J. H. & Mandell, J. B. 2003. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science,* Vol. 302, No. 5645**,** 643-646.

Kingsmore, S. F. & Saunders, C. J. 2011. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci Transl Med,* Vol. 3, No. 87**,** 87ps23.

Kinzler, K. W. & Vogelstein, B. 1997. Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature,* Vol. 386, No. 6627**,** 761, 763.

Kittaneh, M., Montero, A. J. & Gluck, S. 2013. Molecular profiling for breast cancer: a comprehensive review. *Biomark Cancer,* Vol. 5, No**.,** 61-70.

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E. & Topper, S. E. 2017. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Medicine,* Vol. 9, No. 1**,** 13.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature,* Vol. 533, No. 7603**,** 420-424.

Kovaleva, V., Geissler, A.-L., Lutz, L., Fritsch, R., Makowiec, F., Wiesemann, S., Hopt, U. T., Passlick, B., Werner, M. & Lassmann, S. 2016. Spatio-temporal mutation profiles of case-matched colorectal carcinomas and their metastases reveal unique de novo mutations in metachronous lung metastases by targeted next generation sequencing. *Molecular Cancer,* Vol. 15, No**.,** 63.

Krepischi, A. C., Achatz, M. I., Santos, E. M., Costa, S. S., Lisboa, B. C., Brentani, H., Santos, T. M., Goncalves, A., Nobrega, A. F., Pearson, P. L., Vianna-Morgante, A. M., Carraro, D. M., Brentani, R. R. & Rosenberg, C. 2012. Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res,* Vol. 14, No. 1**,** R24.

Ku, C. S., Cooper, D. N. & Patrinos, G. P. 2016. The Rise and Rise of Exome Sequencing. *Public Health Genomics,* Vol. 19, No. 6**,** 315-324.

Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., Etzioni, R., Bolouri, H., Montgomery, B., White, T., Lucas, J. M., Brown, L. G., Dumpit, R. F., Desarkar, N., Higano, C., Yu, E. Y., Coleman, R., Schultz, N., Fang, M., Lange, P. H., Shendure, J., Vessella, R. L. & Nelson, P. S. 2016. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat Med,* Vol. 22, No. 4**,** 369-378.

Kumaran, M., Cass, C. E., Graham, K., Mackey, J. R., Hubaux, R., Lam, W., Yasui, Y. & Damaraju, S. 2017. Germline copy number variations are associated with breast cancer risk and prognosis. *Scientific Reports,* Vol. 7, No. 1**,** 14621.

Kuscu, C., Parlak, M., Tufan, T., Yang, J., Szlachta, K., Wei, X., Mammadov, R. & Adli, M. 2017. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nat Methods,* Vol. 14, No. 7**,** 710-712.

Kuusisto, K. M., Akinrinade, O., Vihinen, M., Kankuri-Tammilehto, M., Laasanen, S. L. & Schleutker, J. 2013. copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS One,* Vol. 8, No. 8**,** e71802.

Kwong, A., Shin, V. Y., Cheuk, I. W. Y., Chen, J., Au, C. H., Ho, D. N., Chan, T. L., Ma, E. S. K., Akbari, M. R. & Narod, S. A. 2016. Germline RECQL mutations in high risk Chinese breast cancer patients. *Breast Cancer Res Treat,* Vol. 157, No. 2**,** 211-215.

Lalloo, F. & Evans, D. G. 2012. Familial breast cancer. *Clin Genet,* Vol. 82, No. 2**,** 105-114.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. & Maglott, D. R. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research,* Vol. 42, No. Database issue**,** D980-D985.

Lara, R., Mauri, F. A., Taylor, H., Derua, R., Shia, A., Gray, C., Nicols, A., Shiner, R. J., Schofield, E., Bates, P. A., Waelkens, E., Dallman, M., Lamb, J., Zicha, D., Downward, J., Seckl, M. J. & Pardo, O. E. 2011. An siRNA screen identifies RSK1 as a key modulator of lung cancer metastasis. *Oncogene,* Vol. 30, No. 32**,** 3513-3521.

Larsen, M. J., Kruse, T. A., Tan, Q., Laenkholm, A. V., Bak, M., Lykkesfeldt, A. E., Sorensen, K. P., Hansen, T. V., Ejlertsen, B., Gerdes, A. M. & Thomassen, M. 2013. Classifications within molecular subtypes enables identification of BRCA1/BRCA2 mutation carriers by RNA tumor profiling. *PLoS One,* Vol. 8, No. 5**,** e64268.

Larson, G. P., Zhang, G., Ding, S., Foldenauer, K., Udar, N., Gatti, R. A., Neuberg, D., Lunetta, K. L., Ruckdeschel, J. C., Longmate, J., Flanagan, S. & Krontiris, T. G. 1997. An allelic variant at the ATM locus is implicated in breast cancer susceptibility. *Genet Test,* Vol. 1, No. 3**,** 165-170.

Le Gallo, M., O'hara, A. J., Rudd, M. L., Urick, M. E., Hansen, N. F., O'neil, N. J., Price, J. C., Zhang, S., England, B. M., Godwin, A. K., Sgroi, D. C., Hieter, P., Mullikin, J. C., Merino, M. J. & Bell, D. W. 2012. Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat Genet,* Vol. 44, No. 12**,** 1310-1315.

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb De Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F. & Antoniou, A. C. 2019. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine,* Vol. 21, No. 8**,** 1708-1718.

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,* Vol. 404, No. 6774**,** 201-204.

Lengyel, E. 2010. Ovarian cancer development and metastasis. *Am J Pathol,* Vol. 177, No. 3**,** 1053-1064.

Li, D., Wang, Q., Liu, C., Duan, H., Zeng, X., Zhang, B., Li, X., Zhao, J., Tang, S., Li, Z., Xing, X., Yang, P., Chen, L., Zeng, J., Zhu, X., Zhang, S., Zhang, Z., Ma, L., He, Z., Wang, E., Xiao, Y., Zheng, Y. & Chen, W. 2012. Aberrant expression of miR-638 contributes to benzo(a)pyrene-induced human cell transformation. *Toxicol Sci,* Vol. 125, No. 2**,** 382-391.

Li, H., Ruan, J. & Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research,* Vol. 18, No. 11**,** 1851-1858.

Li, N., Rowley, S. M., Goode, D. L., Amarasinghe, K. C., Mcinerny, S., Devereux, L., Wong-Brown, M. W., Lupat, R., Lee, J. E. A., Hughes, S., Thompson, E. R., Zethoven, M., Li, J., Trainer, A. H., Gorringe, K. L., Scott,

R. J., James, P. A., Campbell, I. G. & Lifepool, I. 2018. Mutations in RECQL are not associated with breast cancer risk in an Australian population. *Nature Genetics,* Vol. 50, No. 10**,** 1346-1348.

Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., Carte, J., Chen, W., Roark, N., Ranganathan, S., Ravinder, N. & Chesnut, J. D. 2015. Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *Journal of Biotechnology,* Vol. 208, No. 1**,** 44-53.

Liberal, V., Martinsson-Ahlzen, H. S., Liberal, J., Spruck, C. H., Widschwendter, M., Mcgowan, C. H. & Reed, S. I. 2012. Cyclin-dependent kinase subunit (Cks) 1 or Cks2 overexpression overrides the DNA damage response barrier triggered by activated oncoproteins. *Proc Natl Acad Sci U S A,* Vol. 109, No. 8**,** 2754-2759.

Liebens, F. P., Carly, B., Pastijn, A. & Rozenberg, S. 2007. Management of BRCA1/2 associated breast cancer: a systematic qualitative review of the state of knowledge in 2006. *Eur J Cancer,* Vol. 43, No. 2**,** 238-257.

Lieber, M. R. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual review of biochemistry,* Vol. 79, No. 1**,** 181-211.

Liehr, J. G. 1990. Genotoxic effects of estrogens. *Mutat Res,* Vol. 238, No. 3**,** 269-276.

Lifton, R. P. 2010. Individual genomes on the horizon. *N Engl J Med,* Vol. 362, No. 13**,** 1235-1236.

Lin, D. C., Hao, J. J., Nagata, Y., Xu, L., Shang, L., Meng, X., Sato, Y., Okuno, Y., Varela, A. M., Ding, L. W., Garg, M., Liu, L. Z., Yang, H., Yin, D., Shi, Z. Z., Jiang, Y. Y., Gu, W. Y., Gong, T., Zhang, Y., Xu, X., Kalid, O., Shacham, S., Ogawa, S., Wang, M. R. & Koeffler, H. P. 2014a. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet,* Vol. 46, No. 5**,** 467-473.

Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. 2014b. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife,* Vol. 15, No. 3**,** e04766.

Lin, S. Y., Rai, R., Li, K., Xu, Z. X. & Elledge, S. J. 2005. BRIT1/MCPH1 is a DNA damage responsive protein that regulates the Brca1-Chk1 pathway, implicating checkpoint dysfunction in microcephaly. *Proc Natl Acad Sci U S A,* Vol. 102, No. 42**,** 15105-15109.

Lin, Y., Chen, Y., Chang, H. & Li, S. 2009. Nature of genetic variants in the *BRCA1* and *BRCA2* genes from breast caqncer families in Taiwan. *Life Science Journal,* Vol. 6, No. 3**,** 99-103.

Lincoln, S. E., Kobayashi, Y., Anderson, M. J., Yang, S., Desmond, A. J., Mills, M. A., Nilsen, G. B., Jacobs, K. B., Monzon, F. A., Kurian, A. W., Ford, J. M. & Ellisen, L. W. 2015. A Systematic Comparison of Traditional and Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Genes in More Than 1000 Patients. *The Journal of Molecular Diagnostics,* Vol. 17, No. 5**,** 533-544.

Liu, C., Wang, Y., Wang, Q. S. & Wang, Y. J. 2012. The CHEK2 I157T variant and breast cancer susceptibility: a systematic review and meta-analysis. *Asian Pac J Cancer Prev,* Vol. 13, No. 4**,** 1355-1360.

Livraghi, L. & Garber, J. E. 2015. PARP inhibitors in the management of breast cancer: current data and future prospects. *BMC Med,* Vol. 13, No. 13**,** 188.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. & Pallen, M. J. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol,* Vol. 30, No. 5**,** 434-439.

Lord, C. J., Tutt, A. N. & Ashworth, A. 2015. Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annu Rev Med,* Vol. 66, No. 1**,** 455-470.

Lv, X.-B., Jiao, Y., Qing, Y., Hu, H., Cui, X., Lin, T., Song, E. & Yu, F. 2011. miR-124 suppresses multiple steps of breast cancer metastasis by targeting a cohort of pro-metastatic genes in vitro. *Chinese journal of cancer,* Vol. 30, No. 12**,** 821-830.

Macarthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A., Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W. & Gunter, C. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature,* Vol. 508, No. 7497**,** 469-476.

Macklin, S., Durand, N., Atwal, P. & Hines, S. 2018. Observed frequency and challenges of variant reclassification in a hereditary cancer clinic. *Genet Med,* Vol. 20, No. 3**,** 346-350.

Magnussen, G. I., Hellesylt, E., Nesland, J. M., Trope, C. G., Florenes, V. A. & Holm, R. 2013. High expression of wee1 is associated with malignancy in vulvar squamous cell carcinoma patients. *BMC Cancer,* Vol. 13, No. 14**,** 288.

Mak, C. C. Y., Leung, G. K. C., Mok, G. T. K., Yeung, K. S., Yang, W., Fung, C.-W., Chan, S. H. S., Lee, S.-L., Lee, N.-C., Pfundt, R., Lau, Y.-L. & Chung, B. H. Y. 2018. Exome sequencing for paediatric-onset diseases: impact of the extensive involvement of medical geneticists in the diagnostic odyssey. *npj Genomic Medicine,* Vol. 3, No. 1**,** 1-19.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J., Charpentier, E., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J., Terns, R. M., Terns, M. P., White, M. F., Yakunin, A. F., Garrett, R. A., Van Der Oost, J., Backofen, R. & Koonin, E. V. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol,* Vol. 13, No. 11**,** 722-736.

Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., Yang, L. & Church, G. M. 2013a. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature biotechnology,* Vol. 31, No. 9**,** 833-838.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E. & Church, G. M. 2013b. RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.),* Vol. 339, No. 6121**,** 823-826.

Malina, A., Mills, J. R., Cencic, R., Yan, Y., Fraser, J., Schippers, L. M., Paquet, M., Dostie, J. & Pelletier, J. 2013. Repurposing CRISPR/Cas9 for in situ functional assays. *Genes & development,* Vol. 27, No. 23**,** 2602-2614.

Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Jr., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A. & Et Al. 1990. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science,* Vol. 250, No. 4985**,** 1233-1238.

Marella, N. V., Malyavantham, K. S., Wang, J., Matsui, S.-I., Liang, P. & Berezney, R. 2009. Cytogenetic and cDNA microarray expression analysis of MCF10 human breast cancer progression cell lines. *Cancer research,* Vol. 69, No. 14**,** 5946-5953.

Mariotti, L. G., Pirovano, G., Savage, K. I., Ghita, M., Ottolenghi, A., Prise, K. M. & Schettino, G. 2013. Use of the γ-H2AX assay to investigate DNA repair dynamics following multiple radiation exposures. *PloS one,* Vol. 8, No. 11**,** e79541-e79541.

Marraffini, L. A. & Sontheimer, E. J. 2008. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science,* Vol. 322, No. 5909**,** 1843-1845.

Marroni, F., Pinosio, S. & Morgante, M. 2012. The quest for rare variants: pooled multiplexed next generation sequencing in plants. *Front Plant Sci,* Vol. 3, No. 1**,** 133.

Martin-Ezquerra, G., Salgado, R., Toll, A., Baro, T., Mojal, S., Yebenes, M., Garcia-Muret, M. P., Sole, F., Quitllet, F. A., Espinet, B. & Pujol, R. M. 2011. CDC28 protein kinase regulatory subunit 1B (CKS1B) expression and genetic status analysis in oral squamous cell carcinoma. *Histol Histopathol,* Vol. 26, No. 1**,** 71-77.

Masciari, S., Larsson, N., Senz, J., Boyd, N., Kaurah, P., Kandel, M. J., Harris, L. N., Pinheiro, H. C., Troussard, A., Miron, P., Tung, N., Oliveira, C., Collins, L., Schnitt, S., Garber, J. E. & Huntsman, D. 2007. Germline E-cadherin mutations in familial lobular breast cancer. *J Med Genet,* Vol. 44, No. 11**,** 726-731.

Masson, A. L., Talseth-Palmer, B. A., Evans, T. J., Grice, D. M., Hannan, G. N. & Scott, R. J. 2014. Expanding the genetic basis of copy number variation in familial breast cancer. *Hered Cancer Clin Pract,* Vol. 12, No. 1**,** 15.

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., Watanabe, T., Kanai, T. & Sato, T. 2015. Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat Med,* Vol. 21, No. 3**,** 256-262.

Matsuoka, S., Huang, M. & Elledge, S. J. 1998. Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science,* Vol. 282, No. 5395**,** 1893-1897.

Mavaddat, N., Barrowdale, D., Andrulis, I. L., Domchek, S. M., Eccles, D., Nevanlinna, H., Ramus, S. J., Spurdle, A., Robson, M., Sherman, M., Mulligan, A. M., Couch, F. J., Engel, C., Mcguffog, L., Healey, S., Sinilnikova, O. M., Southey, M. C., Terry, M. B., Goldgar, D., Malley, F., John, E. M., Janavicius, R., Tihomirova, L., Hansen, T. V. O., Nielsen, F. C., Osorio, A., Stavropoulou, A., Benítez, J., Manoukian, S., Peissel, B., Barile, M., Volorio, S., Pasini, B., Dolcetti, R., Putignano, A. L., Ottini, L., Radice, P., Hamann, U., Rashid, M. U., Hogervorst, F. B., Kriege, M., Van Der Luijt, R. B., Peock, S., Frost, D., Evans, D. G., Brewer, C., Walker, L., Rogers, M. T., Side, L. E., Houghton, C., Weaver, J., Godwin, A. K., Schmutzler, R. K., Wappenschmidt, B., Meindl, A., Kast, K., Arnold, N., Niederacher, D., Sutter, C., Deissler, H., Gadzicki, D., Preisler-Adams, S., Varon-Mateeva, R., Schönbuchner, I., Gevensleben, H., Stoppa-Lyonnet, D., Belotti, M., Barjhoux, L., Isaacs, C., Peshkin, B. N., Caldes, T., De La Hoya, M.,

Cañadas, C., Heikkinen, T., Heikkilä, P., Aittomäki, K., Blanco, I., Lazaro, C., Brunet, J., Agnarsson, B. A., Arason, A., Barkardottir, R. B., Dumont, M., Simard, J., Montagna, M., Agata, S., Andrea, E., Yan, M., Fox, S., Rebbeck, T. R., Rubinstein, W., Tung, N., Garber, J. E., Wang, X., Fredericksen, Z., Pankratz, V. S., Lindor, N. M., Szabo, C., Offit, K., Sakr, R.*, et al.* 2012. Pathology of Breast and Ovarian Cancers among BRCA1 and BRCA2 Mutation Carriers: Results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiology Biomarkers and Prevention,* Vol. 21, No. 1**,** 134-147.

Maxwell, C. A., Benítez, J., Gómez-Baldó, L., Osorio, A., Bonifaci, N., Fernández-Ramires, R., Costes, S. V., Guinó, E., Chen, H., Evans, G. J. R., Mohan, P., Català, I., Petit, A., Aguilar, H., Villanueva, A., Aytes, A., Serra-Musach, J., Rennert, G., Lejbkowicz, F., Peterlongo, P., Manoukian, S., Peissel, B., Ripamonti, C. B., Bonanni, B., Viel, A., Allavena, A., Bernard, L., Radice, P., Friedman, E., Kaufman, B., Laitman, Y., Dubrovsky, M., Milgrom, R., Jakubowska, A., Cybulski, C., Gorski, B., Jaworska, K., Durda, K., Sukiennicki, G., Lubiński, J., Shugart, Y. Y., Domchek, S. M., Letrero, R., Weber, B. L., Hogervorst, F. B. L., Rookus, M. A., Collee, J. M., Devilee, P., Ligtenberg, M. J., Van Der Luijt, R. B., Aalfs, C. M., Waisfisz, Q., Wijnen, J., Van Roozendaal, C. E. P., Hebon, Embrace, Easton, D. F., Peock, S., Cook, M., Oliver, C., Frost, D., Harrington, P., Evans, D. G., Lalloo, F., Eeles, R., Izatt, L., Chu, C., Eccles, D., Douglas, F., Brewer, C., Nevanlinna, H., Heikkinen, T., Couch, F. J., Lindor, N. M., Wang, X., Godwin, A. K., Caligo, M. A., Lombardi, G., Loman, N., Karlsson, P., Ehrencrona, H., Von Wachenfeldt, A., Swe, B., Bjork Barkardottir, R., Hamann, U., Rashid, M. U., Lasa, A., Caldés, T., Andrés, R., Schmitt, M., Assmann, V., Stevens, K., Offit, K., Curado, J., Tilgner, H., Guigó, R., Aiza, G., Brunet, J., Castellsagué, J., Martrat, G.*, et al.* 2011. Interplay between BRCA1 and RHAMM Regulates Epithelial Apicobasal Polarization and May Influence Risk of Breast Cancer. *PLOS Biology,* Vol. 9, No. 11**,** e1001199.

Maxwell, K. N., Hart, S. N., Vijai, J., Schrader, K. A., Slavin, T. P., Thomas, T., Wubbenhorst, B., Ravichandran, V., Moore, R. M., Hu, C., Guidugli, L., Wenz, B., Domchek, S. M., Robson, M. E., Szabo, C., Neuhausen, S. L., Weitzel, J. N., Offit, K., Couch, F. J. & Nathanson, K. L. 2016. Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. *American journal of human genetics,* Vol. 98, No. 5**,** 801-817.

Mccall, C. M., Mosier, S., Thiess, M., Debeljak, M., Pallavajjala, A., Beierl, K., Deak, K. L., Datto, M. B., Gocke, C. D., Lin, M.-T. & Eshleman, J. R. 2014. False Positives in Multiplex PCR-Based Next-Generation Sequencing Have Unique Signatures. *The Journal of Molecular Diagnostics,* Vol. 16, No. 5**,** 541-549.

Mccarthy, H. J., Bierzynska, A., Wherlock, M., Ognjanovic, M., Kerecuk, L., Hegde, S., Feather, S., Gilbert, R. D., Krischock, L., Jones, C., Sinha, M. D., Webb, N. J., Christian, M., Williams, M. M., Marks, S., Koziell, A., Welsh, G. I. & Saleem, M. A. 2013. Simultaneous sequencing of 24 genes associated with steroid-resistant nephrotic syndrome. *Clin J Am Soc Nephrol,* Vol. 8, No. 4**,** 637-648.

Mcgowan, C. H. & Russell, P. 1993. Human Wee1 kinase inhibits cell division by phosphorylating p34cdc2 exclusively on Tyr15. *Embo j,* Vol. 12, No. 1**,** 75-85.

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & Depristo, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* Vol. 20, No. 9**,** 1297-1303.

Mefford, H. C., Baumbach, L., Panguluri, R. C., Whitfield-Broome, C., Szabo, C., Smith, S., King, M. C., Dunston, G., Stoppa-Lyonnet, D. & Arena, F. 1999. Evidence for a BRCA1 founder mutation in families of West African ancestry. *Am J Hum Genet,* Vol. 65, No. 2**,** 575-578.

Meijers-Heijboer, H., Van Den Ouweland, A., Klijn, J., Wasielewski, M., De Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., Van Veghel-Plandsoen, M., Elstrodt, F., Van Duijn, C., Bartels, C., Meijers, C., Schutte, M., Mcguffog, L., Thompson, D., Easton, D., Sodha, N., Seal, S., Barfoot, R., Mangion, J., Chang-Claude, J., Eccles, D., Eeles, R., Evans, D. G., Houlston, R., Murday, V., Narod, S., Peretz, T., Peto, J., Phelan, C., Zhang, H. X., Szabo, C., Devilee, P., Goldgar, D., Futreal, P. A., Nathanson, K. L., Weber, B., Rahman, N. & Stratton, M. R. 2002. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet,* Vol. 31, No. 1**,** 55-59.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* Vol. 11, No. 1**,** 31-46.

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., Rahman, N., Turnbull, C., Fletcher, O., Peto, J., Gibson, L., Dos-Santos-Silva, I., Chang-Claude, J., Flesch-Janys, D., Rudolph, A., Eilber, U., Behrens, S.,

Nevanlinna, H., Muranen, T. A., Aittomaki, K., Blomqvist, C., Khan, S., Aaltonen, K., Ahsan, H., Kibriya, M. G., Whittemore, A. S., John, E. M., Malone, K. E., Gammon, M. D., Santella, R. M., Ursin, G., Makalic, E., Schmidt, D. F., Casey, G., Hunter, D. J., Gapstur, S. M., Gaudet, M. M., Diver, W. R., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Berg, C. D., Chanock, S. J., Figueroa, J., Hoover, R. N., Lambrechts, D., Neven, P., Wildiers, H., Van Limbergen, E., Schmidt, M. K., Broeks, A., Verhoef, S., Cornelissen, S., Couch, F. J., Olson, J. E., Hallberg, E., Vachon, C., Waisfisz, Q., Meijers-Heijboer, H., Adank, M. A., Van Der Luijt, R. B., Li, J., Liu, J., Humphreys, K., Kang, D., Choi, J. Y., Park, S. K., Yoo, K. Y., Matsuo, K., Ito, H., Iwata, H., Tajima, K., Guenel, P., Truong, T., Mulot, C., Sanchez, M., Burwinkel, B., Marme, F., Surowy, H., Sohn, C., Wu, A. H., Tseng, C. C., Van Den Berg, D., Stram, D. O., Gonzalez-Neira, A., Benitez, J.*, et al.* 2015. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet,* Vol. 47, No. 4**,** 373-380.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., Bolla, M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., Wang, Z., Allen, J., Keeman, R., Eilber, U., French, J. D., Qing Chen, X., Fachal, L., Mccue, K., Mccart Reed, A. E., Ghoussaini, M., Carroll, J. S., Jiang, X., Finucane, H., Adams, M., Adank, M. A., Ahsan, H., Aittomäki, K., Anton-Culver, H., Antonenkova, N. N., Arndt, V., Aronson, K. J., Arun, B., Auer, P. L., Bacot, F., Barrdahl, M., Baynes, C., Beckmann, M. W., Behrens, S., Benitez, J., Bermisheva, M., Bernstein, L., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bonanni, B., Børresen-Dale, A.-L., Brand, J. S., Brauch, H., Brennan, P., Brenner, H., Brinton, L., Broberg, P., Brock, I. W., Broeks, A., Brooks-Wilson, A., Brucker, S. Y., Brüning, T., Burwinkel, B., Butterbach, K., Cai, Q., Cai, H., Caldés, T., Canzian, F., Carracedo, A., Carter, B. D., Castelao, J. E., Chan, T. L., David Cheng, T.-Y., Seng Chia, K., Choi, J.-Y., Christiansen, H., Clarke, C. L., Collaborators, N., Collée, M., Conroy, D. M., Cordina-Duverger, E., Cornelissen, S., Cox, D. G., Cox, A., Cross, S. S., Cunningham, J. M., Czene, K., Daly, M. B., Devilee, P., Doheny, K. F., Dörk, T., Dos-Santos-Silva, I., Dumont, M., Durcan, L., Dwek, M., Eccles, D. M., Ekici, A. B., Eliassen, A. H., Ellberg, C., Elvira, M.*, et al.* 2017. Association analysis identifies 65 new breast cancer risk loci. *Nature,* Vol. 551, No. 7678**,** 92-94.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W. & Et Al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science,* Vol. 266, No. 5182**,** 66-71.

Miller, C. W., Ikezoe, T., Krug, U., Hofmann, W. K., Tavor, S., Vegesna, V., Tsukasaki, K., Takeuchi, S. & Koeffler, H. P. 2002. Mutations of the CHK2 gene are found in some osteosarcomas, but are rare in breast, lung, and ovarian tumors. *Genes Chromosomes Cancer,* Vol. 33, No. 1**,** 17-21.

Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A. B., Kramer, J. L., Rowland, J. H., Stein, K. D., Alteri, R. & Jemal, A. 2016. Cancer treatment and survivorship statistics, 2016. *CA Cancer J Clin,* Vol. 66, No. 4**,** 271-289.

Mitrunen, K. & Hirvonen, A. 2003. Molecular epidemiology of sporadic breast cancer: The role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutation Research/Reviews in Mutation Research,* Vol. 544, No. 1**,** 9-41.

Moinova, H. R., Chen, W. D., Shen, L., Smiraglia, D., Olechnowicz, J., Ravi, L., Kasturi, L., Myeroff, L., Plass, C., Parsons, R., Minna, J., Willson, J. K., Green, S. B., Issa, J. P. & Markowitz, S. D. 2002. HLTF gene silencing in human colon cancer. *Proc Natl Acad Sci U S A,* Vol. 99, No. 7**,** 4562-4567.

Momozawa, Y., Iwasaki, Y., Parsons, M. T., Kamatani, Y., Takahashi, A., Tamura, C., Katagiri, T., Yoshida, T., Nakamura, S., Sugano, K., Miki, Y., Hirata, M., Matsuda, K., Spurdle, A. B. & Kubo, M. 2018. Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. *Nature Communications,* Vol. 9, No. 1**,** 4083.

Montano, M. M., Krishnamurthy, N. & Sripathy, S. 2012. Targeting the Genotoxic effects of Estrogens. *Drug discovery today. Disease mechanisms,* Vol. 9, No. 1-2**,** e29-e33.

Morris, E. J., Ji, J. Y., Yang, F., Di Stefano, L., Herr, A., Moon, N. S., Kwon, E. J., Haigis, K. M., Naar, A. M. & Dyson, N. J. 2008. E2F1 represses beta-catenin transcription and is antagonized by both pRB and CDK8. *Nature,* Vol. 455, No. 7212**,** 552-556.

Morris, M. C., Kaiser, P., Rudyak, S., Baskerville, C., Watson, M. H. & Reed, S. I. 2003. Cks1-dependent proteasome recruitment and activation of CDC20 transcription in budding yeast. *Nature,* Vol. 423, No. 6943**,** 1009-1013.

Mosesson, Y., Mills, G. B. & Yarden, Y. 2008. Derailed endocytosis: an emerging feature of cancer. *Nat Rev Cancer,* Vol. 8, No. 11**,** 835-850.

Mouradov, D., Sloggett, C., Jorissen, R. N., Love, C. G., Li, S., Burgess, A. W., Arango, D., Strausberg, R. L., Buchanan, D., Wormald, S., O'connor, L., Wilding, J. L., Bicknell, D., Tomlinson, I. P., Bodmer, W. F., Mariadason, J. M. & Sieber, O. M. 2014. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res,* Vol. 74, No. 12**,** 3238-3247.

Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. 1999. Brca1 controls homology-directed DNA repair. *Mol Cell,* Vol. 4, No. 4**,** 511-518.

Mu, W., Lu, H.-M., Chen, J., Li, S. & Elliott, A. 2016. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *J Mol Diagn,* Vol. 18, No. 6**,** 923-932.

Muranen, T. A., Blomqvist, C., Dork, T., Jakubowska, A., Heikkila, P., Fagerholm, R., Greco, D., Aittomaki, K., Bojesen, S. E., Shah, M., Dunning, A. M., Rhenius, V., Hall, P., Czene, K., Brand, J. S., Darabi, H., Chang-Claude, J., Rudolph, A., Nordestgaard, B. G., Couch, F. J., Hart, S. N., Figueroa, J., Garcia-Closas, M., Fasching, P. A., Beckmann, M. W., Li, J., Liu, J., Andrulis, I. L., Winqvist, R., Pylkas, K., Mannermaa, A., Kataja, V., Lindblom, A., Margolin, S., Lubinski, J., Dubrowinskaja, N., Bolla, M. K., Dennis, J., Michailidou, K., Wang, Q., Easton, D. F., Pharoah, P. D., Schmidt, M. K. & Nevanlinna, H. 2016. Patient survival and tumor characteristics associated with CHEK2:p.I157T - findings from the Breast Cancer Association Consortium. *Breast Cancer Res,* Vol. 18, No. 1**,** 98.

Muraoka, M., Konishi, M., Kikuchi-Yanoshita, R., Tanaka, K., Shitara, N., Chong, J. M., Iwama, T. & Miyaki, M. 1996. p300 gene alterations in colorectal and gastric carcinomas. *Oncogene,* Vol. 12, No. 7**,** 1565-1569.

Murria Estal, R., Palanca Suela, S., De Juan Jimenez, I., Egoavil Rojas, C., Garcia-Casado, Z., Juan Fita, M. J., Sanchez Heras, A. B., Segura Huerta, A., Chirivella Gonzalez, I., Sanchez-Izquierdo, D., Llop Garcia, M., Barragan Gonzalez, E. & Bolufer Gilabert, P. 2013. MicroRNA signatures in hereditary breast cancer. *Breast Cancer Res Treat,* Vol. 142, No. 1**,** 19-30.

Nagarajan, N. & Pop, M. 2010. Sequencing and genome assembly using next-generation technologies. *Methods Mol Biol,* Vol. 673, No. 1**,** 1-17.

National Breast Cancer Foundation. 2014. *About Breast Cancer* [Online]. National Breast Cancer Foundation, . Available: http://www.nbcf.org.au/research/about-breast-cancer.aspx [Accessed 1 September 2014].

Need, A. C., Shashi, V., Hitomi, Y., Schoch, K., Shianna, K. V., Mcdonald, M. T., Meisler, M. H. & Goldstein, D. B. 2012. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet,* Vol. 49, No. 6**,** 353-361.

Nevanlinna, H. & Bartek, J. 2006. The CHEK2 gene and inherited breast cancer susceptibility. *Oncogene,* Vol. 25, No. 43**,** 5912-5919.

Nguyen-Dumont, T., Myszka, A., Karpinski, P., Sasiadek, M. M., Akopyan, H., Hammet, F., Tsimiklis, H., Park, D. J., Pope, B. J., Slezak, R., Kitsera, N., Siekierzynska, A. & Southey, M. C. 2018. FANCM and RECQL genetic variants and breast cancer susceptibility: relevance to South Poland and West Ukraine. *BMC Med Genet,* Vol. 19, No. 1**,** 12.

Nikkilä, J., Coleman, K. A., Morrissey, D., Pylkäs, K., Erkko, H., Messick, T. E., Karppinen, S. M., Amelina, A., Winqvist, R. & Greenberg, R. A. 2009. Familial breast cancer screening reveals an alteration in the RAP80 UIM domain that impairs DNA damage response function. *Oncogene,* Vol. 28, No. 16**,** 1843-1852.

Norton, N., Robertson, P. D., Rieder, M. J., Züchner, S., Rampersaud, E., Martin, E., Li, D., Nickerson, D. A., Hershberger, R. E., National Heart, L. & Blood Institute, G. O. E. S. P. 2012. Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. *Circulation. Cardiovascular genetics,* Vol. 5, No. 2**,** 167-174.

O'leary, E., Iacoboni, D., Holle, J., Michalski, S. T., Esplin, E. D., Yang, S. & Ouyang, K. 2017. Expanded Gene Panel Use for Women With Breast Cancer: Identification and Intervention Beyond Breast Cancer Risk. *Annals of surgical oncology,* Vol. 24, No. 10**,** 3060-3066.

Oakley, G. G. & Patrick, S. M. 2010. Replication protein A: directing traffic at the intersection of replication and repair. *Front Biosci (Landmark Ed),* Vol. 15, No. 1**,** 883-900.

Ogawa, H., Ishiguro, K., Gaubatz, S., Livingston, D. M. & Nakatani, Y. 2002. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. *Science,* Vol. 296, No. 5570**,** 1132-1136.

Okumura, K., Nogami, M., Taguchi, H., Dean, F. B., Chen, M., Pan, Z. Q., Hurwitz, J., Shiratori, A., Murakami, Y. & Ozawa, K. 1995. Assignment of the 36.5-kDa (RFC5), 37-kDa (RFC4), 38-kDa (RFC3), and 40-kDa (RFC2) subunit genes of human replication factor C to chromosome bands 12q24.2-q24.3, 3q27, 13q12.3-q13, and 7q11.23. *Genomics,* Vol. 25, No. 1**,** 274-278.

Oliveira, A. M., Ross, J. S. & Fletcher, J. A. 2005. Tumor suppressor genes in breast cancer: the gatekeepers and the caretakers. *Am J Clin Pathol,* Vol. 124, No. 1**,** 16-28.

Olivier, M., Hollstein, M. & Hainaut, P. 2010. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology,* Vol. 2, No. 1**,** a001008-a001008.

Olsen, J. H., Hahnemann, J. M., Børresen-Dale, A.-L., Brøndum-Nielsen, K., Hammarström, L., Kleinerman, R., Kääriäinen, H., Lönnqvist, T., Sankila, R., Seersholm, N., Tretli, S., Yuen, J., Boice, J. D. & Tucker, M. 2001. Cancer in Patients With Ataxia-Telangiectasia and in Their Relatives in the Nordic Countries. *Journal of the National Cancer Institute,* Vol. 93, No. 2**,** 121-127.

Ooki, A., Yamashita, K., Yamaguchi, K., Mondal, A., Nishimiya, H. & Watanabe, M. 2013. DNA damage-inducible gene, reprimo functions as a tumor suppressor and is suppressed by promoter methylation in gastric cancer. *Mol Cancer Res,* Vol. 11, No. 11**,** 1362-1374.

Pang, A. W., Macdonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L. & Scherer, S. W. 2010. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol,* Vol. 11, No. 5**,** R52.

Park, D. J., Lesueur, F., Nguyen-Dumont, T., Pertesi, M., Odefrey, F., Hammet, F., Neuhausen, S. L., John, E. M., Andrulis, I. L., Terry, M. B., Daly, M., Buys, S., Le Calvez-Kelm, F., Lonie, A., Pope, B. J., Tsimiklis, H., Voegele, C., Hilbers, F. M., Hoogerbrugge, N., Barroso, A., Osorio, A., Giles, G. G., Devilee, P., Benitez, J., Hopper, J. L., Tavtigian, S. V., Goldgar, D. E. & Southey, M. C. 2012. Rare mutations in XRCC2 increase the risk of breast cancer. *Am J Hum Genet,* Vol. 90, No. 4**,** 734-739.

Parris, C. N., Adam Zahir, S., Al-Ali, H., Bourton, E. C., Plowman, C. & Plowman, P. N. 2015. Enhanced γ-H2AX DNA damage foci detection using multimagnification and extended depth of field in imaging flow cytometry. *Cytometry Part A,* Vol. 87, No. 8**,** 717-723.

Patel, K. J., Yu, V. P., Lee, H., Corcoran, A., Thistlethwaite, F. C., Evans, M. J., Colledge, W. H., Friedman, L. S., Ponder, B. A. & Venkitaraman, A. R. 1998. Involvement of Brca2 in DNA repair. *Mol Cell,* Vol. 1, No. 3**,** 347-357.

Paull, T. T. & Gellert, M. 1998. The 3' to 5' exonuclease activity of Mre 11 facilitates repair of DNA double-strand breaks. *Mol Cell,* Vol. 1, No. 7**,** 969-979.

Petitjean, A., Achatz, M. I., Borresen-Dale, A. L., Hainaut, P. & Olivier, M. 2007. TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene,* Vol. 26, No. 15**,** 2157-2165.

Petrovic, N., Davidovic, R., Bajic, V., Obradovic, M. & Isenovic, R. E. 2017. MicroRNA in breast cancer: The association with BRCA1/2. *Cancer Biomark,* Vol. 19, No. 2**,** 119-128.

Pharoah, P. D., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F. & Ponder, B. A. 2002. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet,* Vol. 31, No. 1**,** 33-36.

Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., Hogervorst, F. B., Hoogerbrugge, N., Spurdle, A. B. & Tavtigian, S. V. 2008. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat,* Vol. 29, No. 11**,** 1282-1291.

Portela, A. & Esteller, M. 2010. Epigenetic modifications and human disease. *Nat Biotechnol,* Vol. 28, No. 10**,** 1057-1068.

Porter, C. C., Kim, J., Fosmire, S., Gearheart, C. M., Van Linden, A., Baturin, D., Zaberezhnyy, V., Patel, P. R., Gao, D., Tan, A. C. & Degregori, J. 2012. Integrated genomic analyses identify WEE1 as a critical mediator of cell fate and a novel therapeutic target in acute myeloid leukemia. *Leukemia,* Vol. 26, No. 6**,** 1266-1276.

Prapa, M., Solomons, J. & Tischkowitz, M. 2017. The use of panel testing in familial breast and ovarian cancer. *Clinical medicine (London, England),* Vol. 17, No. 6**,** 568-572.

Pujana, M. A., Han, J. D., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J. F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Sole, X., Hernandez, P., Lazaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D. & Vidal, M. 2007. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet,* Vol. 39, No. 11**,** 1338-1349.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. & Turner, D. J. 2008. A large genome center's improvements to the Illumina sequencing system. *Nat Methods,* Vol. 5, No. 12**,** 1005-1010.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* Vol. 13, No. 1**,** 341.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., Jayatilake, H., Mcguffog, L., Hanks, S., Evans, D. G., Eccles, D., Easton, D. F. & Stratton, M. R. 2007. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet,* Vol. 39, No. 2**,** 165-167.

Rajan, J. V., Wang, M., Marquis, S. T. & Chodosh, L. A. 1996. Brca2 is coordinately regulated with Brca1 during proliferation and differentiation in mammary epithelial cells. *Proc Natl Acad Sci U S A,* Vol. 93, No. 23**,** 13078-13083.

Ramus, S. J. & Gayther, S. A. 2009. The contribution of BRCA1 and BRCA2 to ovarian cancer. *Mol Oncol,* Vol. 3, No. 2**,** 138-150.

Ran, F. A., Hsu, Patrick d., Lin, C.-Y., Gootenberg, Jonathan s., Konermann, S., Trevino, A. E., Scott, David a., Inoue, A., Matoba, S., Zhang, Y. & Zhang, F. 2013a. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell,* Vol. 154, No. 6**,** 1380-1389.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. & Zhang, F. 2013b. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols,* Vol. 8, No. 11**,** 2281-2308.

Rebbeck, T. R., Friebel, T., Lynch, H. T., Neuhausen, S. L., Van 'T Veer, L., Garber, J. E., Evans, G. R., Narod, S. A., Isaacs, C., Matloff, E., Daly, M. B., Olopade, O. I. & Weber, B. L. 2004. Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. *J Clin Oncol,* Vol. 22, No. 6**,** 1055-1062.

Reid, E. S., Papandreou, A., Drury, S., Boustred, C., Yue, W. W., Wedatilake, Y., Beesley, C., Jacques, T. S., Anderson, G., Abulhoul, L., Broomfield, A., Cleary, M., Grunewald, S., Varadkar, S. M., Lench, N., Rahman, S., Gissen, P., Clayton, P. T. & Mills, P. B. 2016. Advantages and pitfalls of an extended gene panel for investigating complex neurometabolic phenotypes. *Brain : a journal of neurology,* Vol. 139, No. 11**,** 2844-2854.

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., North, B., Jayatilake, H., Barfoot, R., Spanova, K., Mcguffog, L., Evans, D. G., Eccles, D., Easton, D. F., Stratton, M. R. & Rahman, N. 2006. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet,* Vol. 38, No. 8**,** 873-875.

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J. M., Kim, J., Lawrence, M. S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., Hess, J., Stewart, C., Maruvka, Y. E., Stojanov, P., Cortes, M. L., Seepo, S., Cibulskis, C., Tracy, A., Pugh, T. J., Lee, J., Zheng, Z., Ellisen, L. W., Iafrate, A. J., Boehm, J. S., Gabriel, S. B., Meyerson, M., Golub, T. R., Baselga, J., Hidalgo-Miranda, A., Shioda, T., Bernards, A., Lander, E. S. & Getz, G. 2017. Recurrent and functional regulatory mutations in breast cancer. *Nature,* Vol. 547, No. 7661**,** 55-60.

Riaz, M., Van Jaarsveld, M. T. M., Hollestelle, A., Prager-Van Der Smissen, W. J. C., Heine, A. a. J., Boersma, A. W. M., Liu, J., Helmijr, J., Ozturk, B., Smid, M., Wiemer, E. A., Foekens, J. A. & Martens, J. W. M. 2013. miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs. *Breast cancer research : BCR,* Vol. 15, No. 2**,** R33.

Richard, J. L. C. & Eichhorn, P. J. A. 2018. Deciphering the roles of lncRNAs in breast development and disease. *Oncotarget,* Vol. 9, No. 28**,** 20179-20212.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. & Rehm, H. L. 2015. Standards and guidelines for the interpretation of sequence

variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med,* Vol. 17, No. 5**,** 405-424.

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J. M., Kuruvilla, F., Lagacé, C., Neale, B., Lo, K. S., Schumm, P., Törkvist, L., National Institute Of, D., Digestive Kidney Diseases Inflammatory Bowel Disease Genetics, C., United Kingdom Inflammatory Bowel Disease Genetics, C., International Inflammatory Bowel Disease Genetics, C., Dubinsky, M. C., Brant, S. R., Silverberg, M. S., Duerr, R. H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'amato, M., Mcgovern, D. P. B., Cho, J. H., Rioux, J. D., Xavier, R. J. & Daly, M. J. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics,* Vol. 43, No. 11**,** 1066-1073.

Rizzo, J. M. & Buck, M. J. 2012. Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev Res (Phila),* Vol. 5, No. 7**,** 887-900.

Roberts, M. E., Jackson, S. A., Susswein, L. R., Zeinomar, N., Ma, X., Marshall, M. L., Stettner, A. R., Milewski, B., Xu, Z., Solomon, B. D., Terry, M. B., Hruska, K. S., Klein, R. T. & Chung, W. K. 2018. MSH6 and PMS2 germ-line pathogenic variants implicated in Lynch syndrome are associated with breast cancer. *Genet Med,* Vol. 20, No. 10**,** 1167-1174.

Robertson, K. D. & Jones, P. A. 1999. Tissue-specific alternative splicing in the human INK4a/ARF cell cycle regulatory locus. *Oncogene,* Vol. 18, No. 26**,** 3810-3820.

Robin, J. D., Ludlow, A. T., Laranger, R., Wright, W. E. & Shay, J. W. 2016. Comparison of DNA Quantification Methods for Next Generation Sequencing. *Scientific Reports,* Vol. 6, No. 1**,** 24067.

Rodriguez-Lopez, R., Osorio, A., Ribas, G., Pollan, M., Sanchez-Pulido, L., De La Hoya, M., Ruibal, A., Zamora, P., Arias, J. I., Salazar, R., Vega, A., Martinez, J. I., Esteban-Cardenosa, E., Alonso, C., Leton, R., Urioste Azcorra, M., Miner, C., Armengod, M. E., Carracedo, A., Gonzalez-Sarmiento, R., Caldes, T., Diez, O. & Benitez, J. 2004. The variant E233G of the RAD51D gene could be a low-penetrance allele in high-risk breast cancer families without BRCA1/2 mutations. *Int J Cancer,* Vol. 110, No. 6**,** 845-849.

Rosenthal, E. T., Bernhisel, R., Brown, K., Kidd, J. & Manley, S. 2017. Clinical testing with a panel of 25 genes associated with increased cancer risk results in a significant increase in clinically significant findings across a broad range of cancer histories. *Cancer Genetics,* Vol. 218, No. 219**,** 58-68.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C. & Jaffe, D. B. 2013. Characterizing and measuring bias in sequence data. *Genome biology,* Vol. 14, No. 5**,** R51.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., Hoon, J., Simons, J. F., Marran, D., Myers, J. W., Davidson, J. F., Branting, A., Nobile, J. R., Puc, B. P., Light, D., Clark, T. A., Huber, M., Branciforte, J. T., Stoner, I. B., Cawley, S. E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J. A., Namsaraev, E., Mckernan, K. J., Williams, A., Roth, G. T. & Bustillo, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature,* Vol. 475, No. 7356**,** 348-352.

Rudd, M. F., Webb, E. L., Matakidou, A., Sellick, G. S., Williams, R. D., Bridle, H., Eisen, T. & Houlston, R. S. 2006. Variants in the GH-IGF axis confer susceptibility to lung cancer. *Genome Res,* Vol. 16, No. 6**,** 693-701.

Rusmini, M., Federici, S., Caroli, F., Grossi, A., Baldi, M., Obici, L., Insalaco, A., Tommasini, A., Caorsi, R., Gallo, E., Olivieri, A. N., Marzano, A., Coviello, D., Ravazzolo, R., Martini, A., Gattorno, M. & Ceccherini, I. 2016. Next-generation sequencing and its initial applications for molecular diagnosis of systemic auto-inflammatory diseases. *Annals of the Rheumatic Diseases,* Vol. 75, No. 8**,** 1550-1557.

Ryu, S., Han, J., Norden-Krichmar, T. M., Schork, N. J. & Suh, Y. 2018. Effective discovery of rare variants by pooled target capture sequencing: A comparative analysis with individually indexed target capture sequencing. *Mutat Res,* Vol. 809, No. 1**,** 24-31.

Saitsu, H., Nishimura, T., Muramatsu, K., Kodera, H., Kumada, S., Sugai, K., Kasai-Yoshida, E., Sawaura, N., Nishida, H., Hoshino, A., Ryujin, F., Yoshioka, S., Nishiyama, K., Kondo, Y., Tsurusaki, Y., Nakashima, M., Miyake, N., Arakawa, H., Kato, M., Mizushima, N. & Matsumoto, N. 2013. De novo mutations in the autophagy gene WDR45 cause static encephalopathy of childhood with neurodegeneration in adulthood. *Nat Genet,* Vol. 45, No. 4**,** 445-449.

Sánchez-Rivera, F. J. & Jacks, T. 2015. Applications of the CRISPR-Cas9 system in cancer biology. *Nature reviews. Cancer,* Vol. 15, No. 7**,** 387-395.

Sankaranarayanan, R. & Ferlay, J. 2006. Worldwide burden of gynaecological cancer: the size of the problem. *Best Pract Res Clin Obstet Gynaecol,* Vol. 20, No. 2**,** 207-225.

Sato, N., Fukushima, N., Matsubayashi, H., Iacobuzio-Donahue, C. A., Yeo, C. J. & Goggins, M. 2006. Aberrant methylation of Reprimo correlates with genetic instability and predicts poor prognosis in pancreatic ductal adenocarcinoma. *Cancer,* Vol. 107, No. 2**,** 251-257.

Savage, K. I. & Harkin, D. P. 2015. BRCA1, a 'complex' protein involved in the maintenance of genomic stability. *The FEBS Journal,* Vol. 282, No. 4**,** 630-646.

Savitsky, K., Sfez, S., Tagle, D. A., Ziv, Y., Sartiel, A., Collins, F. S., Shiloh, Y. & Rotman, G. 1995. The complete sequence of the coding region of the ATM gene reveals similarity to cell cycle regulators in different species. *Hum Mol Genet,* Vol. 4, No. 11**,** 2025-2032.

Sawyer, S. L., Hartley, T., Dyment, D. A., Beaulieu, C. L., Schwartzentruber, J., Smith, A., Bedford, H. M., Bernard, G., Bernier, F. P., Brais, B., Bulman, D. E., Warman Chardon, J., Chitayat, D., Deladoëy, J., Fernandez, B. A., Frosk, P., Geraghty, M. T., Gerull, B., Gibson, W., Gow, R. M., Graham, G. E., Green, J. S., Heon, E., Horvath, G., Innes, A. M., Jabado, N., Kim, R. H., Koenekoop, R. K., Khan, A., Lehmann, O. J., Mendoza-Londono, R., Michaud, J. L., Nikkel, S. M., Penney, L. S., Polychronakos, C., Richer, J., Rouleau, G. A., Samuels, M. E., Siu, V. M., Suchowersky, O., Tarnopolsky, M. A., Yoon, G., Zahir, F. R., Consortium, F. C., Consortium, C. R. C., Majewski, J. & Boycott, K. M. 2016. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clinical Genetics,* Vol. 89, No. 3**,** 275-284.

Schlotterer, C., Tobler, R., Kofler, R. & Nolte, V. 2014. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev Genet,* Vol. 15, No. 11**,** 749-763.

Schmidt, A. Y., Hansen, T. V. O., Ahlborn, L. B., Jonson, L., Yde, C. W. & Nielsen, F. C. 2017. Next-Generation Sequencing-Based Detection of Germline Copy Number Variations in BRCA1/BRCA2: Validation of a One-Step Diagnostic Workflow. *J Mol Diagn,* Vol. 19, No. 6**,** 809-816.

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,* Vol. 30, No. 12**,** e57.

Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., North, B., Mcguffog, L., Evans, D. G., Eccles, D., Easton, D. F., Stratton, M. R. & Rahman, N. 2006. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet,* Vol. 38, No. 11**,** 1239-1241.

Seitz, S., Rohde, K., Bender, E., Nothnagel, A., Pidde, H., Ullrich, O. M., El-Zehairy, A., Haensch, W., Jandrig, B., Kolble, K., Schlag, P. M. & Scherneck, S. 1997. Deletion mapping and linkage analysis provide strong indication for the involvement of the human chromosome region 8p12-p22 in breast carcinogenesis. *Br J Cancer,* Vol. 76, No. 8**,** 983-991.

Sellner, L. N. & Taylor, G. R. 2004. MLPA and MAPH: new techniques for detection of gene deletions. *Hum Mutat,* Vol. 23, No. 5**,** 413-419.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. & Zhang, F. 2014. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science,* Vol. 343, No. 6166**,** 84-87.

Shankar, G. M., Taylor-Weiner, A., Lelic, N., Jones, R. T., Kim, J. C., Francis, J. M., Abedalthagafi, M., Borges, L. F., Coumans, J. V., Curry, W. T., Nahed, B. V., Shin, J. H., Paek, S. H., Park, S. H., Stewart, C., Lawrence, M. S., Cibulskis, K., Thorner, A. R., Van Hummelen, P., Stemmer-Rachamimov, A. O., Batchelor, T. T., Carter, S. L., Hoang, M. P., Santagata, S., Louis, D. N., Barker, F. G., Meyerson, M., Getz, G., Brastianos, P. K. & Cahill, D. P. 2014. Sporadic hemangioblastomas are characterized by cryptic VHL inactivation. *Acta Neuropathol Commun,* Vol. 2, No. 1**,** 167.

Sharan, S. K., Morimatsu, M., Albrecht, U., Lim, D. S., Regel, E., Dinh, C., Sands, A., Eichele, G., Hasty, P. & Bradley, A. 1997. Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking Brca2. *Nature,* Vol. 386, No. 6627**,** 804-810.

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P. & Sharma, K. K. 2010. Various types and management of breast cancer: an overview. *Journal of advanced pharmaceutical technology & research,* Vol. 1, No. 2**,** 109-126.

Sharpe, H. J., Pau, G., Dijkgraaf, G. J., Basset-Seguin, N., Modrusan, Z., Januario, T., Tsui, V., Durham, A. B., Dlugosz, A. A., Haverty, P. M., Bourgon, R., Tang, J. Y., Sarin, K. Y., Dirix, L., Fisher, D. C., Rudin, C. M., Sofen, H., Migden, M. R., Yauch, R. L. & De Sauvage, F. J. 2015. Genomic analysis of smoothened inhibitor resistance in basal cell carcinoma. *Cancer Cell,* Vol. 27, No. 3**,** 327-341.

Shen, J., Ambrosone, C. B., Dicioccio, R. A., Odunsi, K., Lele, S. B. & Zhao, H. 2008. A functional polymorphism in the miR-146a gene and age of familial breast/ovarian cancer diagnosis. *Carcinogenesis,* Vol. 29, No. 10**,** 1963-1966.

Shi, X. B., Xue, L., Ma, A. H., Tepper, C. G., Gandour-Edwards, R., Kung, H. J. & Devere White, R. W. 2012. Tumor suppressive miR-124 targets androgen receptor and inhibits proliferation of prostate cancer cells. *Oncogene,* Vol. 32, No. 35**,** 4130-4138.

Shiovitz, S. & Korde, L. A. 2015. Genetics of breast cancer: a topic in evolution. *Annals of oncology : official journal of the European Society for Medical Oncology,* Vol. 26, No. 7**,** 1291-1299.

Silver, D. P., Richardson, A. L., Eklund, A. C., Wang, Z. C., Szallasi, Z., Li, Q., Juul, N., Leong, C. O., Calogrias, D., Buraimoh, A., Fatima, A., Gelman, R. S., Ryan, P. D., Tung, N. M., De Nicolo, A., Ganesan, S., Miron, A., Colin, C., Sgroi, D. C., Ellisen, L. W., Winer, E. P. & Garber, J. E. 2010. Efficacy of neoadjuvant Cisplatin in triple-negative breast cancer. *J Clin Oncol,* Vol. 28, No. 7**,** 1145-1153.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. & Ng, P. C. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research,* Vol. 40, No. 1**,** 452-457.

Slattery, M. L., Lundgreen, A., Herrick, J. S. & Wolff, R. K. 2011. Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. *Mutat Res,* Vol. 706, No. 1-2**,** 13-20.

Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X. & Zhang, F. 2016. Rationally engineered Cas9 nucleases with improved specificity. *Science,* Vol. 351, No. 6268**,** 84-88.

Sluiter, M., Mew, S. & Van Rensburg, E. J. 2009. PALB2 sequence variants in young South African breast cancer patients. *Fam Cancer,* Vol. 8, No. 4**,** 347-353.

Smith, M. L., Chen, I. T., Zhan, Q., Bae, I., Chen, C. Y., Gilmer, T. M., Kastan, M. B., O'connor, P. M. & Fornace, A. J., Jr. 1994. Interaction of the p53-regulated protein Gadd45 with proliferating cell nuclear antigen. *Science,* Vol. 266, No. 5189**,** 1376-1380.

Smith, P., Mcguffog, L., Easton, D. F., Mann, G. J., Pupo, G. M., Newman, B., Chenevix-Trench, G., Szabo, C., Southey, M., Renard, H., Odefrey, F., Lynch, H., Stoppa-Lyonnet, D., Couch, F., Hopper, J. L., Giles, G. G., Mccredie, M. R., Buys, S., Andrulis, I., Senie, R., Goldgar, D. E., Oldenburg, R., Kroeze-Jansema, K., Kraan, J., Meijers-Heijboer, H., Klijn, J. G., Van Asperen, C., Van Leeuwen, I., Vasen, H. F., Cornelisse, C. J., Devilee, P., Baskcomb, L., Seal, S., Barfoot, R., Mangion, J., Hall, A., Edkins, S., Rapley, E., Wooster, R., Chang-Claude, J., Eccles, D., Evans, D. G., Futreal, P., Nathanson, K. L., Weber, B. L., Rahman, N. & Stratton, M. R. 2006. A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosomes Cancer,* Vol. 45, No. 7**,** 646-655.

Soares, P., Berx, G., Van Roy, F. & Sobrinho-Simoes, M. 1997. E-cadherin gene alterations are rare events in thyroid tumors. *Int J Cancer,* Vol. 70, No. 1**,** 32-38.

Sobhian, B., Shao, G., Lilli, D. R., Culhane, A. C., Moreau, L. A., Xia, B., Livingston, D. M. & Greenberg, R. A. 2007. RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science,* Vol. 316, No. 5828**,** 1198-1202.

Sodha, N., Mantoni, T. S., Tavtigian, S. V., Eeles, R. & Garrett, M. D. 2006. Rare Germ Line CHEK2 Variants Identified in Breast Cancer Families Encode Proteins That Show Impaired Activation. *Cancer Research,* Vol. 66, No. 18**,** 8966-8970.

Solyom, S., Aressy, B., Pylkas, K., Patterson-Fortin, J., Hartikainen, J. M., Kallioniemi, A., Kauppila, S., Nikkila, J., Kosma, V. M., Mannermaa, A., Greenberg, R. A. & Winqvist, R. 2012. Breast cancer-associated Abraxas mutation disrupts nuclear localization and DNA damage response functions. *Sci Transl Med,* Vol. 4, No. 122**,** 122-123.

Song, L., Huang, W., Kang, J., Huang, Y., Ren, H. & Ding, K. 2017. Comparison of error correction algorithms for Ion Torrent PGM data: application to hepatitis B virus. *Scientific reports,* Vol. 7, No. 1**,** 8106.

Sood, A. K., Seftor, E. A., Fletcher, M. S., Gardner, L. M. G., Heidger, P. M., Buller, R. E., Seftor, R. E. B. & Hendrix, M. J. C. 2001. Molecular Determinants of Ovarian Cancer Plasticity. *The American Journal of Pathology,* Vol. 158, No. 4**,** 1279-1288.

Soule, H. D., Maloney, T. M., Wolman, S. R., Peterson, W. D., Jr., Brenz, R., Mcgrath, C. M., Russo, J., Pauley, R. J., Jones, R. F. & Brooks, S. C. 1990. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer Res,* Vol. 50, No. 18**,** 6075-6086.

Southey, M. C., Goldgar, D. E., Winqvist, R., Pylkas, K., Couch, F., Tischkowitz, M., Foulkes, W. D., Dennis, J., Michailidou, K., Van Rensburg, E. J., Heikkinen, T., Nevanlinna, H., Hopper, J. L., Dork, T., Claes, K. B., Reis-Filho, J., Teo, Z. L., Radice, P., Catucci, I., Peterlongo, P., Tsimiklis, H., Odefrey, F. A., Dowty, J. G., Schmidt, M. K., Broeks, A., Hogervorst, F. B., Verhoef, S., Carpenter, J., Clarke, C., Scott, R. J., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Peto, J., Dos-Santos-Silva, I., Fletcher, O., Johnson, N., Bolla, M. K., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Marme, F., Burwinkel, B., Yang, R., Guenel, P., Truong, T., Menegaux, F., Sanchez, M., Bojesen, S., Nielsen, S. F., Flyger, H., Benitez, J., Zamora, M. P., Perez, J. I., Menendez, P., Anton-Culver, H., Neuhausen, S., Ziogas, A., Clarke, C. A., Brenner, H., Arndt, V., Stegmaier, C., Brauch, H., Bruning, T., Ko, Y. D., Muranen, T. A., Aittomaki, K., Blomqvist, C., Bogdanova, N. V., Antonenkova, N. N., Lindblom, A., Margolin, S., Mannermaa, A., Kataja, V., Kosma, V. M., Hartikainen, J. M., Spurdle, A. B., Investigators, K., Wauters, E., Smeets, D., Beuselinck, B., Floris, G., Chang-Claude, J., Rudolph, A., Seibold, P., Flesch-Janys, D., Olson, J. E., Vachon, C., Pankratz, V. S., Mclean, C., Haiman, C. A., Henderson, B. E., Schumacher, F., Le Marchand, L., Kristensen, V., Alnaes, G. G., Zheng, W., Hunter, D. J*., et al.* 2016. PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. *J Med Genet,* Vol. 53, No. 12**,** 800-811.

Starink, T. M., Van Der Veen, J. P., Arwert, F., De Waal, L. P., De Lange, G. G., Gille, J. J. & Eriksson, A. W. 1986. The Cowden syndrome: a clinical and genetic study in 21 patients. *Clin Genet,* Vol. 29, No. 3**,** 222-233.

Starita, L. M. & Parvin, J. D. 2003. The multiple nuclear functions of BRCA1: transcription, ubiquitination and DNA repair. *Curr Opin Cell Biol,* Vol. 15, No. 3**,** 345-350.

Stenson, P. D., Ball, E. V., Howells, K., Phillips, A. D., Mort, M. & Cooper, D. N. 2009. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. *Human Genomics,* Vol. 4, No. 2**,** 69-72.

Stepanenko, A. A. & Dmitrenko, V. V. 2015. HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene,* Vol. 569, No. 2**,** 182-190.

Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. & Elledge, S. J. 2003. MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature,* Vol. 421, No. 6926**,** 961-966.

Stork, C. T., Bocek, M., Crossley, M. P., Sollier, J., Sanz, L. A., Chédin, F., Swigut, T. & Cimprich, K. A. 2016. Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *eLife,* Vol. 5, No.**,** e17548.

Strom, C. M., Rivera, S., Elzinga, C., Angeloni, T., Rosenthal, S. H., Goos-Root, D., Siaw, M., Platt, J., Braastadt, C., Cheng, L., Ross, D. & Sun, W. 2015. Development and Validation of a Next-Generation Sequencing Assay for BRCA1 and BRCA2 Variants for the Clinical Laboratory. *PloS one,* Vol. 10, No. 8**,** e0136419.

Sun, J., Wang, Y., Xia, Y., Xu, Y., Ouyang, T., Li, J., Wang, T., Fan, Z., Fan, T., Lin, B., Lou, H. & Xie, Y. 2015a. Mutations in RECQL Gene Are Associated with Predisposition to Breast Cancer. *PLoS Genet,* Vol. 11, No. 5**,** e1005228.

Sun, Y., Ruivenkamp, C. A., Hoffer, M. J., Vrijenhoek, T., Kriek, M., Van Asperen, C. J., Den Dunnen, J. T. & Santen, G. W. 2015b. Next-generation diagnostics: gene panel, exome, or whole genome? *Hum Mutat,* Vol. 36, No. 6**,** 648-655.

Suter, R. & Marcum, J. A. 2007. The molecular genetics of breast cancer and targeted therapy. *Biologics,* Vol. 1, No. 3**,** 241-258.

Tambini, C. E., George, A. M., Rommens, J. M., Tsui, L. C., Scherer, S. W. & Thacker, J. 1997. The XRCC2 DNA repair gene: identification of a positional candidate. *Genomics,* Vol. 41, No. 1**,** 84-92.

Tamimi, R. M., Baer, H. J., Marotti, J., Galan, M., Galaburda, L., Fu, Y., Deitz, A. C., Connolly, J. L., Schnitt, S. J., Colditz, G. A. & Collins, L. C. 2008. Comparison of molecular phenotypes of ductal carcinoma in situ and invasive breast cancer. *Breast Cancer Res,* Vol. 10, No. 4**,** R67.

Tang, J., Erikson, R. L. & Liu, X. 2006. Checkpoint kinase 1 (Chk1) is required for mitotic progression through negative regulation of polo-like kinase 1 (Plk1). *Proc Natl Acad Sci U S A,* Vol. 103, No. 32**,** 11964-11969.

Tarabeux, J., Zeitouni, B., Moncoutier, V., Tenreiro, H., Abidallah, K., Lair, S., Legoix-Ne, P., Leroy, Q., Rouleau, E., Golmard, L., Barillot, E., Stern, M. H., Rio-Frio, T., Stoppa-Lyonnet, D. & Houdayer, C. 2014. Streamlined ion torrent PGM-based diagnostics: BRCA1 and BRCA2 genes as a model. *Eur J Hum Genet,* Vol. 22, No. 4**,** 535-541.

Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E. & Thomas, A. 2008a. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Human Mutation,* Vol. 29, No. 11**,** 1342-1354.

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., De Silva, D., Zharkikh, A. & Thomas, A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet,* Vol. 43, No. 4**,** 295-305.

Tavtigian, S. V., Greenblatt, M. S., Goldgar, D. E., Boffetta, P. & For The, I. U. G. V. W. G. 2008b. Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group. *Human Mutation,* Vol. 29, No. 11**,** 1261-1264.

Tavtigian, S. V., Oefner, P. J., Babikyan, D., Hartmann, A., Healey, S., Le Calvez-Kelm, F., Lesueur, F., Byrnes, G. B., Chuang, S. C., Forey, N., Feuchtinger, C., Gioia, L., Hall, J., Hashibe, M., Herte, B., Mckay-Chopin, S., Thomas, A., Vallee, M. P., Voegele, C., Webb, P. M., Whiteman, D. C., Sangrajrang, S., Hopper, J. L., Southey, M. C., Andrulis, I. L., John, E. M. & Chenevix-Trench, G. 2009. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am J Hum Genet,* Vol. 85, No. 4**,** 427-446.

Tavtigian, S. V., Simard, J., Rommens, J., Couch, F., Shattuck-Eidens, D., Neuhausen, S., Merajver, S., Thorlacius, S., Offit, K., Stoppa-Lyonnet, D., Belanger, C., Bell, R., Berry, S., Bogden, R., Chen, Q., Davis, T., Dumont, M., Frye, C., Hattier, T., Jammulapati, S., Janecki, T., Jiang, P., Kehrer, R., Leblanc, J. F., Mitchell, J. T., Mcarthur-Morrison, J., Nguyen, K., Peng, Y., Samson, C., Schroeder, M., Snyder, S. C., Steele, L., Stringfellow, M., Stroup, C., Swedlund, B., Swense, J., Teng, D., Thomas, A., Tran, T., Tranchant, M., Weaver-Feldhaus, J., Wong, A. K. C., Shizuya, H., Eyfjord, J. E., Cannon-Albright, L., Tranchant, M., Labrie, F., Skolnick, M. H., Weber, B., Kamb, A. & Goldgar, D. E. 1996. The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genetics,* Vol. 12, No. 3**,** 333-337.

Ten Bosch, J. R. & Grody, W. W. 2008. Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *J Mol Diagn,* Vol. 10, No. 6**,** 484-492.

Teo, Z. L., Park, D. J., Provenzano, E., Chatfield, C. A., Odefrey, F. A., Nguyen-Dumont, T., Dowty, J. G., Hopper, J. L., Winship, I., Goldgar, D. E. & Southey, M. C. 2013. Prevalence of PALB2 mutations in Australasian multiple-case breast cancer families. *Breast Cancer Res,* Vol. 15, No. 1**,** R17.

Thangaraju, M., Kaufmann, S. H. & Couch, F. J. 2000. BRCA1 facilitates stress-induced apoptosis in breast and ovarian cancer cell lines. *J Biol Chem,* Vol. 275, No. 43**,** 33487-33496.

The Breast Cancer Linkage Consortium 1999. Cancer Risks in BRCA2 Mutation Carriers. *Journal of the National Cancer Institute,* Vol. 91, No. 15**,** 1310-1316.

The Chek2 Cancer Consortium 2004. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am J Hum Genet,* Vol. 74, No. 6**,** 1175-1182.

Thompson, B. A., Greenblatt, M. S., Vallee, M. P., Herkert, J. C., Tessereau, C., Young, E. L., Adzhubey, I. A., Li, B., Bell, R., Feng, B., Mooney, S. D., Radivojac, P., Sunyaev, S. R., Frebourg, T., Hofstra, R. M. W., Sijmons, R. H., Boucher, K., Thomas, A., Goldgar, D. E., Spurdle, A. B. & Tavtigian, S. V. 2013a. Calibration of Multiple In Silico Tools for Predicting Pathogenicity of Mismatch Repair Gene Missense Substitutions. *Human Mutation,* Vol. 34, No. 1**,** 255-265.

Thompson, D., Duedal, S., Kirner, J., Mcguffog, L., Last, J., Reiman, A., Byrd, P., Taylor, M. & Easton, D. F. 2005. Cancer risks and mortality in heterozygous ATM mutation carriers. *J Natl Cancer Inst,* Vol. 97, No. 11**,** 813-822.

Thompson, D. & Easton, D. 2002. Variation in BRCA1 cancer risks by mutation position. *Cancer Epidemiol Biomarkers Prev,* Vol. 11, No. 4**,** 329-336.

Thompson, E. A., Graham, E., Macneill, C. M., Young, M., Donati, G., Wailes, E. M., Jones, B. T. & Levi-Polyachenko, N. H. 2014. Differential response of MCF7, MDA-MB-231, and MCF 10A cells to hyperthermia, silver nanoparticles and silver nanoparticle-induced photothermal therapy. *International Journal of Hyperthermia,* Vol. 30, No. 5**,** 312-323.

Thompson, E. R., Rowley, S. M., Sawyer, S., Kconfab, Eccles, D. M., Trainer, A. H., Mitchell, G., James, P. A. & Campbell, I. G. 2013b. Analysis of RAD51D in ovarian cancer patients and families with a history of ovarian or breast cancer. *PLoS One,* Vol. 8, No. 1**,** e54772.

Thorstenson, Y. R., Roxas, A., Kroiss, R., Jenkins, M. A., Yu, K. M., Bachrich, T., Muhr, D., Wayne, T. L., Chu, G., Davis, R. W., Wagner, T. M. & Oefner, P. J. 2003. Contributions of ATM mutations to familial breast and ovarian cancer. *Cancer Res,* Vol. 63, No. 12**,** 3325-3333.

Toledo, F. & Wahl, G. M. 2006. Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nat Rev Cancer,* Vol. 6, No. 12**,** 909-923.

Tommiska, J., Seal, S., Renwick, A., Barfoot, R., Baskcomb, L., Jayatilake, H., Bartkova, J., Tallila, J., Kaare, M., Tamminen, A., Heikkila, P., Evans, D. G., Eccles, D., Aittomaki, K., Blomqvist, C., Bartek, J., Stratton, M. R., Nevanlinna, H. & Rahman, N. 2006. Evaluation of RAD50 in familial breast cancer predisposition. *Int J Cancer,* Vol. 118, No. 11**,** 2911-2916.

Tonin, P., Weber, B., Offit, K., Couch, F., Rebbeck, T. R., Neuhausen, S., Godwin, A. K., Daly, M., Wagner-Costalos, J., Berman, D., Grana, G., Fox, E., Kane, M. F., Kolodner, R. D., Krainer, M., Haber, D. A., Struewing, J. P., Warner, E., Rosen, B., Lerman, C., Peshkin, B., Norton, L., Serova, O., Foulkes, W. D., Garber, J. E. & Et Al. 1996. Frequency of recurrent BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer families. *Nat Med,* Vol. 2, No. 11**,** 1179-1183.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J. & Jemal, A. 2015. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians,* Vol. 65, No. 2**,** 87-108.

Trujillano, D., Weiss, M. E., Schneider, J., Koster, J., Papachristos, E. B., Saviouk, V., Zakharkina, T., Nahavandi, N., Kovacevic, L. & Rolfs, A. 2015. Next-generation sequencing of the BRCA1 and BRCA2 genes for the genetic diagnostics of hereditary breast and/or ovarian cancer. *J Mol Diagn,* Vol. 17, No. 2**,** 162-170.

Tucker, T., Marra, M. & Friedman, J. M. 2009. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *American Journal of Human Genetics,* Vol. 85, No. 2**,** 142-154.

Tung, N., Battelli, C., Allen, B., Kaldate, R., Bhatnagar, S., Bowles, K., Timms, K., Garber, J. E., Herold, C., Ellisen, L., Krejdovsky, J., Deleonardis, K., Sedgwick, K., Soltis, K., Roa, B., Wenstrup, R. J. & Hartman, A. R. 2015. Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel. *Cancer,* Vol. 121, No. 1**,** 25-33.

Turnbull, C. & Rahman, N. 2008. Genetic predisposition to breast cancer: past, present, and future. *Annu Rev Genomics Hum Genet,* Vol. 9, No.**,** 321-345.

Van Der Groep, P., Bouter, A., Van Der Zanden, R., Siccama, I., Menko, F. H., Gille, J. J., Van Kalken, C., Van Der Wall, E., Verheijen, R. H. & Van Diest, P. J. 2006. Distinction between hereditary and sporadic breast cancer on the basis of clinicopathological data. *J Clin Pathol,* Vol. 59, No. 6**,** 611-617.

Van Der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol,* Vol. 12, No. 7**,** 479-492.

Vaughn, J. P., Davis, P. L., Jarboe, M. D., Huper, G., Evans, A. C., Wiseman, R. W., Berchuck, A., Iglehart, J. D., Futreal, P. A. & Marks, J. R. 1996. BRCA1 expression is induced before DNA synthesis in both normal and tumor-derived breast cells. *Cell Growth Differ,* Vol. 7, No. 6**,** 711-715.

Venkitaraman, A. R. 2002. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell,* Vol. 108, No. 2**,** 171-182.

Venkitaraman, A. R. 2019. How do mutations affecting the breast cancer genes BRCA1 and BRCA2 cause cancer susceptibility? *DNA Repair,* Vol. 81, No.**,** 102668.

Vogel, U., Szczepanowski, R., Claus, H., Junemann, S., Prior, K. & Harmsen, D. 2012. Ion torrent personal genome machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information. *J Clin Microbiol,* Vol. 50, No. 6**,** 1889-1894.

Vogelstein, B. & Kinzler, K. W. 1994. Tumour-suppressor genes. X-rays strike p53 again. *Nature,* Vol. 370, No. 6486**,** 174-175.

Walker, L. C., Pearson, J. F., Wiggins, G. A., Giles, G. G., Hopper, J. L. & Southey, M. C. 2017. Increased genomic burden of germline copy number variants is associated with early onset breast cancer: Australian breast cancer family registry. *Breast Cancer Res,* Vol. 19, No. 1**,** 30.

Walsh, T., Casadei, S., Coats, K. H., Swisher, E., Stray, S. M., Higgins, J., Roach, K. C., Mandell, J., Lee, M. K., Ciernikova, S., Foretova, L., Soucek, P. & King, M.-C. 2006. Spectrum of Mutations in BRCA1, BRCA2, CHEK2, and TP53 in Families at High Risk of Breast Cancer. *JAMA,* Vol. 295, No. 12**,** 1379-1388.

Wang, B., Matsuoka, S., Ballif, B. A., Zhang, D., Smogorzewska, A., Gygi, S. P. & Elledge, S. J. 2007. Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. *Science,* Vol. 316, No. 5828**,** 1194-1198.

Wang, H., Russa, M. L. & Qi, L. S. 2016. CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry,* Vol. 85, No. 1**,** 227-264.

Wang, J. J., Fang, Z. X., Ye, H. M., You, P., Cai, M. J., Duan, H. B., Wang, F. & Zhang, Z. Y. 2013. Clinical significance of overexpressed cyclin-dependent kinase subunits 1 and 2 in esophageal carcinoma. *Dis Esophagus,* Vol. 26, No. 7**,** 729-736.

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H., Shi, S. T., Siu, H. C., Deng, S., Chu, K. M., Law, S., Chan, K. H., Chan, A. S., Tsui, W. Y., Ho, S. L., Chan, A. K., Man, J. L., Foglizzo, V., Ng, M. K., Chan, A. S., Ching, Y. P., Cheng, G. H., Xie, T., Fernandez, J., Li, V. S., Clevers, H., Rejto, P. A., Mao, M. & Leung, S. Y. 2014. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet,* Vol. 46, No. 6**,** 573-582.

Wang, W., Wei, Z., Lam, T.-W. & Wang, J. 2011. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Scientific Reports,* Vol. 1, No. 1**,** 55.

Wang, X., Wang, R. H., Li, W., Xu, X., Hollander, M. C., Fornace, A. J., Jr. & Deng, C. X. 2004. Genetic interactions between Brca1 and Gadd45a in centrosome duplication, genetic stability, and neural tube closure. *J Biol Chem,* Vol. 279, No. 28**,** 29606-29614.

Wang, Y., Cortez, D., Yazdi, P., Neff, N., Elledge, S. J. & Qin, J. 2000. BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev,* Vol. 14, No. 8**,** 927-939.

Wang, Y., Putnam, C. D., Kane, M. F., Zhang, W., Edelmann, L., Russell, R., Carrion, D. V., Chin, L., Kucherlapati, R., Kolodner, R. D. & Edelmann, W. 2005. Mutation in Rpa1 results in defective DNA double-strand break repair, chromosomal instability and cancer in mice. *Nat Genet,* Vol. 37, No. 7**,** 750-755.

Weitzel, J. N., Lagos, V. I., Herzog, J. S., Judkins, T., Hendrickson, B., Ho, J. S., Ricker, C. N., Lowstuter, K. J., Blazer, K. R., Tomlinson, G. & Scholl, T. 2007. Evidence for common ancestral origin of a recurring BRCA1 genomic rearrangement identified in high-risk Hispanic families. *Cancer Epidemiol Biomarkers Prev,* Vol. 16, No. 8**,** 1615-1620.

Welcsh, P. L. & King, M.-C. 2001. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Human Molecular Genetics,* Vol. 10, No. 7**,** 705-713.

Wellings, S. R. & Jensen, H. M. 1973. On the origin and progression of ductal carcinoma in the human breast. *J Natl Cancer Inst,* Vol. 50, No. 5**,** 1111-1118.

Wells, N. J., Watanabe, N., Tokusumi, T., Jiang, W., Verdecia, M. A. & Hunter, T. 1999. The C-terminal domain of the Cdc2 inhibitory kinase Myt1 interacts with Cdc2 complexes and is required for inhibition of G(2)/M progression. *J Cell Sci,* Vol. 112, No. 19**,** 3361-3371.

Wenger, S. L., Senft, J. R., Sargent, L. M., Bamezai, R., Bairwa, N. & Grant, S. G. 2004. Comparison of established cell lines at different passages by karyotype and comparative genomic hybridization. *Biosci Rep,* Vol. 24, No. 6**,** 631-639.

Wertz, I. E. & Dixit, V. M. 2010. Regulation of death receptor signaling by the ubiquitin system. *Cell Death Differ,* Vol. 17, No. 1**,** 14-24.

Whittemore, A. S., Gong, G., John, E. M., Mcguire, V., Li, F. P., Ostrow, K. L., Dicioccio, R., Felberg, A. & West, D. W. 2004. Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic Whites. *Cancer Epidemiol Biomarkers Prev,* Vol. 13, No. 12**,** 2078-2083.

Wiles, M. V., Qin, W., Cheng, A. W. & Wang, H. 2015. CRISPR-Cas9-mediated genome editing and guide RNA design. *Mamm Genome,* Vol. 26, No. 9-10**,** 501-510.

Winkler, E. C. & Wiemann, S. 2016. Findings made in gene panel to whole genome sequencing: data, knowledge, ethics – and consequences? *Expert Review of Molecular Diagnostics,* Vol. 16, No. 12**,** 1259-1270.

Winship, I. & Southey, M. C. 2016. Gene panel testing for hereditary breast cancer. *Med J Aust,* Vol. 204, No. 5**,** 188-190.

Winslow, T. 2011. *Anatomy of the Female Breast* [Online]. Terese Winslow Medical and Scientific Illustration. Available: https://www.teresewinslow.com/breast/5cfygq3a1tvi8s0ps7uo25fn9guuq5 [Accessed 23 May 2014].

Woerner, S. M., Benner, A., Sutter, C., Schiller, M., Yuan, Y. P., Keller, G., Bork, P., Doeberitz, M. V. K. & Gebert, J. F. 2003. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene,* Vol. 22, No. 15**,** 2226-2235.

Wong, K. Y., So, C. C., Loong, F., Chung, L. P., Lam, W. W., Liang, R., Li, G. K., Jin, D. Y. & Chim, C. S. 2011a. Epigenetic inactivation of the miR-124-1 in haematological malignancies. *PLoS One,* Vol. 6, No. 4**,** e19027.

Wong, M. W., Nordfors, C., Mossman, D., Pecenpetelovska, G., Avery-Kiejda, K. A., Talseth-Palmer, B., Bowden, N. A. & Scott, R. J. 2011b. BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. *Breast Cancer Res Treat,* Vol. 127, No. 3**,** 853-859.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. & Micklem, G. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature,* Vol. 378, No. 6559**,** 789-792.

Wooster, R. & Weber, B. L. 2003. Breast and ovarian cancer. *N Engl J Med,* Vol. 348, No. 23**,** 2339-2347.

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M. K., Shu, X.-O., Lu, Y., Cai, Q., Al-Ejeh, F., Rozali, E., Wang, Q., Dennis, J., Li, B., Zeng, C., Feng, H., Gusev, A., Barfield, R. T., Andrulis, I. L., Anton-Culver, H., Arndt, V., Aronson, K. J., Auer, P. L., Barrdahl, M., Baynes, C., Beckmann, M. W., Benitez, J., Bermisheva, M., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Brauch, H., Brenner, H., Brinton, L., Broberg, P., Brucker, S. Y., Burwinkel, B., Caldés, T., Canzian, F., Carter, B. D., Castelao, J. E., Chang-Claude, J., Chen, X., Cheng, T.-Y. D., Christiansen, H., Clarke, C. L., Collaborators, N., Collée, M., Cornelissen, S., Couch, F. J., Cox, D., Cox, A., Cross, S. S., Cunningham, J. M., Czene, K., Daly, M. B., Devilee, P., Doheny, K. F., Dörk, T., Dos-Santos-Silva, I., Dumont, M., Dwek, M., Eccles, D. M., Eilber, U., Eliassen, A. H., Engel, C., Eriksson, M., Fachal, L., Fasching, P. A., Figueroa, J., Flesch-Janys, D., Fletcher, O., Flyger, H., Fritschi, L., Gabrielson, M., Gago-Dominguez, M., Gapstur, S. M., García-Closas, M., Gaudet, M. M., Ghoussaini, M., Giles, G. G., Goldberg, M. S., Goldgar, D. E., González-Neira, A., Guénel, P., Hahnen, E., Haiman, C. A., Håkansson, N., Hall, P., Hallberg, E., Hamann, U., Harrington, P., Hein, A., Hicks, B., Hillemanns, P., Hollestelle, A., Hoover, R. N., Hopper, J. L., Huang, G.*, et al.* 2018. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics,* Vol. 50, No. 7**,** 968-978.

Wu, L., Timmers, C., Maiti, B., Saavedra, H. I., Sang, L., Chong, G. T., Nuckolls, F., Giangrande, P., Wright, F. A., Field, S. J., Greenberg, M. E., Orkin, S., Nevins, J. R., Robinson, M. L. & Leone, G. 2001. The E2F1-3 transcription factors are essential for cellular proliferation. *Nature,* Vol. 414, No. 6862**,** 457-462.

Wu, L. C., Wang, Z. W., Tsan, J. T., Spillman, M. A., Phung, A., Xu, X. L., Yang, M. C., Hwang, L. Y., Bowcock, A. M. & Baer, R. 1996. Identification of a RING protein that can interact in vivo with the BRCA1 gene product. *Nat Genet,* Vol. 14, No. 4**,** 430-440.

Xia, B., Sheng, Q., Nakanishi, K., Ohashi, A., Wu, J., Christ, N., Liu, X., Jasin, M., Couch, F. J. & Livingston, D. M. 2006. Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol Cell,* Vol. 22, No. 6**,** 719-729.

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,* Vol. 98, No. 15**,** 8644-8649.

Xiao, M., Shen, Y., Chen, L., Liao, Z. & Wen, F. 2014. The rs7003908 (T>G) polymorphism in the XRCC7 gene and the risk of cancers. *Mol Biol Rep,* Vol. 41, No. 6**,** 3577-3582.

Xu, X., Gao, D., Wang, P., Chen, J., Ruan, J., Xu, J. & Xia, X. 2018. Efficient homology-directed gene editing by CRISPR/Cas9 in human stem and primary cells using tube electroporation. *Scientific Reports,* Vol. 8, No. 1**,** 11649.

Xu, X., Wagner, K. U., Larson, D., Weaver, Z., Li, C., Ried, T., Hennighausen, L., Wynshaw-Boris, A. & Deng, C. X. 1999. Conditional mutation of Brca1 in mammary epithelial cells results in blunted ductal morphogenesis and tumour formation. *Nat Genet,* Vol. 22, No. 1**,** 37-43.

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., Stenson, P. D., Cooper, D. N., Tyler-Smith, C. & Genomes Project, C. 2012. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *American journal of human genetics,* Vol. 91, No. 6**,** 1022-1032.

Yan, J., Kim, Y.-S., Yang, X.-P., Li, L.-P., Liao, G., Xia, F. & Jetten, A. M. 2007a. The ubiquitin-interacting motif containing protein RAP80 interacts with BRCA1 and functions in DNA damage repair response. *Cancer research,* Vol. 67, No. 14**,** 6647-6656.

Yan, J., Yang, X. P., Kim, Y. S., Joo, J. H. & Jetten, A. M. 2007b. RAP80 interacts with the SUMO-conjugating enzyme UBC9 and is a novel target for sumoylation. *Biochem Biophys Res Commun,* Vol. 362, No. 1**,** 132-138.

Yan, Z., Kim, Y.-S. & Jetten, A. M. 2002. RAP80, a Novel Nuclear Protein That Interacts with the Retinoid-related Testis-associated Receptor. *Journal of Biological Chemistry,* Vol. 277, No. 35**,** 32379-32388.

Yang, H., Li, Q., Fan, J., Holloman, W. K. & Pavletich, N. P. 2005. The BRCA2 homologue Brh2 nucleates RAD51 filament formation at a dsDNA-ssDNA junction. *Nature,* Vol. 433, No. 7026**,** 653-657.

Yang, X. J., Ogryzko, V. V., Nishikawa, J., Howard, B. H. & Nakatani, Y. 1996. A p300/CBP-associated factor that competes with the adenoviral oncoprotein E1A. *Nature,* Vol. 382, No. 6589**,** 319-324.

Yao, R., Zhang, C., Yu, T., Li, N., Hu, X., Wang, X., Wang, J. & Shen, Y. 2017. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics,* Vol. 10, No. 1**,** 30.

Yeo, Z. X., Chan, M., Yap, Y. S., Ang, P., Rozen, S. & Lee, A. S. G. 2012. Improving Indel Detection Specificity of the Ion Torrent PGM Benchtop Sequencer. *PLOS ONE,* Vol. 7, No. 9**,** e45798.

Yeo, Z. X., Wong, J. C. L., Rozen, S. G. & Lee, A. S. G. 2014. Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes. *BMC Genomics,* Vol. 15, No. 1**,** 516.

Yin, Y., Stephen, C. W., Luciani, M. G. & Fahraeus, R. 2002. p53 Stability and activity is regulated by Mdm2-mediated induction of alternative p53 translation products. *Nat Cell Biol,* Vol. 4, No. 6**,** 462-467.

Ying, M. Z., Wang, J. J., Li, D. W., Yu, G., Wang, X., Pan, J., Chen, Y. & He, M. X. 2010. The p300/CBP associated factor is frequently downregulated in intestinal-type gastric carcinoma and constitutes a biomarker for clinical outcome. *Cancer Biol Ther,* Vol. 9, No. 4**,** 312-320.

Yohe, S. & Thyagarajan, B. 2017. Review of Clinical Next-Generation Sequencing. *Arch Pathol Lab Med,* Vol. 141, No. 11**,** 1544-1557.

Young, E. L., Feng, B. J., Stark, A. W., Damiola, F., Durand, G., Forey, N., Francy, T. C., Gammon, A., Kohlmann, W. K., Kaphingst, K. A., Mckay-Chopin, S., Nguyen-Dumont, T., Oliver, J., Paquette, A. M., Pertesi, M., Robinot, N., Rosenthal, J. S., Vallee, M., Voegele, C., Hopper, J. L., Southey, M. C., Andrulis, I. L., John, E. M., Hashibe, M., Gertz, J., Le Calvez-Kelm, F., Lesueur, F., Goldgar, D. E. & Tavtigian, S. V. 2016. Multigene testing of moderate-risk genes: be mindful of the missense. *Journal of Medical Genetics,* Vol. 53, No. 6**,** 366-376.

Yu, H., Cook, T. J. & Sinko, P. J. 1997. Evidence for diminished functional expression of intestinal transporters in Caco-2 cell monolayers at high passages. *Pharm Res,* Vol. 14, No. 6**,** 757-762.

Yu, K. D., Di, G. H., Yuan, W. T., Fan, L., Wu, J., Hu, Z., Shen, Z. Z., Zheng, Y., Huang, W. & Shao, Z. M. 2009. Functional polymorphisms, altered gene expression and genetic association link NRH:quinone oxidoreductase 2 to breast cancer with wild-type p53. *Hum Mol Genet,* Vol. 18, No. 13**,** 2502-2517.

Yuan, S. S., Hou, M. F., Hsieh, Y. C., Huang, C. Y., Lee, Y. C., Chen, Y. J. & Lo, S. 2012. Role of MRE11 in cell proliferation, tumor invasion, and DNA repair in breast cancer. *J Natl Cancer Inst,* Vol. 104, No. 19**,** 1485-1502.

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., Hellmann, M. D., Barron, D. A., Schram, A. M., Hameed, M., Dogan, S., Ross, D. S., Hechtman, J. F., Delair, D. F., Yao, J., Mandelker, D. L., Cheng, D. T., Chandramohan, R., Mohanty, A. S., Ptashkin, R. N., Jayakumaran, G., Prasad, M., Syed, M. H., Rema, A. B., Liu, Z. Y., Nafa, K., Borsu, L., Sadowska, J., Casanova, J., Bacares, R., Kiecka, I. J., Razumova, A., Son, J. B., Stewart, L., Baldi, T., Mullaney, K. A., Al-Ahmadie, H., Vakiani, E., Abeshouse, A. A., Penson, A. V., Jonsson, P., Camacho, N., Chang, M. T., Won, H. H., Gross, B. E., Kundra, R., Heins, Z. J., Chen, H. W., Phillips, S., Zhang, H., Wang, J., Ochoa, A., Wills, J., Eubank, M., Thomas, S. B., Gardos, S. M., Reales, D. N., Galle, J., Durany, R., Cambria, R., Abida, W., Cercek, A., Feldman, D. R., Gounder, M. M., Hakimi, A. A., Harding, J. J., Iyer, G., Janjigian, Y. Y., Jordan, E. J., Kelly, C. M., Lowery, M. A., Morris, L. G. T., Omuro, A. M., Raj, N., Razavi, P., Shoushtari, A. N., Shukla, N., Soumerai, T. E., Varghese, A. M., Yaeger, R., Coleman, J., Bochner, B., Riely, G. J., Saltz, L. B., Scher, H. I., Sabbatini, P. J., Robson, M. E., Klimstra, D. S., Taylor, B. S., Baselga, J., Schultz, N., Hyman, D. M., Arcila, M. E., Solit, D. B., Ladanyi, M. & Berger, M. F. 2017.

Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med,* Vol. 23, No. 6**,** 703-713.

Zhang, C., Konermann, S., Brideau, N. J., Lotfy, P., Wu, X., Novick, S. J., Strutzenberg, T., Griffin, P. R., Hsu, P. D. & Lyumkis, D. 2018a. Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. *Cell,* Vol. 175, No. 17**,** 212-223.

Zhang, G., Wang, J., Yang, J., Li, W., Deng, Y., Li, J., Huang, J., Hu, S. & Zhang, B. 2015a. Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. *BMC Genomics,* Vol. 16, No. 1**,** 581.

Zhang, G., Zeng, Y., Liu, Z. & Wei, W. 2013a. Significant association between Nijmegen breakage syndrome 1 657del5 polymorphism and breast cancer risk. *Tumour Biol,* Vol. 34, No. 5**,** 2753-2757.

Zhang, H., Zhang, X., Ji, S., Hao, C., Mu, Y., Sun, J. & Hao, J. 2014. Sohlh2 inhibits ovarian cancer cell proliferation by upregulation of p21 and downregulation of cyclin D1. *Carcinogenesis,* Vol. 35, No. 8**,** 1863-1871.

Zhang, J., Willers, H., Feng, Z., Ghosh, J. C., Kim, S., Weaver, D. T., Chung, J. H., Powell, S. N. & Xia, F. 2004. Chk2 phosphorylation of BRCA1 regulates DNA double-strand break repair. *Mol Cell Biol,* Vol. 24, No. 2**,** 708-718.

Zhang, J., Wu, X. H. & Gan, Y. 2013b. Current evidence on the relationship between three polymorphisms in the XRCC7 gene and cancer risk. *Mol Biol Rep,* Vol. 40, No. 1**,** 81-86.

Zhang, K., Shi, L., Zhao, T., Wang, X., Zhang, Y., Chen, Y., Li, J. & Sun, Z. 2018b. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Research,* Vol. 46, No. 15**,** 7793-7804.

Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. 2015b. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Molecular Therapy - Nucleic Acids,* Vol. 4, No. 1**,** e264.

Zhang, X., Chiang, H.-C., Wang, Y., Zhang, C., Smith, S., Zhao, X., Nair, S. J., Michalek, J., Jatoi, I., Lautner, M., Oliver, B., Wang, H., Petit, A., Soler, T., Brunet, J., Mateo, F., Angel Pujana, M., Poggi, E., Chaldekas, K., Isaacs, C., Peshkin, B. N., Ochoa, O., Chedin, F., Theoharis, C., Sun, L.-Z., Curiel, T. J., Elledge, R., Jin, V. X., Hu, Y. & Li, R. 2017. Attenuation of RNA polymerase II pausing mitigates BRCA1-associated R-loop accumulation and tumorigenesis. *Nature Communications,* Vol. 8, No. 1**,** 15908.

Zhao, H., Watkins, J. L. & Piwnica-Worms, H. 2002. Disruption of the checkpoint kinase 1/cell division cycle 25A pathway abrogates ionizing radiation-induced S and G2 checkpoints. *Proc Natl Acad Sci U S A,* Vol. 99, No. 23**,** 14795-14800.

Zheng, X., Gai, X., Ding, F., Lu, Z., Tu, K., Yao, Y. & Liu, Q. 2013. Histone acetyltransferase PCAF up-regulated cell apoptosis in hepatocellular carcinoma via acetylating histone H4 and inactivating AKT signaling. *Mol Cancer,* Vol. 12, No. 1**,** 96.

Zhong, Q., Chen, C. F., Li, S., Chen, Y., Wang, C. C., Xiao, J., Chen, P. L., Sharp, Z. D. & Lee, W. H. 1999. Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. *Science,* Vol. 285, No. 5428**,** 747-750.

Zhou, L. P., Luan, H., Dong, X. H., Jin, G. J., Man, D. L. & Shang, H. 2012. Association between XRCC5, 6 and 7 gene polymorphisms and the risk of breast cancer: a HuGE review and meta-analysis. *Asian Pac J Cancer Prev,* Vol. 13, No. 8**,** 3637-3643.

Appendices

Appendix A: Pathways selected for AmpliSeq panel design

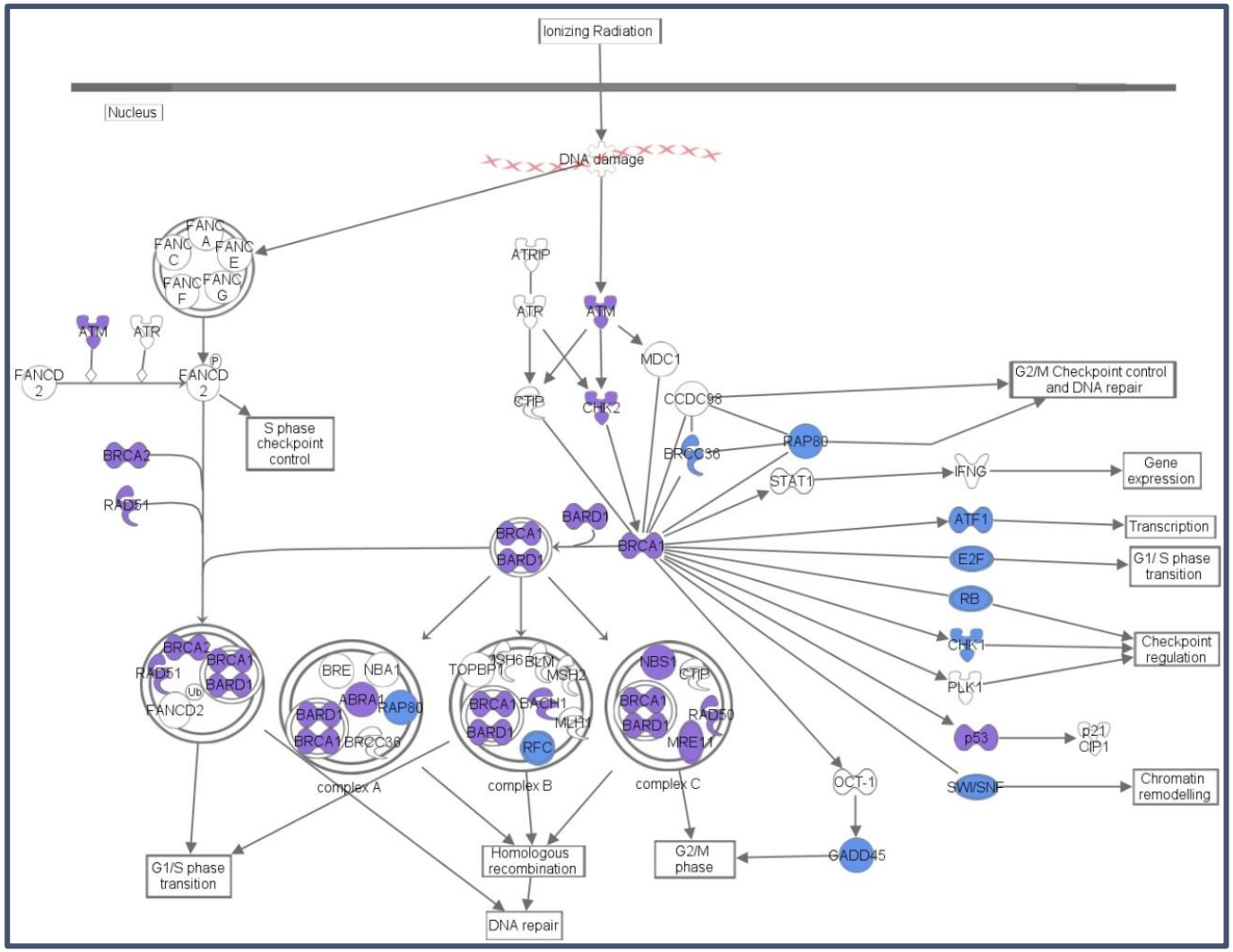


Figure A.1: Diagrammatic representation of the pathway depicting the role of BRCA1 in DNA damage response. Proteins highlighted in purple indicate known breast cancer predisposition genes which have been included on the diagnostic panel. Proteins shown in blue highlight potential novel breast cancer susceptibility genes which have been included on the discovery panel. Pathway generated using Ingenuity Pathway Analysis software.

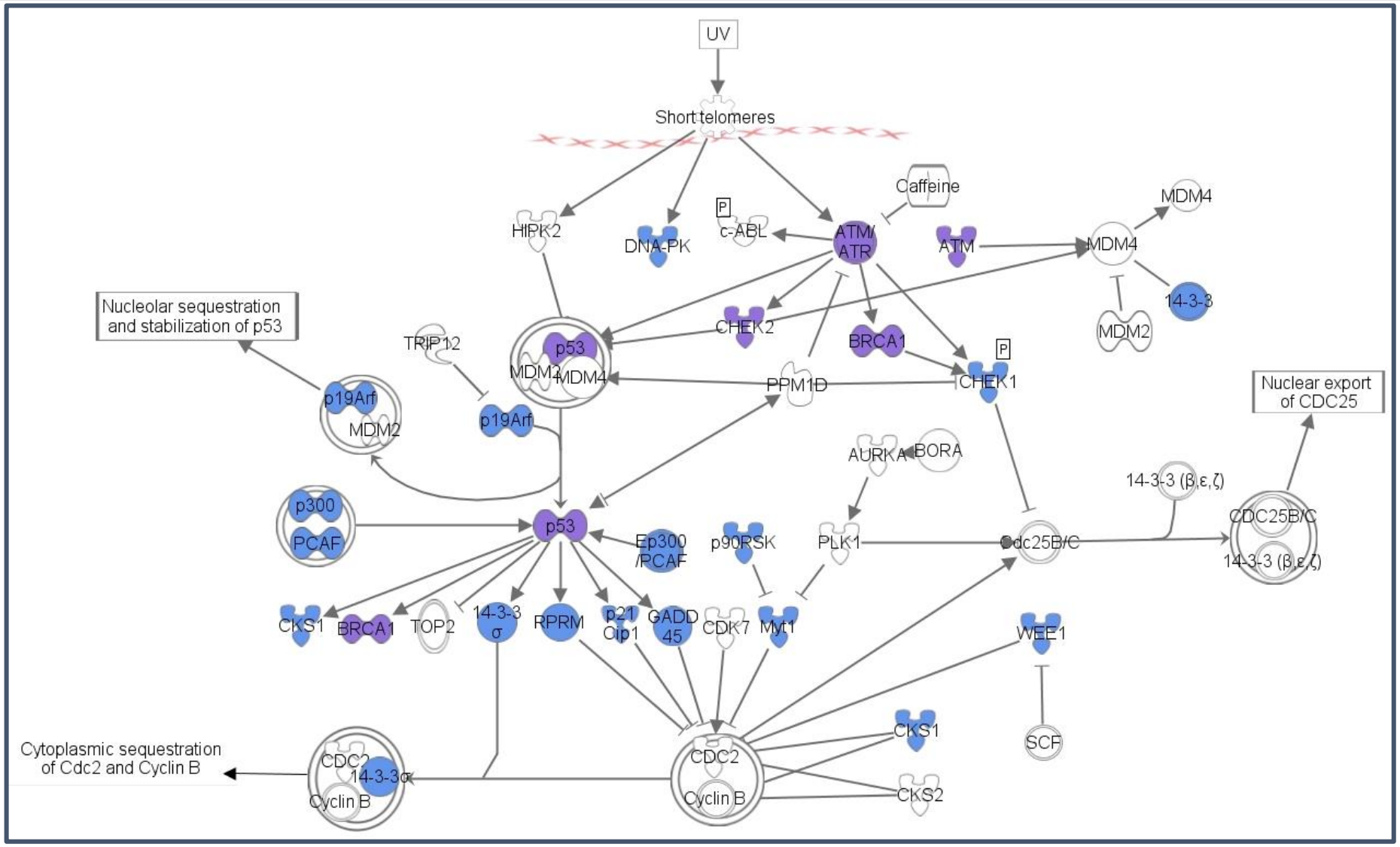


Figure A.2: Diagrammatic representation of G2/M checkpoint control pathway. Proteins highlighted in purple indicate known breast cancer predisposition genes which have been included on the diagnostic panel. Proteins shown in blue highlight potential novel breast cancer susceptibility genes which have been included on the discovery panel. Pathway generated using Ingenuity Pathway Analysis software.

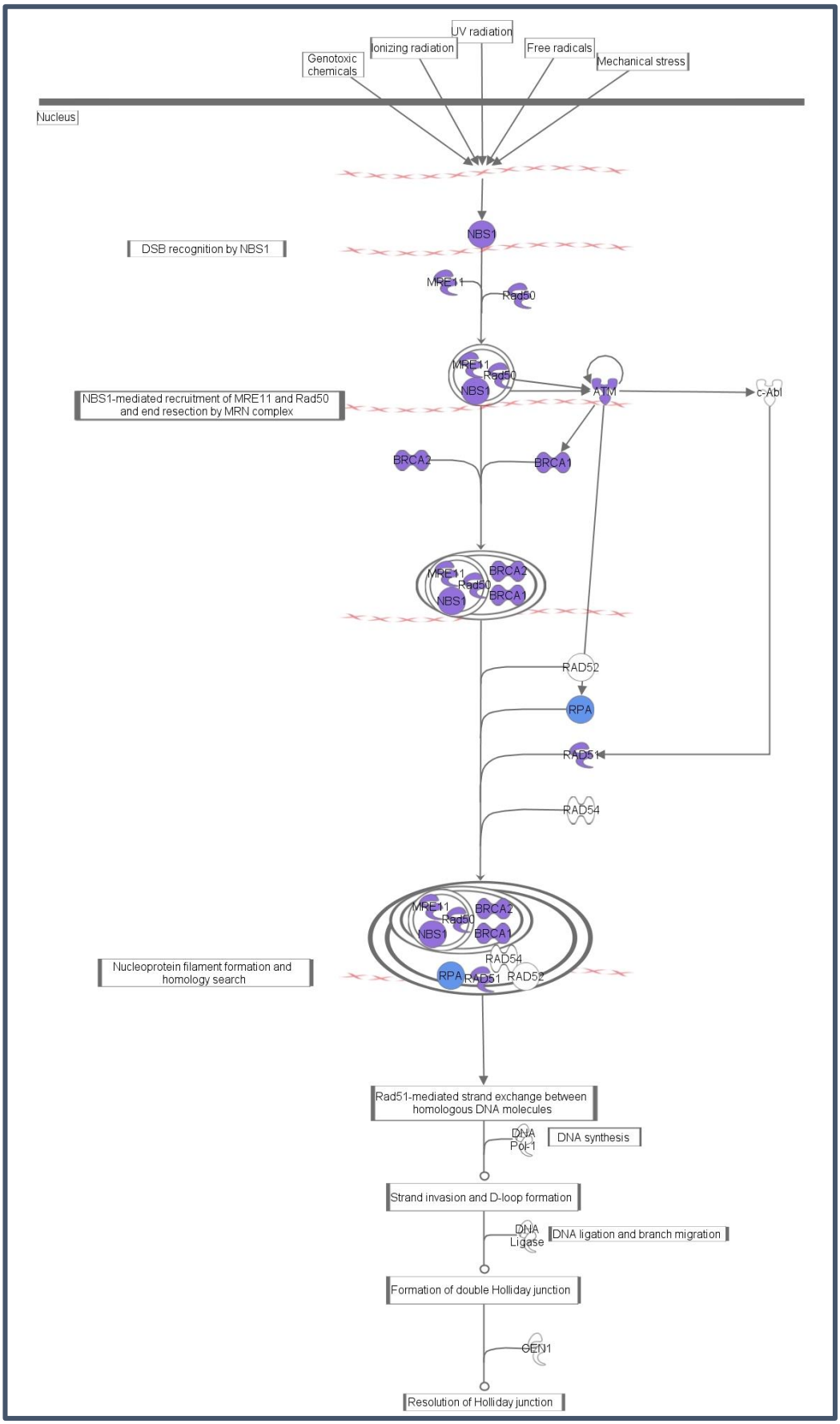


Figure A.3: Diagrammatic representation the homologous recombination pathway. Proteins highlighted in purple indicate known breast cancer predisposition genes which have been included on the diagnostic panel. Proteins shown in blue highlight potential novel breast cancer susceptibility genes which have been included on the discovery panel. Pathway generated using Ingenuity Pathway Analysis software.

Appendix B: Ion AmpliSeq™ Coverage Statistics

Table B.1: Ion AmpliSeq multiplexed primer statistics Coverage statistics including total number of bases, missed bases and overall coverage obtained for each of the genes included within the multiplex PCR.

Gene	Chromosome	Amplicons	Total Bases	Covered Bases	Missed Bases	Overall Coverage	Number of Exons
CKS1B	1	4	393	393	0	1.000	3
E2F2	1	14	1671	1671	0	1.000	7
GADD45A	1	9	790	790	0	1.000	5
RPS6KA1	1	35	3516	3463	53	0.985	23
SFN	1	6	798	798	0	1.000	1
BARD1	2	29	2895	2872	23	0.992	11
E2F6	2	13	1203	1202	1	0.999	7
RPRM	2	3	381	381	0	1.000	1
HLTF	3	51	4305	3989	316	0.927	25
KAT2B	3	35	3417	3158	259	0.924	18
RFC4	3	21	1602	1567	35	0.978	10
FAM175A	4	19	1689	1586	103	0.939	9
HMMR	5	33	3255	2897	358	0.890	20
RAD50	5	56	5214	4864	350	0.933	25
UIMC1	5	28	2874	2746	128	0.955	14
CDKN1A	6	4	597	597	0	1.000	2
E2F3	6	16	2077	1948	129	0.938	9
NQO2	6	12	1002	976	26	0.974	6
RFC2	7	17	1626	1551	75	0.954	11
XRCC2	7	8	996	995	1	0.999	3
E2F5	8	15	1619	1291	328	0.797	10
NBN	8	35	3081	3081	0	1.000	16
PRKDC	8	173	16772	16345	427	0.975	86
CDKN2A	9	12	1625	1625	0	1.000	7
ATM	11	127	12333	11526	807	0.935	62
CHEK1	11	26	2043	1952	91	0.955	12
MRE11A	11	36	3096	3045	51	0.984	19
WEE1	11	23	2693	2466	227	0.916	12
ATF1	12	11	1122	981	141	0.874	6
RFC5	12	15	1820	1740	80	0.956	14
BRCA2	13	113	12763	11714	1049	0.918	27
RFC3	13	18	1620	1561	59	0.964	10
RAD51	15	16	1771	1599	172	0.903	11
CDH1	16	32	3465	3465	0	1.000	16
E2F4	16	17	1752	1752	0	1.000	10
PALB2	16	39	4224	4148	76	0.982	13
PKMYT1	16	17	2143	2136	7	0.997	10
RBL2	16	42	4542	4337	205	0.955	22
BRCA1	17	71	9066	8636	430	0.953	24
BRIP1	17	48	4719	4528	191	0.960	19
RAD51C	17	17	1904	1803	101	0.947	10
RAD51D	17	16	1727	1611	116	0.933	11
RPA1	17	26	2718	2714	4	0.999	17
SMARCD2	17	19	2259	2008	251	0.889	13
TP53	17	17	2089	1846	243	0.884	13
E2F1	20	12	1671	1268	403	0.759	7
RBL1	20	45	4554	4170	384	0.916	23
SLC19A1	21	20	2379	2272	107	0.955	7
CHEK2	22	23	2526	2225	301	0.881	15
EP300	22	82	8826	8679	147	0.983	31
BRCC3	X	17	1621	1470	151	0.907	12

Appendix C: Manchester Scores

Table C.1: Manchester Scores of all individuals included in patient cohort.

Patient ID	Manchester Score	Patient ID	Manchester Score	Patient ID	Manchester Score	Patient ID	Manchester Score
SABC001	34	SABC035	28	SABC068	12	SABC101	21
SABC002	53	SABC036	19	SABC069	17	SABC102	13
SABC003	37	SABC037	19	SABC070	54	SABC103	20
SABC004	36	SABC038	8	SABC071	17	SABC104	25
SABC005	52	SABC039	33	SABC072	30	SABC105	18
SABC006	33	SABC040	8	SABC073	14	SABC106	10
SABC007	36	SABC041	25	SABC074	9	SABC107	34
SABC008	20	SABC042	34	SABC075	10	SABC108	21
SABC009	39	SABC043	29	SABC076	25	SABC109	26
SABC010	20	SABC044	26	SABC077	10	SABC110	26
SABC011	15	SABC045	41	SABC078	10	SABC111	22
SABC012	14	SABC046	12	SABC079	13	SABC112	11
SABC013	23	SABC047	18	SABC080	16	SABC113	21
SABC014	27	SABC048	16	SABC081	12	SABC114	13
SABC015	10	SABC049	10	SABC082	29	SABC115	36
SABC016	26	SABC050	13	SABC083	18	SABC116	30
SABC017	28	SABC051	10	SABC084	9	SABC117	21
SABC018	19	SABC052	16	SABC085	10	SABC118	27
SABC019	8	SABC053	9	SABC086	16	SABC119	17
SABC020	21	SABC054	28	SABC087	Unknown	SABC120	15
SABC021	25	SABC055	13	SABC088	12	SABC121	14
SABC022	34	SABC056	17	SABC089	14	SABC122	16
SABC023	22	SABC057	17	SABC090	14	SABC123	28
SABC024	29	SABC058	24	SABC091	29	SABC124	33
SABC025	15	SABC059	10	SABC092	13	SABC125	18
SABC026	22	SABC060	8	SABC093	26	SABC126	12
SABC027	34	SABC061	12	SABC094	16	SABC127	16
SABC028	24	SABC062	18	SABC095	32	SABC128	30
SABC029	18	SABC063	18	SABC096	61	SABC129	29
SABC030	31	SABC064	17	SABC097	10	SABC130	14
SABC031	29	SABC065	15	SABC098	14	SABC131	15
SABC032	7	SABC066	17	SABC099	16	SABC132	19
SABC033	24	SABC067	19	SABC100	16	SABC133	5
SABC034	22						

Appendix D: MPS run summaries

All raw data generated and analysed as part of this thesis is de-identified and publicly available (in addition to the required BED files) at https://doi.org/10.25957/5d916a7a7b101

Individual sequencing runs

Table D.1: ISP Sequencing Summary from Run 1. Run 1 contained 3 patient samples run on an Ion 318 chip.

Addressable Wells: 11,304,277		
	Number of Wells	**% of wells**
With ISPs	7,040,590	62.3
Live[1]	6,975,792	99.0
Test Fragment [2]	49,625	0.07
Library	6,923,167	99.3
Library ISPs: 6,923,167		
Filtered: Polyclonal[3]	1,094,275	15.8
Filtered: Low Quality	1,356,441	19.6
Filtered: Adapter Dimer	452	<1
Final Library ISPs	4,471,999	64.6

[1]LiveISPs identifies ISPs currently being sequenced
[2] Test Fragments are added to the sample ISP mix to allow the assessment of the run quality
[3] Polyclonal ISPs are ISPs with more than one library template attached.

Table D.2: ISP Sequencing Summary from Run 2. Run 2 contained 10 patient samples run on an Ion 318 chip.

Addressable Wells: 11,303,878		
	Number of Wells	**% of wells**
With ISPs	9,281,500	82.1
Live	9,281,325	100
Test Fragment	56,048	0.06
Library	9,225,277	99.4
Library ISPs: 9,225,277		
Filtered: Polyclonal	2,399,182	26.0
Filtered: Low Quality	1,124,926	12.2
Filtered: Adapter Dimer	2,915	<1
Final Library ISPs	5,698,254	61.8

Table D.3: ISP Sequencing Summary from Run 3. Run 3 contained 29 patient samples run on an Ion 318 chip

Addressable Wells: 11,287,275		
	Number of Wells	**% of wells**
With ISPs	8,436,034	74.7
Live	8,409,627	99.7
Test Fragment	15,228	0.2
Library	8,394,399	99.8
Library ISPs: 8,394,399		
Filtered: Polyclonal	3,697,537	44.0
Filtered: Low Quality	479,353	5.7
Filtered: Adapter Dimer	7,346	0.1
Final Library ISPs	4,210,163	50.2

Table D.4: ISP Sequencing Summary from Run 4. Run 4 contained 29 patient samples run on an Ion 318 chip

Addressable Wells: 11,287,275		
	Number of Wells	**% of wells**
With ISPs	9,460,569	83.8
Live	9,455,545	99.9
Test Fragment	10,014	0.1
Library	9,445,591	99.9
Library ISPs: 9,445,591		
Filtered: Polyclonal	6,379,504	67.5
Filtered: Low Quality	778,679	8.2
Filtered: Adapter Dimer	2,272	0.0
Final Library ISPs	2,285,076	24.2

Majority of the samples sequenced on this run did not achieve coverage greater than 50 X due to the high level of polyclonality (67%), with only 24% usable reads. Consultation with the ThermoFisher application specialist advised to decrease the amount of template loaded onto the 318v2 sequencing chip in an attempt to prevent the high level of polyclonality observed due to excess template.

Table D.5: ISP Sequencing Summary from Reattempt of MPS sequencing for Run 4. Run 4 contained 29 patient samples run on an Ion 318 chip.

Addressable Wells: 11,287,275		
	Number of Wells	**% of wells**
With ISPs	9,464,989	83.9
Live	9,462,149	100
Test Fragment	251	0
Library	9,461,898	100
Library ISPs: 9,461,898		
Filtered: Polyclonal	4,921,559	52
Filtered: Low Quality	598,601	6.3
Filtered: Adapter Dimer	5,874	0.1
Final Library ISPs	3,935,864	41.6

Polyclonality within this sequencing run was still high. Six sequenced samples failed to have a mean read depth greater than 50X coverage. This was the level of coverage that had been deemed appropriate for the analysis of germline mutations within our patent cohort. However, it was possible to combine the sequencing analysis from both attempts of sequencing these patient samples for a sufficient level of coverage for all 29 patients sequenced on this run.

Table D.6: ISP Sequencing Summary from Run 5. Run 5 contained 30 patient samples run on an Ion 318 chip

Addressable Wells: 11,287,275		
	Number of Wells	**% of wells**
With ISPs	7,529,355	66.7
Live	7,506,035	99.7
Test Fragment	124,675	1.7
Library	7,381,360	98.3
Library ISPs: 7,381,360		
Filtered: Polyclonal	3,186,085	43.2
Filtered: Low Quality	699,142	9.5
Filtered: Adapter Dimer	5,844	0.1
Final Library ISPs	3,490,289	47.3

Table D.7: ISP Sequencing Summary from Run 6. Run 6 contained 30 patient samples run on an Ion P1 Chip.

Addressable Wells: 61,764,148		
	Number of Wells	**% of wells**
With ISPs	140,878,841	93.0
Live	140,794,880	99.9
Test Fragment	693,118	0.5
Library	140,101,762	99.5
Library ISPs: 140,101,762		
Filtered: Polyclonal	65,139,623	46.5
Filtered: Low Quality	11,547,078	8.2
Filtered: Adapter Dimer	1,650,913	1.2
Final Library ISPs	61,764,148	44.1

Pooled DNA sequencing runs

Table D.8: ISP Sequencing Summary from pilot Tri-Pool-Seq MPS Run 1. Run contained 3 patient pools, each containing equimolar concentrations of 25 patient samples run on an Ion318 chip.

Addressable Wells: 11,287,275		
	Number of Wells	**% of wells**
With ISPs	8,436,034	74.7
Live	8,409,627	99.7
Test Fragment	15,228	0.2
Library	8,394,399	99.8
Library ISPs: 8,394,399		
Filtered: Polyclonal	3,697,537	44.0
Filtered: Low Quality	479,353	5.7
Filtered: Primer Dimer	7,346	0.1
Final Library ISPs	4,210,163	50.2

Table D.9: ISP Sequencing Summary from Tri-Pool-Seq MPS Run 2. Run contained 2 individually sequenced patients and 5 patient pools, each containing equimolar concentrations of 25 patient samples run on an Ion P1 chip. Tri-Pool-Seq

Addressable Wells: 151,539,288		
	Number of Wells	**% of wells**
With ISPs	139,181,239	91.8
Live	139,076,130	99.9
Test Fragment	1,364,942	1.0
Library	137,711,188	99.0
Library ISPs: 137,711,188		
Filtered: Polyclonal	36,892,448	26.8
Filtered: Low Quality	11,152,469	8.1
Filtered: Adapter Dimer	398,559	0.3
Final Library ISPs	89,267,712	64.8

Appendix E: Tri-Pool-Seq data analysis

Table E.1: Analysis of variants identified through Tri-Pool-Seq methodology. Initial analysis included all variants identified in pools. Analysis then narrowed down to rare variants (MAF<5%, as determined by gnomAD). True variants; variants detected by all three pools and present in individually sequenced patient, False positives; variants identified through pooling analysis that were not present in individually sequenced patient data. False negatives; variants identified in individual sequencing analysis but missed through Tri-Pool-Seq methodology.

Patient ID	Pools	Total number of true variants	False Positives				False Negatives			Total number of variants identified in through pooled analysis	Total number of variants identified in patient sample	Rare variants			
			Present in all 3 pools	Present in 2 pools only	Present in 1 pool only	Total	Present in patient only	Present in patient and 1 or 2 pools only	Total			Total number of rare variants in patient sample	Variants in all pools and patient	Variant in patient and 1 or 2 pools only	Variant in patient only
SABC007	2, 8, 15	65	78	34	34	146	74	14	88	299	153	93	13	7	73
SABC013	3, 8, 15	86	57	28	51	136	22	9	31	253	117	52	26	6	20
SABC025	2, 7, 15	82	69	28	37	134	34	12	46	262	128	57	18	7	32
SABC031	1, 7, 11	82	52	54	72	178	37	18	55	315	137	74	25	16	33
SABC042	1, 5, 11	107	49	27	70	146	55	25	80	333	187	103	30	18	55
SABC050	1, 8, 11	82	57	41	56	154	21	26	47	283	129	52	14	18	20
SABC059	3, 6, 11	88	52	51	78	181	25	13	38	307	126	54	22	8	24
SABC064	2, 8, 11	95	44	23	70	137	19	26	45	277	140	49	20	11	18
SABC065	2, 7, 11	112	28	36	64	128	24	25	49	289	161	70	30	16	24
SABC070	3, 7, 15	83	66	28	46	140	27	10	37	260	120	50	19	8	23
SABC071	2, 6, 15	81	61	33	62	156	39	25	64	301	145	73	22	14	38
SABC077	3, 6, 15	97	43	49	78	170	19	14	33	300	130	55	25	11	19
SABC085	1, 7, 15	69	67	28	83	178	28	19	47	294	116	58	20	10	28
SABC098	1, 6, 15	90	52	26	83	161	24	26	50	301	140	71	25	22	24
SABC102	3, 7, 11	101	37	43	69	149	30	19	49	299	150	64	23	14	27
SABC114	3, 8, 11	84	52	41	63	156	21	19	40	280	124	51	19	13	19
SABC127	1, 8, 15	98	40	27	74	141	21	18	39	278	137	61	29	11	21
SABC131	2, 6, 11	92	50	43	67	160	22	26	48	300	140	60	16	22	22

Appendix F: Primer sequences

Table F.1:Primer sequences for variant confirmation identified from MPS individual runs. Sequence variants listed in HGVS nomenclature. Forward and reverse primers listed in 5' – 3' direction. Optimised cycling conditions for each amplicon listed in **Appendix G.**

Amplicon Name	Variant confirmation	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Size	Cycling conditions
ATF1-571	ATF1:c.571C>G	ATA CTT GTG CCC AGC AAT CAG	CAT GAC TGT ATA CAG TTC CAG	372bp	Standard 58
ATM-2	ATM:c.2T>C	AGA ATG TGC CTC TAA TTG TAC AG	CAG GAT CTC GAA TCA GGC GCT	366bp	Standard 56
ATM-1010	ATM:c.1010G>A	AGC TAG CAG TGT AAA CAG AG	GTT GCA TGT ACA GAG TCA T	349bp	Standard 58
ATM-2084	ATM:c.2084C>T	AGA CAT GCT CAA GTT CTT GTG	CTA GTC TCA GGT TCA TTT CTC	281bp	Standard 58
ATM-2119	ATM:c.2119T>C	AAC CAT TGT GAG AGA ATG TGG	TAA AAT GAA GCC TCC CAC CA	250bp	Touchdown 60-58
ATM-2572	ATM:c.2572T>C	CTA CAG CAT GCT CCT GCA AG	GAG GCC TCT TAT ACT GCC A	298bp	Touchdown 63-61
ATM-3161	ATM:c.3161C>G	CTC TGT AAG AAT GGC CCT AG	GAC ATT CTA CTG CCA TCT GCA	347bp	Touchdown 62-60
BARD1-1670	BARD1:c.1670G>C	CTG CCT AAT ATG AGT TCT GAG	GGT CGT ACT GTG ATT ATG TC	402bp	Standard 58
BARD1-1993	BARD1:c.1993T>C	TGC CAT GAA GAA GAA AAA CCA	TGT CAT AAT AAG AAC AAT GAA AGT TGT	246bp	Standard 60
BRCA1*1086	BRCA1:c.*1086A>C	AAA AAG GAA AAT GAA ACT AGA AGA GA	CCT TCC AAC AGC TAT AAA CAG TCC	396bp	Touchdown 60-58
BRCA1*1332	BRCA1:c.*1332G>A	ATG ACA GAT CCC ACC AGG AAG	CTC AGA CTT CTG ACC TTG C	478bp	Standard 58
BRCA2	BRCA2:c.8149G>T	TGA TAG AAG CAG AAG ATC GGC	AAC TGT CAG TCT GCC ATT CTT T	250bp	Touchdown 60-58
CDH1-1004	CDH1:c.1004G>A	TAA GCA GTA TTG ACC CAG TCC	CAT GTG CTT CAT GCC AGC CTG	425bp	Standard 60
CDH1-1774	CDH1:c.1774G>A	GGA GTG TGT TCT TGG TGT GAG	GGT GAC ATC TAG AAG TTG AGA	592bp	Standard 58
CDKN2A-442	CDKN2A:c.442G>A	GCA AGT CCA TTT CGG GAT TA	CTT CCT GGA CAC GCT GGT	394bp	Touchdown 62-60
CHEK2-254	CHEK2:254C>T	TAC CAG CAC GAT GCC AAA CTC	CTG GAC AAC TCC AAT CAG AAC CT	290bp	Standard 58
E2F3-838	E2F3:c.838T>A	ATC CTT TGT GCC GCC AGT TCT C	GAC GAA TCT GCT TAG TGT TGT CA	379bp	Touchdown 63-61
EP300-6508	EP300:c.6508A>G	GTA TGC CAA CTC TAA TCC AC	CTG GCC TAT CTG TCC CAT AT	460bp	Touchdown 60-58
FAM175A-1117	FAM175A:c.1117G>A	TCA TCT GTT TCT GGG CTG CT	AAC ACA CTG ACA TTC CTG AAG C	193bp	Standard 60
HLTF-932	HLTF:c.932A>G	GAC ACT TCT CAG AGA TAC TTG	GAC CAT TCA TCG CAT TTC TG	454bp	Standard 58
HLTF-2189	HLTF:c.2189G>A	TTC CTA GCT GAG TCT CAC AC	CTC TGC TGT TTT AGA ATA GGA	339bp	Touchdown 62-60
HLTF-2400	HLTF: c. 2411C>T, HLTF:c. 2456A>G	AAT GAC AGG AAT GAA AGC AG	TGC TTC TAG CTA GTC CAG ATC	467bp	Standard 58
HMMR-300	HMMR:c.274C>T, HMMRc.383G>C	GAC TTT TAA GAT GTA TCA TAG G	CAT GAG GCT CAG ATA CCT TAG	387bp	Standard 58

Amplicon Name	Variant confirmation	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Size	Cycling conditions
HMMR-998	HMMR:c.998G>A	AGA AGA CCA TGT CAA CAG GAA	TTC TCT TGT TGC AGA AGT GAA TC	150bp	Standard 58
HMMR-1783	HMMR:c.1783C>A	TCT CAT AGA GAA TCT ATG GAG	TTA GAA TAA CCC CCA CTC CAA	587bp	Touchdown 58-55
MRE11A-1475	MRE11A:c.1475C>A	GTA TCT TAC AGA ATG TGC AGC	GTA CCA ACC ATA TGC AAG AC	386bp	Standard 58
NQO2-173	NQO2:c.173G>A	CCC TCC CAG ACT GCT TCT CT	CAC TTC CAG AAG CAG CAC AA	399bp	Touchdown 60-58
PALB2-1010	PALB2:c.1010T>C	CAG CAC CTT GAA CAC ATT CC	GAG AGG TTG CTT CCA GGC TA	385bp	Touchdown 62-60
PALB2-3116	PALB2:c.3116delA	CCC ACA GTT CTA CTT TTA CCT AAA TC	CGG GGA AGG TTT GTT CAT TA	349bp	Touchdown 60-58
PRKDC-999	PRKDC:c.999G>A	TGT ATC CTT GGT CTC TTA GTC	GTG CAG TTC ACG CCC ATG TTA	528bp	Touchdown 60-58
PRKDC-2083	PRKDC:c.2083C>T	CTG TGC CTT GAA ATG TCT G	CAG TTC ATG ATA ACA CTA AG	290bp	Touchdown 60-58
PRKDC-3730	PRKD:c.3730T>C	GTG ACT CTG GCA GCA CGA C	AAA AGA TTG TCC CAT AAC ATT TTG A	390bp	Touchdown 65-63
PRKDC-5120	PRKDC:c.5120T>A	CAT GGT GCT GAT TCA GTA TGT G	GTG AAG CCA TTC ATT GAT TCC A	508bp	Standard 56
PRKDC-8659	PRKDC:c.8659C>T	CAT GGT CCA AGT ACC ACA AGC	GCC AAA TGA AGC CAA GTG TT	337bp	Standard 58
PRKDC-9337	PRKDC:c.9337-2A>T	TGA AGT TGA TTA GCA CTC TTG AGG	TGG AGT TTC CAA CCC ATA CA	592bp	Standard 58
PRKDC-9445	PRKDC:c. 9445G>A	ATG AAT CTC TTT GCT GAG ACC	GTC CAG ACA CGG AGA GCT CT	799bp	Touchdown 60-58
RAD50-2794	RAD50:c.2795A>G	TAA TTG GGA GCA CAT GGC CTA G	GTC CGA CGT GGT GCT ATG AA	530bp	Touchdown 63-61
RBL1-2312	RBL1:c.2312C>T	TGG TTA AAA TGG TAA ATT TTG TGT T	GAC AAC TTT CCT AGA ACA TAT GCA G	391bp	Touchdown 60-58
RBL2-2487	RBL2:c.2487A>T	GCT CTC AAC AGG TGA CAG GA	AGA ATT CAA AGC AGG TCC AGA	337bp	Touchdown 60-58
RFC4-365	RFC4:c.365/36TA>CT	TGG TGC TGA TTC GCA ATA AA	AAA AGT GAT CAG TAC TGG TTA GAG AGA	273bp	Standard 58
RPA-1051	RPA:c.1051A>G	GTA GGT AAG CTC GTG TAT GAG	CCA TCC TAT AGG ACA AGA TGC	504bp	Touchdown 60-58
SLC19A1-533	SLC19A1:c.533delG	GCA GCT CAT GGA GCT CTT CT	CAG CAC TGA GTC CCC ACA G	397bp	Touchdown 58-56
TP53-869	TP53:c.869G>A	CTT CTT TGG CTG GGG AGA G	CAA GGG TGG TTG GGA GTA GA	370bp	Touchdown 62-60
TP53-1311	TP53:c.1131insA	GAA GTC CTG GGT GCT TCT G	CGA GAC TAA TAC ACA CTA ATA C	396bp	Touchdown 62-60
UIMC1-1690	UIMC1:c.1690T>C	GGC CTT TCC CAA AAC ACT CT	CTT TTG ATT GGC GTT GGA TT	381bp	Touchdown 60-58
XRCC2-563	XRCC2:c. 563G>A	CTT TTG ATT TTG GAT AGC CTG T	CAA CCC CAC TTT CTC CAA TAA	400bp	Standard 58

Table F.2: Primer sequences for variant confirmation for tri-dimensional pooling analysis. Sequence variants listed in HGVS nomenclature. Forward and reverse primers listed in 5' – 3' direction. Optimised cycling conditions for each amplicon listed in **Appendix G.**

Amplicon Name	Variant confirmation	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Size	Cycling conditions
RFC2-348	RFC2:c.384A>G	GCT CTC TTG TTC ATC TGT C	GTG TAT AAT GTC CCT GGT AC	409bp	Standard 58
CHEK2-1496	CHEK2:c.1496G>A	CTG TGG TGA GGA CTC AGT TG	CGA TTA TCA AGC AGA AGC AC	434bp	Standard 62
ATM_Intronic	ATM:c.8850+60A>G	CCT CTA GTA ATG ATG CTG AC	GCA TGC CTG TAA TCA ATA GTG	736bp	Standard 58
RFC2_Intronic	RFC2:c.354+9G>A	TCG CAG TCC ACC TTC TCA G	GGA TCA AAT AAG ACA CGT CAG	339bp	Standard 58
HMMR_Intronic	HMMR:c.12+3344G>C	GCT CAG TTG CCT GTG TTC C	CCT AAT ATT CCC TGA CAA GTC	594bp	Standard 58
HLTF_Intronic	HLTF:c.1375+53C>A	TTG CTT TGT GCC ATC ACT AG	GTA CCT GTA CTC TAG CAC AG	433bp	Standard 58
BRIP1_Intronic	BRIP1:c.2257+36A>T	TGA ATC AGC ATA CTC AAG TG	CTG GTT TAT GGC ATA ATC TG	330bp	Standard 58
E2F5_Intronic	E2F5:c.344+37T>A E2F5:c.344+40C>T	GAT GCT TTG AAC ATG CTC AG	GTT CTT GCT CTC AAG GAC AG	574bp	Standard 58

Table F.3: STS Marker sequences for linkage analysis of potentially related individuals within the South Australian cohort. Sequences obtained from UCSC. As a di- or tri-nucleotide repeat is being sequenced, the expected product size is indicated as a range. Optimised cycling conditions for each amplicon listed in **Appendix G.**

STS Marker	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Size (bp)	Cycling conditions
Chromosome 5				
D5S211	ACT TTG AAA ACC ACT GGC CT	ATG TAT CTA GCC ATG GTA GC	186-204	Standard 62
D5S2006	TGT ATT CTT AAA TTC TGT GAA GAG G	TGT ATT CTT AAA TTC TGT GAA GAG G	125-155	Standard 58
D5S2030	GAT CAG CAG CTG TGG TGA CA	ACT CCA GTG CAG CCA CGT AA	150-176	Standard 62
D5S2034	TTT AAC AAA ATA TAT AAA ATG CCT G	AAT GAA TCT TAC AAC AAT TTG G	177-211	Standard 58
Chromosome 11				
D11S1391	TGC ATG CAT ACA TAC ATA CAT ACA	CAT CCA TCC CTC TGT CTC	158-178	Standard 60
D11S927	AGT GAG CCG AGT TCG C	ACC AAA AGC CTG GAA TG	129-149	Touchdown 60-58
D11S1781	AGC TGT TCT TGT CAC AGG AGA G	ACA AAT TGT CAG TGC CCC	243-251	Standard 62
D11S2000	AGT AGA GAA CAA AAC ACT GTG GC	TTT GAA GAT CTG TGA AAT GTG C	199-235	Standard 62
D11S1893	TCC CTG GAA CCT GGA T	TGA TGT GGG CTT TTT CAA	206-258	Touchdown 62-60

Table F.4: sgRNA pairs designed for CRISPR/Cas9 genome editing of UIMC1. sgRNA pairs designed for the knockout of *UIMC1* in exon 2 and targeted editing of exon 13 in combination with the HDR template for the introduction of the *UIMC1*:c.1690T>C point mutation.

Oligo Pair	Top Sequence (5' - 3')	Bottom Sequence (5' - 3')
ex2 sgRNA KO-A	CAC CGA GTC TCC GAA TCT CGG AAC C	AAA CGG TTC CGA GAT TCG GAG ACT C
ex2 sgRNA KO-B	CAC CGA TTG TGA TAT CCG ATA GTG A	AAA CTC ACT ATC GGA TAT CAC AAT C
ex13 sgRNA A	CAC CGG ATA CTC TCT AAA TGG GAC C	AAA CGG TCC CAT TTA GAG AGT ATC C
ex13 sgRNA B	CAC CGG TGT TAC CTC TGT AAA TCC C	AAA CGG GAT TTA CAG AGG TAA CAC C
HDR Template	TAT CCA GGG GCC CAG AGG CAA AAG CAC ATC TTA AAA CTG TGT TTT CTT CCT TTT TTC TTA TTG GAT AGG AAT GAG AAG TGT CAC CTC TGT AAA TCT CTC GTC CCA TTT AGA GAG TAT CAG TGT CAT GTG GAC TCC TGT CTC CAG CTT GCA AAG GCT GAC CAA	

Table F.5: Primers for confirmation of digestion and successful modification of CRISPR plasmids (PX330, PX461 and PX462v2.0)

Primer Set	Primer (5' - 3')
U6 Forward	GAG GGC CTA TTT CCC ATG ATT CC
CMV Reverse	CCA TTT ACC GTA AGT TAT GTA AC
bbs1_F	CAC CGG GTC TTC GAG AAG ACC
bbs1_R	GAA AGG ACG AAA CAC CGG GTC

Table F.6: Various amplicons generated through the combination of different primers for the confirmation of digestion and incorporation of sgRNAs into CRISPR cas9 plasmids. Expected product sizes from the specified amplicons. Optimised cycling conditions for each amplicon listed in **Appendix G.**

Amplicon Set	Product size	Cycling Conditions
U6_F + CMV_R	463bp	Standard 58
U6_F + bbs1_R	251bp	Rapid Touchdown (66-55)
U6_F + sgRNA Bottom	251bp	Rapid Touchdown (66-55)
bbs1_F + CMV_R	198bp	Rapid Touchdown (66-55)
sgRNA top + CMV_R	198bp	Rapid Touchdown (66-55)

Table F.7: Amplicons designed to screen the edited CRISPR cells for modification. Amplicons UIMC1.1 – UIMC1.5 designed to screen exon 13 for modifications. Amplicons ex2.KO_1 – ex2.KO_4 designed to screen cells for modifications in exon 2. Forward and reverse primers listed in 5' – 3' direction. Optimised cycling conditions for each amplicon listed in **Appendix G.**

Primer Set	Forward Primer (5' - 3')	Reverse Primer (5' - 3')	Size	Cycling conditions
UIMC1.1	TTG GTG ATC CTG CTG AGT GA	TCC CCT CCA CAG TTG AAC AT	297bp	Touchdown 60-58
UIMC1.2	TGC AAC AGC AGG GTA CAG AG	GAG GGA AAA GCC AGA ACA GA	493bp	Touchdown 63-61
UIMC1.3	GCA ACA GCA GGG TAC AGA GAA	CCC TCC ACA GTT GAA CAT GC	375bp	Touchdown 60-58
UIMC1.4	AGT CCT TTT GAT TGG CGT TGG	CAT ACC TTT GGG TTC TTC AGC CTC	359bp	Touchdown 60-58
UIMC1.5	TGC TCT TGG TGC TCC CTT TTC	CTG ACA ACC AGG AGG GTT AGG	616bp	Touchdown 63-61
UIMC1 ex2.KO_1	TTC ATG GTT TTG GTG AGC TG	GCA ACA AAG CGA GAC CAT CTC	388bp	Standard 58
UIMC1 ex2.KO_2	CAA ACT CCC CAT GGG TAA AG	ATG CCA CGG AGA AAG AAA AA	250bp	Touchdown 63-61
UIMC1 ex2.KO_3	TTC ATG GTT TTG GTG AGC TG	ATG CCA CGG AGA AAG AAA AA	264bp	Standard 58
UIMC1 ex2.KO_4	CAA ACT CCC CAT GGG TAA AG	GCA ACA AAG CGA GAC CAT CT	342bp	Touchdown 63-61

Appendix G: PCR cycling conditions

Table G.1: Standard PCR cycling conditions. Example provided for Standard 58 cycling conditions, with annealing temperature shown in red. This was changed dependent on the temperature indicated on the optimised conditions.

Standard PCR Cycle		
Temperature	**Time**	**Number of repeats**
95°C	10 minutes	1
95°C 58°C * 72°C	30 seconds 30 seconds 1 minute	35
72°C	10 minutes	1
25°C	2 minutes	1

Table G.2: Touchdown PCR cycling conditions. Example provided for Touchdown 60-58 cycling conditions, with annealing temperature shown in red. This was changed dependent on the temperature indicated on the optimised conditions.

Touchdown PCR Cycle		
Temperature	**Time**	**Number of repeats**
95°C	10 minutes	1
95°C 60°C * 72°C	30 seconds 30 seconds 1 minute	5
95°C 59°C * 72°C	30 seconds 30 seconds 1 minute	5
95°C 58°C * 72°C	30 seconds 30 seconds 1 minute	25
72°C	10 minutes	1
25°C	2 minutes	1

Table G.3 Rapid Touchdown (66-55°C) PCR protocol.

Rapid Touchdown PCR		
Temperature	**Time**	**Number of repeats**
95°C	10 minutes	1
95°C 66°C - 55°C* 72°C	5 seconds 5 seconds 5 seconds	11 *annealing temperature decreased 1°C/cycle from 66°C-55°C over 11 cycles.
95°C 58°C 72°C	5 seconds 5 seconds 5 seconds	35
72°C	7 minutes	1
25°C	2 minutes	1

Appendix H: Individual MPS data analysis

Table H.1: Variants identified within the diagnostic genes included on the custom MPS panel from individual sequencing of 133 individuals. Variants broken down by minimum allele frequency (MAF; as determined by gnomAD) with those present in >5% of the population (common) and those present in <5% (rare) which were analysed further. Genes with no variants identified shaded in grey, with patients that failed to successfully sequence filled in grey.

	ATM		BARD1		BRCA1		BRCA2		BRIP1		CDH1		CHEK2		FAM175A		HMMR		MRE11A		NBN		NQO2		PALB2		RAD50		RAD51		RAD51C		RAD51D		TP53		XRCC2	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC001	4	1	12	2	11	0	10	3	3	1	2	0			2	0	1	0			8	1	5	0	0	1	2	1			1	0	0	1	7	1		
SABC002	2	6	5	0	12	0	8	3	4	0	1	0			2	0	0	2			8	0	5	0	0	1	1	0	3	0	2	0	0	1	1	0		
SABC003	4	2	12	0			3	1	4	0	2	0			2	1	0	1	3	0	8	0			0	1	1	0	1	0	1	0			2	1		
SABC004	0	2	9	1			4	0	4	0	2	0			0	1	3	1			0	1	3	0			2	2	1	0			2	0	2	1		
SABC005	2	6	5	0	12	0	8	3	4	0	1	0			2	0	0	2			8	0	5	0	0	1	0	1	3	0	2	0	0	1	1	0		
SABC006	6	2	7	1	11	0	6	4	5	1	1	0			4	0	8	2	3	0			4	0	0	1	1	0	1	0	1	0	1	0	4	1		
SABC007	0	2	7	2			6	3	0	2	1	0			3	0	8	1	4	0	0	2	3	0	0	1	1	0	1	0	2	0			2	0		
SABC008	5	3	4	0	11	1	2	4	1	0	2	0			1	0	0	1					4	0	0	1	2	0	1	0	2	0	1	0	1	3		
SABC009	3	1	4	1	11	0	8	3	3	0	3	0			1	3	7	2	1	0	8	0	5	0	0	1	1	0	2	0	2	0			3	0		
SABC010	5	3	10	2	12	0	7	3	4	1	3	0			1	0	0	1	3	0	10	0	2	3	0	2	1	1					0	1	3	0		
SABC011	3	1	8	0	1	1	7	4	4	0	2	0			1	0	0	2	3	0			2	2	3	1	2	0	2	0	1	0			1	0		
SABC012																																						
SABC013	4	1	3	1	12	1	4	3	4	0	3	1					0	1	3	0	8	1	7	0	0	1	2	1	3	0	2	0			1	1		
SABC014	5	2	7	1	12	0	6	3	4	0	2	1			0	1	0	1	3	0			6	0	0	2	1	0			1	0	0	1	2	2	0	1
SABC015	4	3	11	1	2	0	3	3	1	0	2	1			0	1	0	1	3	0	8	0	5	2	0	1	1	0	1	0	1	3			1	0		
SABC016	4	2	4	1			8	3	4	0	2	0					7	1	1	0			3	0	1	0	1	0							2	0		
SABC017	4	4	11	1	11	0	10	4	5	1	3	0			1	0	1	0					6	0	1	0	2	0			1	0	0	2	2	2		
SABC018	3	2	4	1	1	1	3	3	0	1	1	0			2	1	6	1			8	0	5	0	3	3	1	0	1	0	1	0	0	1	1	4		
SABC019	0	3	11	1	11	0	7	3	5	0	2	0	0	1	1	0	2	1			10	2	6	0	1	1	2	0	1	0	0	1	1	0	2	0		
SABC020	0	3	6	1	12	1	2	3	4	0	3	0			1	1	0	1			8	0	3	0	0	1	2	1	1	1	1	0			2	1		
SABC021	0	2	9	1	10	1	3	3	1	0	2	1			2	0	6	1			8	0	2	1	3	5	2	0	1	0	0	2			1	2		
SABC022	0	1	5	1			9	4	4	1	2	1	0	1	1	0	1	1	1	0	8	1	3	0	0	1	2	0	1	0	1	0			5	1		
SABC023	4	4	5	1	11	2	5	3	0	1	2	1					7	2			0	1	5	0	0	2	2	2	1	1	2	0			1	1	0	1
SABC024	4	2	4	0	1	1	5	3	5	0	2	4	0	1			8	2	3	0			5	0	0	1	2	1			1	0	0	1	2	0		
SABC025	4	3	3	1	11	1	4	3	3	1	2	0			1	0	2	2	3	1			3	1	0	3	1	0	1	0	1	0	1	0	2	0		
SABC026	5	2	4	2	1	0	6	4	3	0	3	1			1	0	0	2	1	1	0	2	3	0	3	4	1	0	1	0	1	0			1	0	0	1
SABC027	7	2	6	0	11	0	6	4	5	3	1	2			4	0	7	2	3	0			4	0	0	1	1	0	1	0	1	0	0	1	4	1		
SABC028	4	3	8	0	1	1	5	3	1	0	2	0					7	1	0	1	0	1	3	0	0	1	1	1	2	0	1	3	0	1	2	0	0	1
SABC029	3	3	7	1	11	2	3	3	4	1	3	0			2	0	0	1	3	0			3	0	0	1	1	0	1	0	2	0			3	3		
SABC030	5	2	8	2	0	1	3	4	5	1	2	1					0	2	2	0	7	0	3	1	0	1	1	2	1	0	1	0			3	0		
SABC031	3	1	8	1	1	0	12	5	1	0	3	0			1	0	4	1			8	0	6	0	0	1	1	0					1	1	4	1		
SABC032	5	1	9	0	11	1	3	3	4	0	1	0	0	1	1	0	4	1			8	1	3	1	1	2	2	2	2	0	1	0			1	0	0	1

	ATM		BARD1		BRCA1		BRCA2		BRIP1		CDH1		CHEK2		FAM175A		HMMR		MRE11A		NBN		NQO2		PALB2		RAD50		RAD51		RAD51C		RAD51D		TP53		XRCC2	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC033	4	1	8	0	11	0	6	5	1	0	2	2			2	0	6	1	1	0	8	1	5	0	3	1	1	0			2	0	2	0	1	0		
SABC034	6	2	2	1			7	3	4	1	3	0			2	0	0	1	3	1	8	0	3	1	0	1	2	1	1	0	1	0	0	1	5	0		
SABC035	4	1	2	0	10	2	10	3	3	0	2	1			1	0	0	1	1	0			6	1	1	1	1	0	1	0	1	0			1	0		
SABC036	0	2	12	1			8	3	5	0	3	0	0	1	2	0	0	2			8	0	6	0	0	1	2	0	1	0					4	0		
SABC037	0	1	13	2	0	1	9	3	4	0	3	1			2	0	7	1			0	1	6	0	3	5	1	0			1	0	1	0	3	0		
SABC038	4	3	12	1	11	1	7	3	5	1	4	1			1	0	7	1	3	0	8	0	4	0	1	1	1	1	1	0	1	0	0	1	2	0		
SABC039	4	3	11	1	1	1	8	3	4	0	3	1	0	1			7	1			8	0	8	0	1	2	1	0	1	0	1	0			3	0		
SABC040	3	1	9	1	1	1	10	4	4	0	2	1			1	0	8	1	1	0	8	0	3	0	2	0	1	0	1	0	1	3	2	0	1	0		
SABC041	0	2	2	1	12	1	3	3	3	0	3	1	0	1	2	0	0	1			8	0	6	0	0	1	2	0	1	0	1	0			3	4		
SABC042	6	1	10	2	1	0	3	3	4	1	2	0	0	1	2	0	0	2	3	0			6	1	1	2	2	0	1	0	2	0	1	0	2	1		
SABC043	5	3	11	1	11	2	3	4	4	0	4	0			2	0	2	2			8	1	3	0	0	1	2	0	1	0			2	0	3	1	0	1
SABC044	4	5	6	2	1	1	9	4	4	0	3	2			2	0	0	1	3	0	8	0	5	0	0	1	1	0			1	0	2	0	2	1		
SABC045	4	2	9	1	12	2	3	5	4	0	3	1	0	1	2	0	3	1	3	1			6	0	3	1	1	1	1	0	1	0			4	1		
SABC046																																						
SABC047	3	1	8	0	11	1	8	5	5	0	1	1					6	1					4	0	0	1	1	0	2	0	0	3			1	0		
SABC048	3	3	9	2	11	0	6	4	4	0	1	1			1	0	0	1					5	0	0	1	2	0	2	0	1	0			2	8		
SABC049	5	3	3	0	1	1	3	3	4	1	2	0					0	1	1	0	8	0	4	0	0	1	1	0	1	0					1	3		
SABC050	0	1	3	0	11	0	6	4	3	1	3	0	0	1	1	0	0	2	1	0			4	0	0	1	1	0	1	0					1	1		
SABC051	5	5	8	0			6	4	4	1	2	1			1	0	0	1			8	0	3	0	3	5	1	0	1	0	1	0			1	0		
SABC052	4	3	7	1			6	5	4	0	2	1			1	0	0	1	3	1	7	0	3	0	0	1	1	0	1	0	1	0			1	4	0	1
SABC053	3	4	3	1			8	3	5	0	2	1			1	0	0	3	3	0	7	0	4	0	1	1	1	0			1	0			1	1		
SABC054	0	2	11	1	11	1	13	4	5	0	3	0	0	1	2	0	2	1	1	0			5	0	3	3	1	0	2	0			1	0	5	0	0	1
SABC055	2	8	10	0	11	2	3	3	4	0	2	0			1	0	8	2	1	0			6	3	1	0	2	0	1	0	1	0			1	0		
SABC056	3	1	8	0	1	1	8	4	5	0	2	0					6	1	3	0	8	0	4	0	0	1	1	1	1	0	1	0	0	1	1	0		
SABC057	5	3	4	0	12	1	3	3	4	0	2	2			1	0	8	1			10	1	2	0	0	1	1	0	1	0	1	0	1	0	1	4		
SABC058	5	2	11	1	1	1	9	5	3	0	3	1			2	0	4	1	3	0	8	0	2	3	0	1	2	0			2	0	2	0	2	0		
SABC059	6	1	7	1	1	0	6	3	4	0	3	1			1	0	1	0	1	0	8	0	5	1	0	1	1	0	1	0	1	0	0	1	2	1		
SABC060	6	4	7	0	11	1	8	3	4	0	2	0			1	0	7	1					4	1	3	5	2	1	2	0	1	0			2	0		
SABC061	5	1	8	1	11	0	8	3	1	0	2	0			1	0	7	1	3	0			5	0	1	1	1	0	2	1	1	0			2	0		
SABC062	4	3	8	1	1	2	7	3	5	1	2	0			5	0	7	1	1	0			3	0	0	1	2	0	1	0	1	0			3	0		
SABC063	0	3	12	1	12	1	2	4	1	0	4	0			1	1	8	2	3	0	8	0	3	0	3	1	2	0	1	0	1	0			2	0	0	1
SABC064	0	4	5	0	11	1	8	4	3	0	3	0			1	0	0	1	3	0			3	2	0	1	2	0	1	0	1	0	1	1	4	0		

	ATM		BARD1		BRCA1		BRCA2		BRIP1		CDH1		CHEK2		FAM175A		HMMR		MRE11A		NBN		NQO2		PALB2		RAD50		RAD51		RAD51C		RAD51D		TP53		XRCC2	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC065	4	2	8	2	11	1	7	4	5	0	1	0			1	0	6	1	3	0	8	1	4	0	3	1	2	1	1	0	1	0	1	0	3	1		
SABC066	4	2	9	4	11	0	3	3	4	0	4	0			3	0	0	2			3	0	4	2	0	1	1	0	1	0	1	0	0	2	4	0		
SABC067	4	2	12	1	11	1	8	2	3	0	2	1			2	0	0	1	0	1	11	0	8	3	0	1	2	0	1	0			1	0	3	0	0	1
SABC068	4	1	2	0	11	1	2	6	4	0	2	0					8	1	3	0	7	0	5	0	1	1	1	0	1	0	1	0			1	1	0	1
SABC069	3	1	7	0	12	0	7	4	3	0	3	1	0	1	1	0	1	0			7	0	3	1	3	1	1	0	1	0	1	0	1	0	5	0		
SABC070	6	2	11	1	1	2	4	3	4	0	4	1			3	0	0	1			3	0	4	2	0	1	1	1			1	0			3	0		
SABC071	4	4	4	1	1	1	4	4	4	0	2	0			2	0	6	1			8	0	3	0	0	1	1	0	1	0	1	0	0	1	1	5		
SABC072	6	2	3	2			4	3	1	0	3	0			1	1	0	1	3	0	8	0	6	0	3	6	1	0	1	0	1	0			2	0		
SABC073	0	1	11	1			7	5	5	0	3	0			1	0	0	1	3	1			1	1	0	1	2	0	1	0	2	0			2	0		
SABC074	4	3	4	2	11	0	5	3	4	0	2	0					6	1					3	1	0	2	1	1	1	1	1	1			2	0		
SABC075	4	5	7	1			6	3	4	0	2	1			2	0	5	2	2	0	8	1	6	0	0	1	2	0	1	0			1	0	3	0		
SABC076	4	3	6	1			3	3	4	0	2	0			1	0	7	1			7	1	3	1	0	1	1	0	1	0			1	0	2	0		
SABC077	6	2	4	0	12	0	7	3	1	0	2	3			1	1	3	3	3	0	8	1	3	2	0	1	1	0	1	0	1	0			2	0	0	1
SABC078	4	4	10	0	12	0	4	5	3	0	2	1	0	1	1	0	6	1	3	0	10	1	4	2	0	1	2	1			1	0	1	1	2	0	0	1
SABC079	4	2	10	1	12	0	5	4	4	0	1	0					7	2	3	0	8	0	4	1	3	1	2	0	1	1	1	0	1	0	5	0		
SABC080	4	3	12	1	2	1	8	3	4	0	4	0	0	1			7	1	3	0	8	0	5	0	0	1	1	0	1	0	1	0	2	0	5	0		
SABC081	7	2	3	2	11	0	6	4			2	1			1	0	0	1	3	0	0	1	6	0	3	5	2	0	1	0	1	0			2	1		
SABC082	4	2	11	2	1	0	6	3	4	0	3	2					0	1	3	0	3	0			4	5	2	0	1	0	2	0			2	0		
SABC083	0	1	3	1	11	0	7	5	5	1	3	0			1	0	0	1	3	0	8	0	5	0	0	1	1	1	3	0	1	0			1	0		
SABC084	4	1	6	2	0	1	9	3	3	0	3	2					0	1	0	1	10	0	3	1	0	1	2	0	1	0	1	0			2	0		
SABC085	5	3	2	0	1	1	3	4	3	1	2	0			1	0	0	1			7	1	2	0	0	1	2	0	1	0	1	0			1	3		
SABC086	0	2	7	1	12	0	4	4	5	0	3	0			2	0	7	1	2	1			6	0	0	3	1	0	1	0	2	0			3	1		
SABC087	4	2	11	1			6	4	3	0	1	0	0	1			7	1	3	0	8	0	3	0	0	1	1	0	1	0	1	0			3	1		
SABC088	6	2	11	2	11	0	8	5	3	0	3	1			0	1	0	2	3	0	8	0			0	1	1	0	1	0	1	0	1	1	3	0		
SABC089	0	2	7	0	2	1	5	4	2	0	3	0			1	0	6	1	1	1	0	1	2	0	0	1	1	1	1	0	1	0			1	4		
SABC090	6	2	4	0			7	5	3	0	2	0			1	0	0	2	1	1			5	0	1	1	1	0	1	0					1	0		
SABC091	6	2	8	1			11	5	5	0	3	0			1	0	0	2					6	1	1	1	1	0	1	0	1	0			2	1		
SABC092	4	2	4	1	11	1	3	3	2	1	3	1					8	1	3	0			5	0	1	1	1	0	3	0	2	0	1	0	2	0		
SABC093	4	2	2	2	11	0	3	3	4	0	3	3			2	0	0	1	1	0	0	1	3	0	0	1	2	0	1	0	1	0			3	0		
SABC094	6	1	2	1	1	1	7	4	1	1	4	0			2	0	4	2	1	0	8	0	6	0	0	1	2	0			1	0	0	1	3	1	0	1
SABC095	4	2	9	4	11	1	7	5	4	0	4	1			2	1	7	1					3	0	0	1	2	1	1	1	1	0			3	0		
SABC096	4	2	12	1	1	0	3	3	4	0	3	0			3	1	2	2					6	0	0	1	2	0	2	0	1	0			3	0		

	ATM		BARD1		BRCA1		BRCA2		BRIP1		CDH1		CHEK2		FAM175A		HMMR		MRE11A		NBN		NQO2		PALB2		RAD50		RAD51		RAD51C		RAD51D		TP53		XRCC2	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC097	4	2	7	1	1	2	7	5	1	1	3	0					3	2	3	0	8	0	3	0	0	1	1	0	1	0			1	0	2	0		
SABC098	3	3	7	1	1	0	6	4	3	0	1	0					0	1	3	0	8	0	3	2	3	1	1	0	1	0					3	5	0	1
SABC099	5	6	6	1			3	3	5	0	3	0			1	1	0	3	3	0	8	0	6	0	3	1	1	1	2	0	1	0	0	1	3	0		
SABC100	5	3	4	0	11	0	6	5	4	1	3	1			1	0	8	1	2	0	8	0	5	0	0	1	2	0	1	0	1	0	2	0	1	4		
SABC101	3	4	8	0	0	1	10	3	5	0	3	0			1	0	6	1					2	0	1	1	1	1			1	0			2	0		
SABC102	4	1	8	0	11	0	10	3	5	0	3	0			1	0	6	2	3	0	8	0	3	1	0	1	1	1			1	0	1	0	2	0		
SABC103	2	8	10	0	11	2	3	3	4	0	2	0			1	0	8	2	1	0			7	2	0	1	2	0	1	0	1	0			1	0		
SABC104	4	3	10	1	11	2	6	4	4	0	3	0					7	2	1	0	8	0	5	0	3	5	1	0	1	0			1	0	3	1	0	1
SABC105	0	1	1	2	11	0	2	3	0	1	2	0			1	0	0	3					3	0	1	1	2	0	1	0	1	0			3	4		
SABC106	0	2	7	2	1	1	3	3	4	0	2	1			0	1	2	2			8	0	8	2	0	1	0	1	1	0	1	0			1	0		
SABC107	4	6	8	1	1	2	6	4	3	0	3	1			1	0	6	2			4	0	6	1	0	1	1	1			1	0	1	0	3	5		
SABC108	0	3	11	1			3	3	3	0	3	1					0	1	1	0	8	0	2	4	3	5	1	0	3	0	2	0			3	1	0	1
SABC109	0	2	8	1	11	1	5	3	3	0	2	1	0	1	1	0	6	2			8	0	6	0	0	1	1	0	1	0	1	0			2	1		
SABC110	4	1	3	0			8	4	4	1	1	2					6	1	1	1	8	0	6	0	0	1	1	0	1	1	1	0	0	1	2	0		
SABC111	4	3	8	1	11	0	3	5	5	0	3	0			1	0	0	2	1	0			4	0	0	1	2	0	1	1	0	3	0	1	2	0		
SABC112	4	4	12	1	11	3	6	6	3	1	3	2	0	1	1	0	7	1	3	1	8	0	7	0	0	2	1	1	2	1	1	0			3	1		
SABC113	0	2	7	1	11	0	8	4			4	1			2	0	0	1	3	0	8	0	3	0	0	1	1	0	1	0	1	0			3	0		
SABC114	4	2	8	1	11	1	7	3	4	0	2	1	0	1			6	1	3	0			6	0	0	1	0	1	3	0	1	0			3	0		
SABC115	7	1	12	1	11	1	10	4	4	1	2	1			2	0	6	1	3	0	3	0	6	0	3	1	1	2	2	0	1	0			4	1		
SABC116	4	2	9	1	1	1	6	3			2	1					0	3	1	0			5	0	3	5	1	0	1	0	1	0	1	0	1	0		
SABC117	4	2	5	0	11	0	12	4	5	0	2	0			1	0	0	1	1	0			3	0	0	1	2	2	0	1	1	0			1	0		
SABC118	0	3	5	0	11	0	8	3	4	0	2	1			1	0	0	2	3	0					0	1	1	0			1	3			4	0		
SABC119	4	1	2	1	1	0	9	3	4	0	1	1			1	0	6	2	3	0			3	0	1	1	1	0			1	0	1	0	3	1		
SABC120	4	1	3	0			6	4	3	1	3	0			1	0	0	2					4	0	3	1	1	0	3	0	2	0			1	1		
SABC121	4	2	8	2	11	1	7	3	3	0	2	0	0	1	1	1	0	1	1	0	8	0	3	0	1	1	1	0			1	0	0	1	3	0		
SABC122	4	1	4	0	11	2	6	4	3	1	3	0	0	1	1	0	4	2	3	0	8	1	2	0	0	1	1	0	1	0	1	0			1	1		
SABC123	4	2	6	1	1	1	8	5	4	1	5	0			2	0	1	0	3	0	8	1	1	1	1	1	1	0	1	0	2	0	0	1	4	0		
SABC124	4	3	3	0	8	0	0	4	0	1	1	0			2	0	0	1					3	0	0	1									1	0		
SABC125	3	1	2	0	11	0	3	3	4	0	3	0			1	0	0	1	3	0			8	0	1	1	1	1	1	2	1	0	2	2	1	0		
SABC126	0	2	9	0	2	0	8	0	3	0	3	0			1	0	8	0	3	0			4	0	3	1	2	2	1	0	1	0			2	4		
SABC127	3	1	8	1	12	0	7	5	1	0	2	1					0	2			10	1	5	0	1	0	1	0	1	0	1	0	0	1	1	1		
SABC128	4	2	11	2	11	2	8	4	3	0	3	0					7	1					3	0	1	1	1	0	1	1	1	0			2	0		
SABC129	4	2	3	0	11	0	9	5	4	0	2	0			1	0	6	1	3	0			4	0	0	1	2	1					1	0	1	1		
SABC130	4	2	5	0	11	0	2	3	4	0	3	1	0	1			6	2	1	0	10	0	3	1	3	5	2	0	2	0	1	0			2	0		
SABC131	4	5	8	1	11	1	8	4	3	0	2	1			1	0	6	1	3	0	1	1	3	1	3	5	1	0	1	0	1	0			1	8		
SABC132	3	2	4	0	13	1	3	3	3	0	2	0	0	1	1	0	6	2	1	1	7	0			0	1	1	0	3	0	1	0			1	1		
SABC133	4	4	4	1	11	1	3	3	5	0	3	1					3	2	3	0			5	4	0	1	1	0	1	0	1	0			2	0		

Table H.2: Variants identified within 16 of the discovery genes (A-K) included on the custom MPS panel from individual sequencing of 133 individuals. Variants broken down by minimum allele frequency (MAF; as determined by gnomAD) with those present in >5% of the population (common) and those present in <5% (rare) which were analysed further. Genes with no variants identified shaded in grey, with patients that failed to successfully sequence filled in grey.

	ATF1		BRCC3		CDKN1A		CDKN2A		CHEK1		CKS1B		E2F1		E2F2		E2F3		E2F4		E2F5		E2F6		EP300		GADD45A		HLTF		KAT2B	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC001	1	1			1	0	1	0	0	1							1	0	0	1	2	0			6	1	1	0	3	3	6	0
SABC002	0	1			1	0	1	0	2	1					1	1	1	0			1	0			0	3	1	0	1	0	8	0
SABC003							1	0	0	1					2	0					2	0			6	2	1	0	3	0	7	0
SABC004							1	0							1	0									2	1			1	1	1	0
SABC005	0	1			1	0	1	0	2	1					1	1	1	0			1	0			0	3	1	0	1	0	8	0
SABC006	1	0					1	0	1	2							0	1			2	0			6	4	1	0	1	1	6	0
SABC007			0	1	1	0	1	0	2	1			0	1	2	0					2	1			7	2	1	0	1	1	3	0
SABC008	1	0					1	1	0	1					2	0					0	2			4	2	2	1	3	0	8	1
SABC009	1	0					1	0	2	1					1	0	1	0			1	0			6	2	1	0	3	0	7	0
SABC010	1	0			1	0	1	0	0	2			0	1			1	0			1	0			6	3	1	0	3	0	5	2
SABC011	2	1					1	0	2	1															4	2	2	1	1	0	5	4
SABC012																																
SABC013	1	1			1	0	1	0	2	6					1	0	1	0			1	0			1	2			3	0	5	1
SABC014	0	1			1	0	1	0	2	1					2	0					1	0			5	2	2	0	1	1	7	0
SABC015	1	0					1	0	0	1			0	1			1	0	0	1	1	0			7	2	1	0	3	1	1	0
SABC016	1	1	0	1			1	0	3	1					2	0			1	0	1	0			7	3	1	0	1	0	10	0
SABC017	2	0					1	1	0	1					2	0	1	0			1	0			7	7	1	0	3	1	7	0
SABC018	1	1					1	0	2	1											1	1			6	3	1	0	1	0	10	0
SABC019	1	1	0	2	1	0	1	0	2	1					2	0					1	0	0	1	10	3	1	0	3	0	1	1
SABC020	2	1	0	1	3	0	1	0	0	1					2	0					1	0			6	3	1	0	1	0	7	1
SABC021					3	0	1	0	1	1			0	1	2	0					1	0			4	2			1	1	7	0
SABC022							1	0	1	1					2	0			0	1	2	0	0	1	1	2	1	0	3	1	10	0
SABC023	1	0			3	0	1	0	2	1					2	0					1	0			4	2	1	0	3	1	8	1
SABC024					0	1	1	0	2	1					2	0	1	0			0	1			0	5	1	0	3	1	7	0
SABC025	2	0	0	1			1	0	0	1			0	1	2	0	1	0							4	0	1	0	3	2	10	0
SABC026	2	0	0	2	1	0	1	0	0	1					2	0	0	1			1	0			1	4	1	0	3	2	8	1
SABC027	1	0					1	0	2	1							0	1			2	0			6	4	1	0	1	0	3	0
SABC028	1	0			1	0	1	1	2	1															7	2	1	1			5	1
SABC029	1	0			1	0	1	1	2	1							1	0	1	1	0	1			3	3			3	1	8	1
SABC030	1	0	0	1			1	1	0	1					2	0	1	0							7	3	1	0	3	1	8	0
SABC031	1	0					1	0	2	1					2	0					0	1			3	0					2	0
SABC032	1	0			1	0	1	0	0	1			0	1	2	0			0	2	1	0			4	2	1	0	1	0	6	0

	ATF1		BRCC3		CDKN1A		CDKN2A		CHEK1		CKS1B		E2F1		E2F2		E2F3		E2F4		E2F5		E2F6		EP300		GADD45A		HLTF		KAT2B	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC033	2	0	0	2			1	0	1	1					2	1	1	0			0	1			6	2	1	0	1	1	5	0
SABC034					1	0	1	0	2	1					2	0			0	2	1	0			1	2	2	0	3	0	7	0
SABC035	2	1			3	0	1	0	0	1					2	0			0	1	1	0			4	2	0	1	2	0	6	0
SABC036					3	0	1	0	2	1					2	0									11	2	1	0	0	1	1	0
SABC037	1	0					1	0	2	1			0	2	2	1	1	1	0	2	0	1			4	2	2	0	1	1	1	0
SABC038	1	0			3	0	2	2	2	1					1	0					1	0			4	6	2	0	3	0	9	0
SABC039	1	0			1	0	1	0	0	1			0	1	2	0					1	0			7	3	1	0	1	0	7	2
SABC040	2	1					1	0	0	1					2	0	1	0			1	0			4	2	0	1	2	1	6	0
SABC041	2	0	0	1	1	0	1	0	0	1					2	0					1	1			4	3	1	0			5	2
SABC042	1	0					1	0	2	1					2	0					2	0			5	2	1	0	3	2	8	1
SABC043							1	0	2	1					2	0			0	1	1	0			6	3	1	0	3	0	7	0
SABC044	1	0					1	0	2	1					2	0	1	0			1	1			7	2	1	0	1	0		
SABC045	1	0	0	1	3	0	1	0	0	1											1	0			1	3	1	0	1	0	7	0
SABC046																																
SABC047					1	0	1	0	1	1					2	0	1	0	0	1					5	2	0	1	1	0	7	0
SABC048	2	1					1	0	0	1					1	0									8	2	1	0			7	1
SABC049	1	0					1	0	1	2							1	0			1	1			8	2	1	0			7	0
SABC050	1	0			1	0	1	0	0	1			0	1	2	0					1	1			7	2	2	0	3	0	6	0
SABC051	1	1			1	0	1	0	2	1							1	0							7	2	1	0	3	2	6	0
SABC052	1	0			1	0	1	0	0	1					1	0									4	2	1	0	3	1	7	1
SABC053	1	0			1	0	1	0	2	1							1	0			1	0			0	2	1	0	1	0	7	1
SABC054					1	0	1	0	0	1					3	0					1	0			6	1	1	0	3	0	8	0
SABC055	1	0	0	2	1	0	1	0	2	1					2	0					1	0			7	4	1	0	3	1	3	0
SABC056	1	0	0	1	1	0	1	0	0	1											1	0			6	2			1	0	8	1
SABC057	2	0	0	1	1	0	1	0	1	1			0	1	2	0	0	1			1	1			7	2			3	1	1	0
SABC058	1	1			1	0	1	0	1	2					2	0	1	0			1	0			4	2	1	0	3	0	6	0
SABC059	2	1					1	0	0	1			0	1	2	1									7	3	1	0	3	0	7	0
SABC060	2	0			1	0	1	0	2	2					2	0	1	1			1	0			5	3	2	0	1	1	5	0
SABC061							1	0	0	1					2	0	1	0			1	0			6	2	2	0	0	1	7	1
SABC062	1	0					1	0	0	1											1	0			1	2	1	0	2	2	7	0
SABC063					1	0	1	0	2	1					2	0	1	0							1	2	2	0	3	0	7	0
SABC064	2	0			1	0	1	0	0	1			0	1	2	0					1	0			7	2	1	0	1	0	6	0

	ATF1		BRCC3		CDKN1A		CDKN2A		CHEK1		CKS1B		E2F1		E2F2		E2F3		E2F4		E2F5		E2F6		EP300		GADD45A		HLTF		KAT2B	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC065	2	1			1	0	1	0	2	1					1	0					1	1			4	2			1	1	8	0
SABC066	2	0			3	0	1	0	2	2											1	0			1	2	1	0	0	1	8	2
SABC067					3	1	1	0	0	1					2	0	1	0			1	0			6	0	1	0	3	1	6	0
SABC068	1	0	0	1			1	0	2	1					2	1			0	1	1	0			7	3	1	0	3	0	5	1
SABC069	2	0			1	0	2	0	1	1					2	0	1	0			1	0			1	2	1	0			6	1
SABC070	1	1					1	0	0	2					2	0	0	1			1	0			0	2	2	0	3	1	3	0
SABC071	2	0					1	0	2	1							0	1			1	1			6	2			1	3	4	0
SABC072	1	0					1	0	3	1			0	1							1	0			5	2	1	0	3	1	7	0
SABC073	1	1			1	0	1	0	0	2			0	1	2	1					1	0			7	2	1	0	1	0	7	0
SABC074	2	0			3	0	1	0	2	1					1	0	2	0	0	1					6	2	1	0	3	0	5	1
SABC075					3	0	1	0	0	1			0	1	2	1	1	1			1	0			0	2	1	0	2	3	5	0
SABC076	2	0	0	1			1	1	2	1							1	0			1	0			0	2	1	0	1	0	6	0
SABC077	2	0	0	1			1	0	0	1									0	1	1	0			1	2	1	0	3	0	6	0
SABC078	1	0	0	1			1	0	2	1					2	0	1	0	0	1	1	0			6	3	1	0	3	1	7	0
SABC079	0	1					1	0	0	1							1	0			1	0	0	1	4	2	2	0	3	0	7	0
SABC080	2	0			1	0	1	0	2	2					2	0	1	0			1	0			5	2	1	0	1	0	8	0
SABC081	1	0			3	0	1	0	2	1			0	1	2	0	1	0			1	0			6	3	1	0	2	1	8	0
SABC082	1	0					1	0	0	2			0	1	2	0	1	0			1	0			5	2	1	0	1	2	8	1
SABC083	2	1	0	1			1	0	1	1					2	0	0	1			1	1			8	3	1	0	2	1	6	1
SABC084	0	1	0	1			1	0	0	1					2	0	1	1							8	2	2	0	1	1	3	0
SABC085	2	0			3	1	1	0	1	1					2	0	0	1					0	1	6	3			3	0	6	0
SABC086	1	1					1	0	0	1					2	0	1	0			1	0			7	3	1	0	3	0	6	0
SABC087	1	1					1	0	1	1					2	0	1	0			1	0			4	2	2	0	1	0	6	0
SABC088	1	0					1	0	0	1											1	0			6	3	1	0	1	0	5	2
SABC089	2	0					1	0	1	1			0	1	2	0	0	1			1	1			4	2	2	0	2	1	7	0
SABC090	2	1			1	0	1	0	0	1					2	0					1	0			7	2	0	1	1	0		
SABC091	1	0					1	0	1	1			0	2	2	0									4	2					7	1
SABC092	1	1	0	2	1	1			0	1					2	0					1	0			7	3	1	0	3	0	5	1
SABC093					3	0	1	0	2	1					2	1					1	0			4	2	1	0	1	0	8	1
SABC094	1	0			3	0	1	1	0	1					2	0					1	0			0	2	1	0	3	0	7	0
SABC095	1	1			1	0	1	0	0	1			0	1	2	0	1	0			1	0			4	3	1	0	3	0	5	1
SABC096	0	1					1	1	0	2					2	1	1	0			1	0			4	2	1	0	3	1	6	0

	ATF1		BRCC3		CDKN1A		CDKN2A		CHEK1		CKS1B		E2F1		E2F2		E2F3		E2F4		E2F5		E2F6		EP300		GADD45A		HLTF		KAT2B	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC097							1	0	0	1			0	1			1	0			1	0			7	5			1	1	6	0
SABC098	1	0			3	0	1	0	1	1			0	1	2	0	1	0			1	1			6	5	2	0	2	0	6	0
SABC099	2	1			1	0	1	0	0	1					2	0	1	0			1	0			1	3	1	0	3	0	6	0
SABC100	1	0			1	0	1	0	0	1					2	0					1	1			7	2	1	0	1	0	7	0
SABC101	1	0			3	0			1	1					2	0	1	0			1	0			6	2	2	0	3	0	6	0
SABC102					3	0	1	0	0	1					2	0									4	2	1	0	1	0	8	0
SABC103	1	0			1	0	1	0	3	0					2	0					1	0			7	4	1	0	3	1	3	0
SABC104	2	0					1	0	1	1			0	1	2	0					1	1			4	2	1	0	3	0	3	0
SABC105	2	0			1	0	1	0	0	1							1	0			1	1			4	2	2	0	1	1	8	0
SABC106	2	0	0	1			1	0	1	1			0	1	2	0					1	0			6	2	1	0	3	0	7	0
SABC107	1	0					1	0	2	1											1	1			6	2	1	0	2	0		
SABC108	1	1			1	0	1	0	2	1							1	0			1	0	0	1	6	3	1	0	7	0		
SABC109	2	0			1	0	1	0	3	1			0	1	1	0					1	1			8	2	1	0	2	1	7	0
SABC110	1	1	0	3			1	0	0	1									0	1	1	0			0	2	1	0	1	1	8	1
SABC111	2	1	0	1			1	0	2	1					2	0			0	1	1	0			1	2			2	1	10	0
SABC112	1	1			0	1	1	0	1	1			0	2	2	0					1	1			7	1	1	0	3	1	8	1
SABC113	1	1			1	0	1	0	2	1			0	1	2	0	1	0			1	0			4	2	1	0	1	0	7	1
SABC114	2	0			3	0	1	1	0	1					2	0					1	0			4	2	2	0	3	1	5	1
SABC115					3	0	1	0	1	1			0	1	2	0					2	0			0	2	1	0	1	1	6	0
SABC116	1	1	0	1	1	0	1	0	3	1											1	0			4	3	2	0	3	0	7	0
SABC117	2	1	0	1			1	1	0	3					2	1	0	1			1	0			4	2	2	0	3	0	3	0
SABC118	1	0	0	1	1	0	1	0	2	1					2	0					1	0			0	2	1	0	1	1	7	0
SABC119	2	0					1	0	1	1					2	0	1	0	0	1					4	2	1	0	3	1	1	0
SABC120	1	0					1	0	2	1											1	0			7	2			1	1	5	0
SABC121	2	0			1	0	1	0	2	1					2	0									0	2	1	0	1	0	3	0
SABC122	1	0			1	0	1	0	2	1			0	1	2	0					1	1			4	2					6	0
SABC123	2	0	0	3	1	0	1	0	2	1							1	0	0	1	1	0			5	2	1	0	1	0	6	0
SABC124	1	0							2	1					1	0					2	0			0	1	2	0			6	0
SABC125	2	2			3	0	1	0	2	1					2	0	1	0	0	1	1	0			4	2	1	0	1	0	7	1
SABC126	1	0					1	0	2	1					2	0					1	1			4	2	2	0	1	0		
SABC127	2	0			1	0	1	0	2	0			0	1	2	0	1	0			2	0			7	3	1	0	3	0	3	1
SABC128			0	1			1	0	2	1									0	1	1	0			4	2	1	0	1	0	7	1
SABC129	1	0			1	0	1	0	2	1			0	1	2	0			0	1	1	1	0	1	7	2	1	0	1	2	7	0
SABC130	1	0	0	2			1	0	2	1					2	0	1	0			1	0			4	4	1	0	3	0	7	0
SABC131	2	0					1	0	0	1					2	0	1	0			1	1			4	2			0	1	7	0
SABC132	1	0			3	0	1	0	0	1					1	0	1	0							5	2			2	0	6	0
SABC133	1	0			1	0	1	0	0	1			0	1	2	0					1	0			4	2	1	0	3	0	8	0

Table H.3: Variants identified within 16 of the discovery genes (P-W) included on the custom MPS panel from individual sequencing of 133 individuals. Variants broken down by minimum allele frequency (MAF; as determined by gnomAD) with those present in >5% of the population (common) and those present in <5% (rare) which were analysed further. Genes with no variants identified shaded in grey, with patients that failed to successfully sequence filled in grey.

	PKMYT1		*PRKDC*		*RBL1*		*RBL2*		*RFC2*		*RFC3*		*RFC4*		*RFC5*		*RPA1*		*RPRM*		*RPS6KA1*		*SFN*		*SLC19A1*		*SMARCD2*		*UIMC1*		*WEE1*	
	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC001	1	0	0	2			7	1	0	1			2	1			2	0			6	1	0	1	4	1	0	1	3	0	1	1
SABC002			0	5			5	0	0	1			2	0			5	1			4	1			2	0					2	0
SABC003			0	4	1	0	6	1					1	0			2	1			4	2			2	0	0	1	2	0	1	0
SABC004			0	1			3	0					3	1	3	0	2	0			3	1			4	0			2	0	1	0
SABC005			0	5			5	0	0	1			2	0			5	1			4	1			2	0					2	0
SABC006			0	3	1	1	7	0	0	2			2	1	4	1	5	1			5	1			4	0			5	1	2	0
SABC007	1	0	8	6									1	1	2	0	5	0			2	1	0	1	2	0			0	1		
SABC008			0	4	1	0	4	1					2	0	2	0	4	1			3	1			4	0			3	1	1	0
SABC009			0	6									1	0			6	0			5	1	0	1	4	0			2	0	2	0
SABC010			0	4			4	0					1	0			2	1			4	2			2	1					1	1
SABC011			8	6			4	0	0	1			2	0	1	0	1	0			2	1			5	1			3	0		
SABC012																																
SABC013			0	6	1	0	5	0	0	1							5	2			3	1			2	1			1	1	1	0
SABC014	0	3	0	5			5	0					2	0	2	0	5	1			3	1			6	0					1	0
SABC015	1	0	0	5			0	1					3	0	0	1	5	1			3	1			4	2			2	0	1	0
SABC016	0	1	0	7			7	0			0	1	2	0	2	0	2	2			4	1			2	1			5	1	2	1
SABC017	1	0	0	5	1	0	6	0					3	0			6	0			5	1			4	0			2	0	1	0
SABC018			0	4			5	0					2	0	2	1	1	2			5	1			2	0			0	1		
SABC019	1	2	3	17	1	0	5	0	0	1			2	0			7	0			5	1			6	0					2	3
SABC020	1	0	0	3			0	1			1	0	1	0	0	1	6	0			3	1			6	0			3	2	1	0
SABC021			0	5	0	1	5	0	0	3	2	0	0	1			1	1			3	1			3	0			3	0	1	0
SABC022	1	0	0	6	1	0	0	1					1	0	2	0	6	0			3	1			4	1			2	0	1	0
SABC023	1	1	0	4			7	0					2	0	4	1	5	0			4	1			6	0			3	0	1	0
SABC024	1	0	0	5			6	0	0	4			2	0			4	0			2	1			5	0			2	0	2	0
SABC025			0	6			5	1					1	1	3	1	5	2			3	1	0	1	6	0			2	0	1	0
SABC026	1	1	1	3	1	0	5	0	0	1			2	0	3	2	1	1			5	1			4	0			3	0	1	0
SABC027			0	4	1	1	7	0	0	2			0	2	4	1	5	1	0	2	5	1			4	0			5	1	2	1
SABC028			0	3	1	0					0	1	2	0	4	1	5	1			4	1			3	0			2	0	1	0
SABC029			0	5	0	2	4	0	0	1			2	0			7	0			3	1			3	1			2	0		
SABC030			0	3	1	0	6	0	0	1	2	0	2	0	2	0	2	2			4	1			6	0					1	0
SABC031			8	5	1	0	5	0					1	0			5	1			4	1			2	0			2	0	1	0
SABC032	1	0	0	6	1	0	5	1			0	1	1	0	2	1	6	0			6	1			2	0			2	0	2	1

	PKMYT1		PRKDC		RBL1		RBL2		RFC2		RFC3		RFC4		RFC5		RPA1		RPRM		RPS6KA1		SFN		SLC19A1		SMARCD2		UIMC1		WEE1	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC033			4	0			5	0					1	0			7	0			5	1			4	0					1	0
SABC034	1	0	0	5									1	0	2	1	6	0			5	1			6	0			3	1	1	0
SABC035			0	4	1	0	6	1					3	0	2	0	1	1			4	1			6	0					1	0
SABC036	1	0	0	4	1	0	6	0					1	0	2	0	2	0			7	1			2	1			2	0	2	0
SABC037			0	4									2	0	4	1	2	0			6	2			6	0			2	0	1	0
SABC038	1	0	0	3	1	0	6	0	0	1	2	0	2	0			6	1			6	1			6	0			3	0	2	0
SABC039	1	0	0	4			7	0			0	1	2	0	2	0	5	0			6	1			6	0			5	1	1	0
SABC040			0	7			6	0	0	2	0	1			1	0	4	1			4	1			6	0			2	0	2	0
SABC041	0	1	0	5			5	1					2	0	1	1	2	0			3	1			2	0			2	0	1	0
SABC042	1	0	0	3			7	0	0	2			2	1	2	0	5	0			5	1			6	0			3	0	0	1
SABC043	1	0	0	5			7	1			0	1	2	0	0	1	5	1	0	1	3	1	0	1	6	1			2	0	1	0
SABC044			0	6	1	0	6	0					2	1			5	0			5	1			2	0			3	0	2	0
SABC045	1	0	0	4			5	0					3	0	0	1	2	2			2	2	0	1	6	0			3	1	1	0
SABC046																																
SABC047	1	0	0	6			6	0					2	0	2	0	4	1			6	2			6	0			3	1		
SABC048			0	4			6	0			0	1	0	2	4	1	4	1			3	1			2	0			2	0	1	0
SABC049			0	6			5	0					2	0	5	1	2	0	0	1	2	1			2	2			4	2	2	0
SABC050			0	5	0	1	5	0			1	2	2	0			5	1			5	1			4	0			2	0	1	0
SABC051	1	0	0	5					0	2			3	0	1	1	5	0			3	1			2	1			2	0		
SABC052			0	3	0	1	5	0					2	0			2	1			2	1			4	0			3	0	1	0
SABC053	1	0	7	5	1	0	6	0					3	0			6	0			6	1			4	1			2	0		
SABC054	1	0	0	3			6	1			1	2	3	0			5	1			6	1			6	0			0	1	1	0
SABC055	1	1	0	3			6	0					2	0	0	1	5	0			4	2			6	0			2	1	1	0
SABC056	0	2	0	3			2	1					1	0	2	0					3	1	0	1	4	1			2	1	1	0
SABC057			0	3	1	1					2	0	3	0			5	1			4	1							2	0		
SABC058	1	0	0	5			7	0					3	0			2	1			4	1			4	0			3	1	1	0
SABC059			0	3			6	0					2	0			5	0			3	1			6	0					1	0
SABC060	1	0	0	4	1	0	6	1	0	1	2	0	1	0	1	1	2	0			5	2			6	0			3	0	1	0
SABC061			0	3			5	0					2	0	4	2	5	0			4	1			2	2			3	0	1	0
SABC062	1	0	0	4			8	0					3	0	2	0	5	0			4	1	0	1	4	0	0	1	0	1	2	0
SABC063	1	0	0	4			5	0	0	2			3	0	2	1	6	1			7	1	0	1	6	0			3	0	0	1
SABC064	1	0	0	3			7	0			2	0	2	0	4	1	8	1			2	1			6	0			3	0	1	0

	PKMYT1		PRKDC		RBL1		RBL2		RFC2		RFC3		RFC4		RFC5		RPA1		RPRM		RPS6KA1		SFN		SLC19A1		SMARCD2		UIMC1		WEE1	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC065	1	0	0	4			6	0					1	0	3	1	1	0			5	1			6	1					2	0
SABC066	1	0	0	3			7	0			2	0	1	0	2	0	2	0			3	2			6	0			2	1	1	0
SABC067	1	0	0	3			6	0					1	0	0	2	2	0			3	1			4	0					1	0
SABC068			0	4			6	0			2	0	1	0	5	1	4	1			2	4			2	0			2	0	2	0
SABC069			0	3			5	0	0	1			2	0	2	0	6	0			6	1			6	0			3	0	1	0
SABC070	1	0	0	4					0	1	0	1	2	0	2	1	5	0			6	1			6	0			3	1		
SABC071			8	5	1	0	6	0					2	0	4	1	4	0			5	1			2	0					1	1
SABC072	1	0	0	3	1	0					2	0	2	0	0	1	5	0			4	1			2	0			3	0	2	0
SABC073	1	1	0	3	1	0							2	0	4	1	2	1			5	1			6	0			2	0	2	0
SABC074	1	0	0	3			6	0			1	2			2	0	5	0			5	1			6	0			5	2	1	0
SABC075	1	0	0	5			7	0			0	1	2	0			5	1			8	1			6	0			3	0	1	0
SABC076	1	0	0	4			6	0					3	0			4	2			5	1			2	0			2	0	2	0
SABC077	1	0	0	4	1	0	6	0									5	1			6	1			6	0			3	0	1	0
SABC078	1	0	0	3			6	0	0	1			2	0	1	1	2	0			3	1			6	0			2	0	1	0
SABC079			0	4			5	0					2	0	0	1	6	0			3	2			2	0			2	0	1	0
SABC080	1	0	0	6			5	0	0	1			1	1	2	1	6	0			3	1			6	0					1	0
SABC081			0	6							1	2	2	0	1	1	5	3			2	1			6	0			2	0	1	1
SABC082	1	0	0	4			7	0	0	1			1	0	2	0	5	0			4	1			6	1			0	1		
SABC083	1	0	8	6	1	0	6	0	0	2	0	1	2	0	2	0	2	4			3	1			6	0			3	0	2	0
SABC084	1	0	0	3			7	0					2	0	0	1	7	1			7	1			4	0					1	0
SABC085	0	1	0	3									2	0	3	0	4	0			4	1			2	0			3	0		
SABC086	1	0	0	3			7	0	0	1			3	0	2	1	2	1			7	1			7	0			3	1	1	0
SABC087	1	0	0	3	1	0	0	1					2	0	2	1	8	1			2	1			2	0			3	0	1	0
SABC088	1	0	0	3			7	1					1	0	4	3	2	2			4	1			6	0						
SABC089			0	4	0	1	5	0					2	1	2	0	1	0			4	2			1	0			3	1	2	0
SABC090			0	5			6	0			2	0	2	0			1	0			3	1			3	0			2	0	1	0
SABC091			0	5	1	0	5	1					2	0	2	0	1	0			4	1			3	0			0	1	1	0
SABC092	1	0	0	3	1	0	4	0			2	1	2	0			2	1			3	2			6	1			2	0	1	1
SABC093	1	0	0	5									3	0	2	0	5	1			4	1	0	1	6	0					2	0
SABC094	1	1	0	3			5	0					2	0	2	1	5	1			3	1			4	1			2	1	1	0
SABC095	1	0	0	4			7	0	0	1			2	0	4	2	2	1			5	1	0	2	6	0			2	0	1	0
SABC096	1	0	0	4	1	0	7	0					3	0	2	1	5	1			5	1			6	1					1	1

	PKMYT1		PRKDC		RBL1		RBL2		RFC2		RFC3		RFC4		RFC5		RPA1		RPRM		RPS6KA1		SFN		SLC19A1		SMARCD2		UIMC1		WEE1	
MAF	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%	>5%	<5%
SABC097	1	0	0	4			7	0			0	1	2	0	2	1	6	0			4	1			4	1			2	0	1	1
SABC098			0	6	1	1	6	0	0	2	2	0	2	1	2	0	0	1			5	1			2	1			2	0	1	0
SABC099	1	1	0	4	0	1	5	0	0	1	0	1	1	0			5	0			3	2			4	1			2	0	1	0
SABC100			0	4			4	0	0	1			2	0	2	1	1	0			2	1	0	1	4	1			2	0	1	0
SABC101			0	3			6	0					2	1			4	0	0	1	5	1			6	0			5	0	1	0
SABC102			0	7	1	0	5	0			1	2	3	0	2	0	6	0			5	1			5	0			2	0		
SABC103	1	1	0	3			6	0					2	0	0	1	5	0			4	2			6	0			2	1	1	0
SABC104	1	0	0	4	1	1	6	1									2	1			4	1			6	2			1	0	2	0
SABC105			0	4	1	0	5	0					2	0	1	0	4	1			3	1			3	0					1	0
SABC106			0	3	1	0	6	0	0	2			2	0	4	1	2	0			6	1			4	0			6	1	1	0
SABC107			0	4			5	0	0	2					0	1	1	0			1	1			4	0			3	2	1	0
SABC108	1	0	0	5			7	0					2	0	4	2	6	0			6	1			6	1			2	0	1	0
SABC109	0	1	0	4			4	0					1	0	0	1	1	0			3	1			6	0			5	1	1	0
SABC110	1	0	0	4									0	2	2	2	5	1			3	2			6	0			2	0	2	0
SABC111			8	4	1	0					0	1	1	0	3	0	5	2			5	1			3	0			2	0	1	0
SABC112	1	0	0	5	0	1	6	1					1	0	0	2	2	1			6	1			4	0			3	1	1	1
SABC113	1	0	0	5			5	0	0	1	2	0	2	0	2	0	2	1			3	2			6	1			3	0	1	0
SABC114	1	1	0	3			0	1			0	1	2	0			2	1			5	1	0	1	2	0			0	1	1	0
SABC115	1	1	0	3			5	0					1	0			5	2			4	1			6	0			3	0	1	1
SABC116			0	4			6	1					3	0			6	0			2	1	0	1	6	1			2	0	1	0
SABC117	1	0	0	3			5	0					3	0	0	1	5	2			6	1			4	1					2	0
SABC118			0	5	0	1	6	0			2	0	1	0	3	1	4	1			3	1			4	0			2	0	1	0
SABC119			0	7			6	1					3	0	2	0	3	0			3	1			3	0					2	0
SABC120			0	5	1	0	5	0					2	0	1	0	5	2			1	1			1	1			3	2		
SABC121			0	3			6	0					2	0	2	0	2	0			4	1			4	1			1	1	1	0
SABC122			0	7			6	0					1	0	0	1	4	1			3	2			4	0			3	0	2	0
SABC123			0	3	0	1							3	0			2	0			4	1			2	0			3	0	1	0
SABC124	1	0	0	3			3	1							1	0	3	0					0	1	2	0			1	1		
SABC125			0	5	1	0	5	0					2	0			1	0			2	1			4	1			2	0	1	0
SABC126			0	4	1	0	6	1					1	0			5	0			5	1			5	0			3	0	2	0
SABC127			0	3	0	1	6	0	0	1			0	1	2	1	4	2			2	1			5	0			3	0	2	0
SABC128			0	4	1	1			0	2			3	0			6	0			5	3			2	1			3	1	1	1
SABC129			3	9	1	0	6	0					1	0			4	1			3	1			4	0			6	0	1	0
SABC130	1	0	0	4	1	0	0	1			2	0			3	0	5	0			6	1	0	1	4	0			2	0	1	1
SABC131			0	4	1	0	6	0	0	1					2	0	4	2			3	1			2	2			0	3	1	0
SABC132			0	5	1	0	6	1					2	0	2	0	1	0			4	2			3	0						
SABC133	1	0	0	5			5	0			1	0	2	0			2	0			5	1			6	0			2	0		

Table H.4: Overall results of individual sequencing data, with total number of variants and rare variants indicated. Variants broken down into diagnostic and discovery genes

Patient ID	Total Number of Variants	Number of diagnostic variants	Number of discovery variants	<5% frequency in general population
SABC001	145	80	65	29
SABC002	118	67	51	27
SABC003	107	54	53	19
SABC004	73	41	32	14
SABC005	118	67	51	28
SABC006	146	75	73	31
SABC007	109	51	58	30
SABC008	111	50	61	29
SABC009	128	73	55	22
SABC010	129	78	51	34
SABC011	112	51	61	29
SABC012	Failed			
SABC013	125	68	57	34
SABC014	125	64	61	30
SABC015	111	59	52	33
SABC016	119	44	75	28
SABC017	142	72	70	30
SABC018	113	58	55	33
SABC019	163	75	88	46
SABC020	120	59	61	29
SABC021	124	69	55	36
SABC022	113	56	57	28
SABC023	138	67	71	34
SABC024	118	58	60	35
SABC025	126	58	68	34
SABC026	120	53	67	39
SABC027	149	77	72	38
SABC028	107	54	53	30
SABC029	122	61	61	34
SABC030	122	58	64	31
SABC031	113	64	49	20
SABC032	133	70	63	31
SABC033	130	74	56	20
SABC034	118	60	58	26
SABC035	112	54	58	23
SABC036	118	61	57	20
SABC037	124	68	56	33
SABC038	164	86	78	29
SABC039	145	74	71	27
SABC040	132	70	62	30
SABC041	113	61	52	32
SABC042	130	60	70	28
SABC043	145	77	68	34
SABC044	123	68	55	29
SABC045	131	76	55	33
SABC046	Failed			

Patient ID	Total Number of Variants	# Diagnostic variants	# Discovery variants	<5% frequency in general population
SABC047	127	63	64	28
SABC048	124	67	57	35
SABC049	110	47	63	31
SABC050	110	46	64	26
SABC051	117	61	56	33
SABC052	108	59	49	29
SABC053	117	54	63	25
SABC054	143	79	64	25
SABC055	137	72	65	34
SABC056	111	62	49	25
SABC057	122	72	50	30
SABC058	134	74	60	28
SABC059	115	58	57	21
SABC060	144	76	68	32
SABC061	125	66	59	21
SABC062	118	62	56	25
SABC063	144	77	67	28
SABC064	128	60	68	24
SABC065	142	83	59	27
SABC066	127	68	59	29
SABC067	136	82	54	23
SABC068	132	65	67	30
SABC069	126	68	58	18
SABC070	113	59	54	30
SABC071	126	60	66	34
SABC072	112	58	54	25
SABC073	110	49	61	25
SABC074	130	59	71	28
SABC075	133	67	66	31
SABC076	105	53	52	23
SABC077	127	72	55	28
SABC078	146	84	62	32
SABC079	135	83	52	24
SABC080	142	78	64	25
SABC081	133	65	68	37
SABC082	124	61	63	31
SABC083	147	63	84	35
SABC084	117	57	60	26
SABC085	100	46	54	27
SABC086	137	68	69	26
SABC087	116	63	53	22
SABC088	132	75	57	31
SABC089	107	49	58	32
SABC090	90	44	46	22
SABC091	108	59	49	27
SABC092	127	64	63	29
SABC093	112	53	59	26
SABC094	119	62	57	27
SABC095	144	76	68	36
SABC096	123	56	67	27

Patient ID	Total Number of Variants	# Diagnostic variants	# Discovery variants	<5% frequency in general population
SABC097	122	59	63	31
SABC098	131	61	70	39
SABC099	124	67	57	33
SABC100	133	80	53	29
SABC101	118	54	64	20
SABC102	139	76	63	22
SABC103	135	72	63	31
SABC104	143	87	56	34
SABC105	95	43	52	26
SABC106	123	54	69	29
SABC107	115	72	43	38
SABC108	134	63	73	35
SABC109	130	68	62	27
SABC110	114	59	55	34
SABC111	127	62	65	32
SABC112	168	97	71	47
SABC113	127	62	65	26
SABC114	122	70	52	28
SABC115	147	91	56	27
SABC116	115	52	63	30
SABC117	121	59	62	29
SABC118	110	53	57	27
SABC119	103	51	52	24
SABC120	96	44	52	25
SABC121	111	67	44	22
SABC122	124	67	57	30
SABC123	114	66	48	26
SABC124	65	32	33	18
SABC125	113	56	57	25
SABC126	111	59	52	19
SABC127	130	66	64	28
SABC128	126	69	57	32
SABC129	135	62	73	31
SABC130	137	74	63	30
SABC131	140	85	55	46
SABC132	111	61	50	23
SABC133	119	63	56	27

Appendix I: Plasmid maps

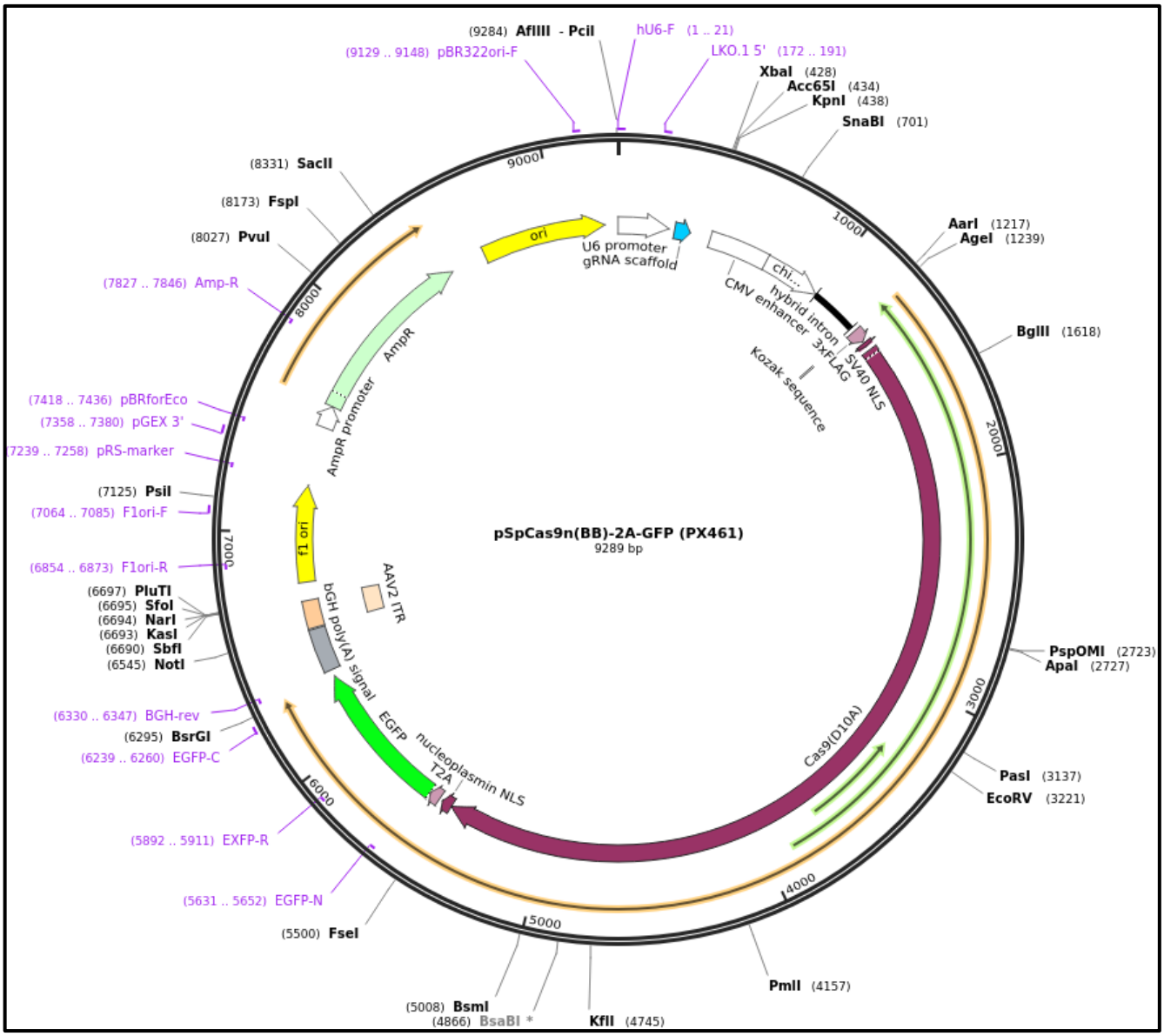


Figure I.1: Plasmid Map of PX461 Plasmid from Addgene. This plasmid encodes a D10A mutated Cas9 nuclease (as indicated by the purple arrow) which results in a cut in a single strand of DNA. Nickase plasmids consist of a pair of plasmids with a 20bp guide RNA (sgRNAs). Annealed oligonucleotides are inserted in the gRNA scaffold region (indicated by the blue arrow) downstream of the U6 promotor. This plasmid allows for selection through fluorescence of Green Fluorescence Protein (GFP) in successfully transfected cells, plasmids have ampicillin resistance.

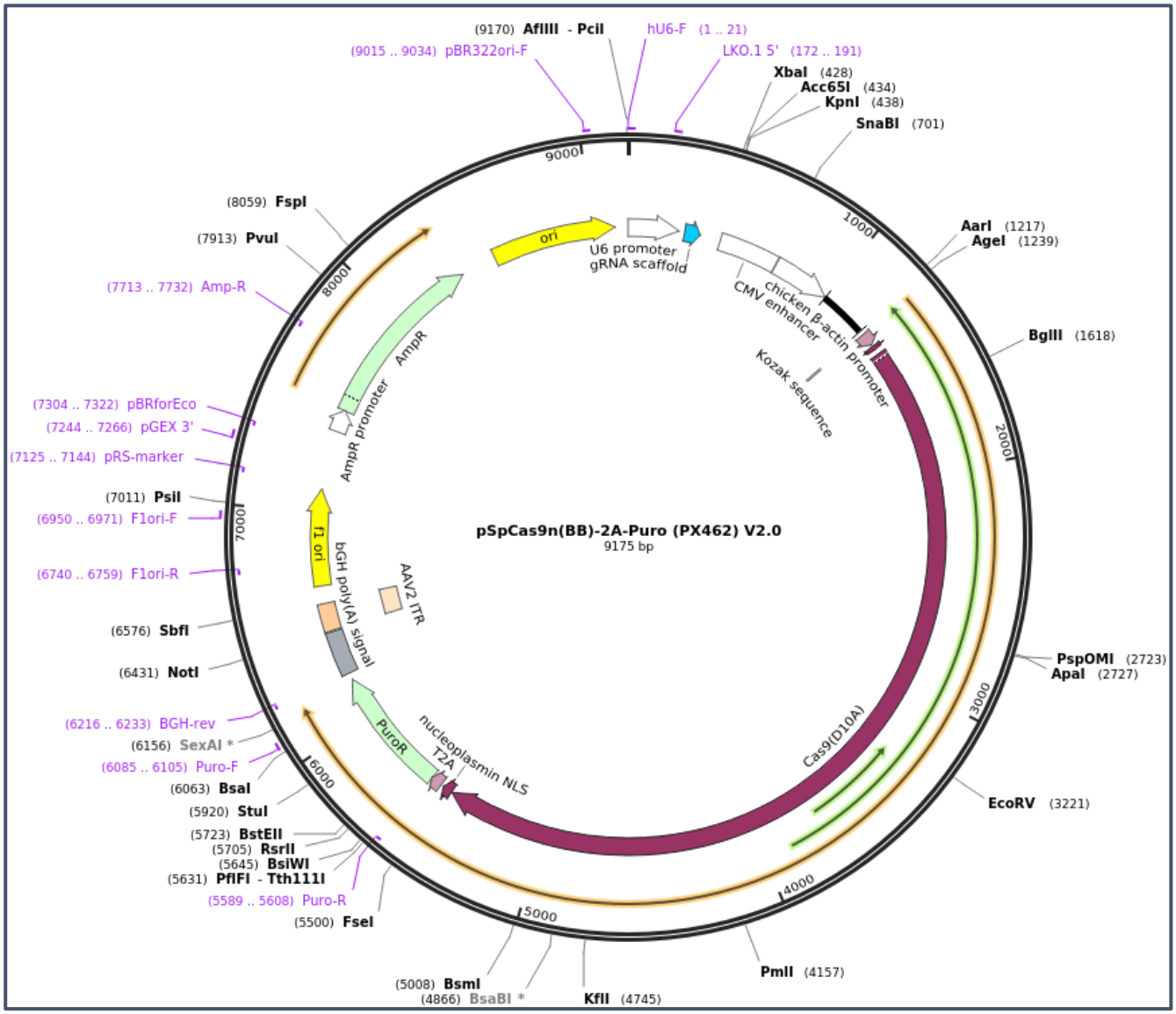


Figure I.2 Plasmid Map of PX462v2.0 This plasmid encodes a D10A mutated Cas9 nuclease (as indicated by the purple arrow) which results in a cut in a single strand of DNA. Nickase plasmids consist of a pair of plasmids with a 20bp guide RNA (sgRNAs). Annealed oligonucleotides are inserted in the gRNA scaffold region (indicated by the blue arrow) downstream of the U6 promotor. This plasmid allows for selection through resistance to puromycin in successfully transfected cells, plasmids have ampicillin resistance.

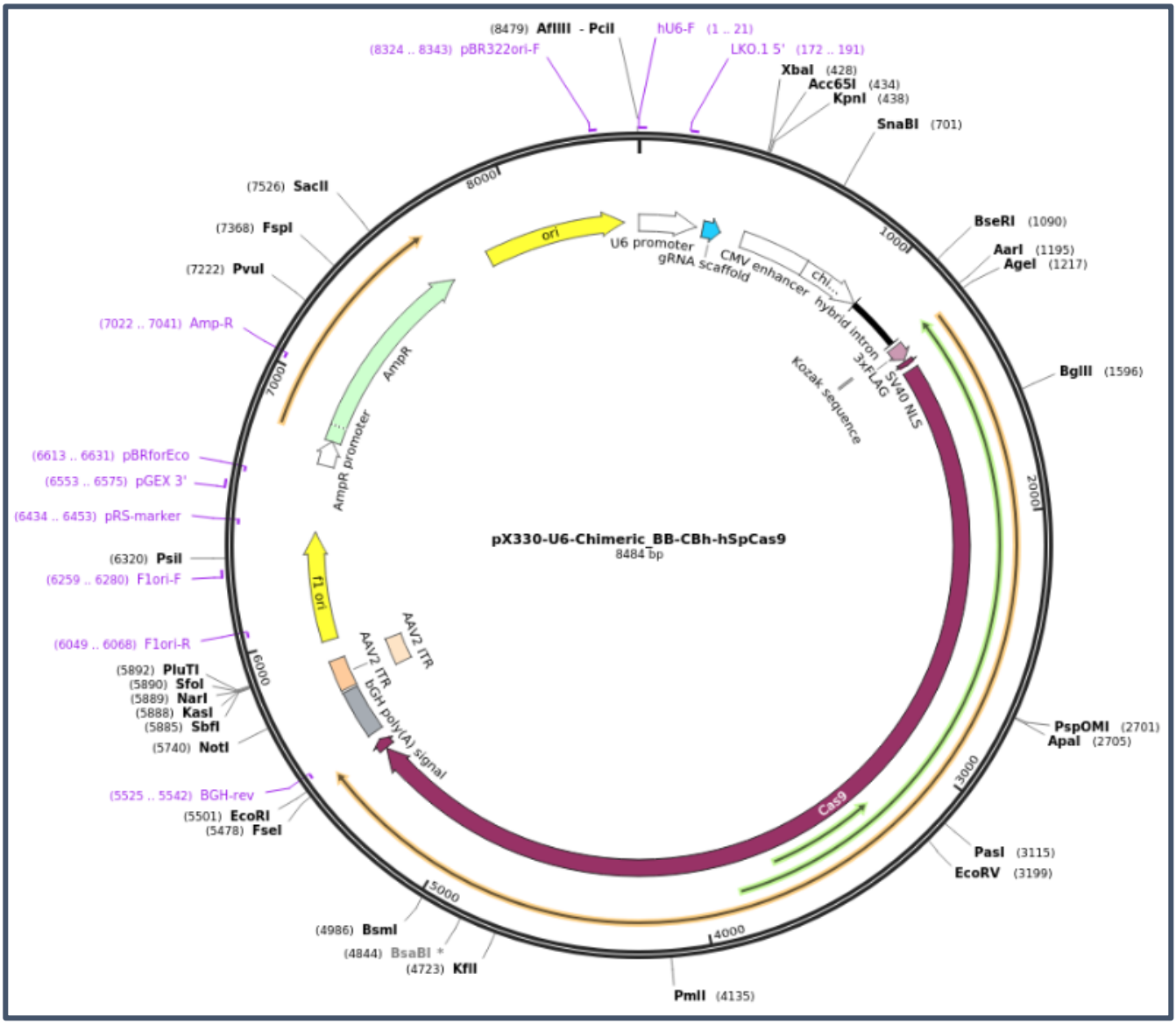


Figure I.3: Plasmid Map of PX330. This plasmid encodes Cas9 nuclease (as indicated by the purple arrow) which results in a double stranded cut of DNA. Annealed oligonucleotides are inserted in the gRNA scaffold region (indicated by the blue arrow) downstream of the U6 promotor. This plasmid has ampicillin resistance, but no means of selection for transfected cells.

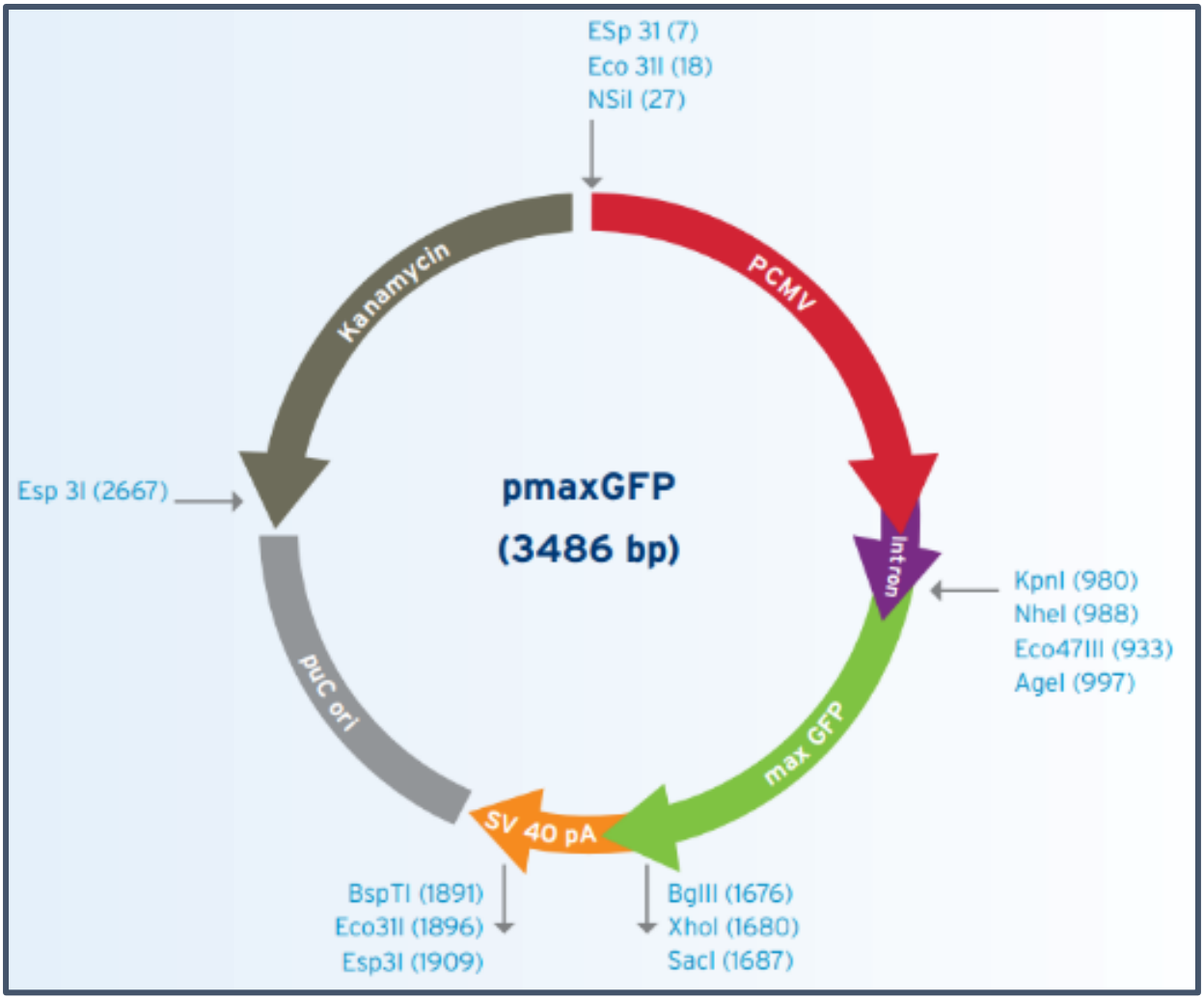


Figure I.4: Plasmid Map of pmaxGFP (Lonza) . This plasmid results in the expression of green florescence protein (GFP; as indicated by the green arrow). This plasmid has Kanamycin resistance and is used as a positive control for all transfection experiments in order to visualise successful transfection of cell lines.

Appendix J: γH2A.X nuclear foci counts

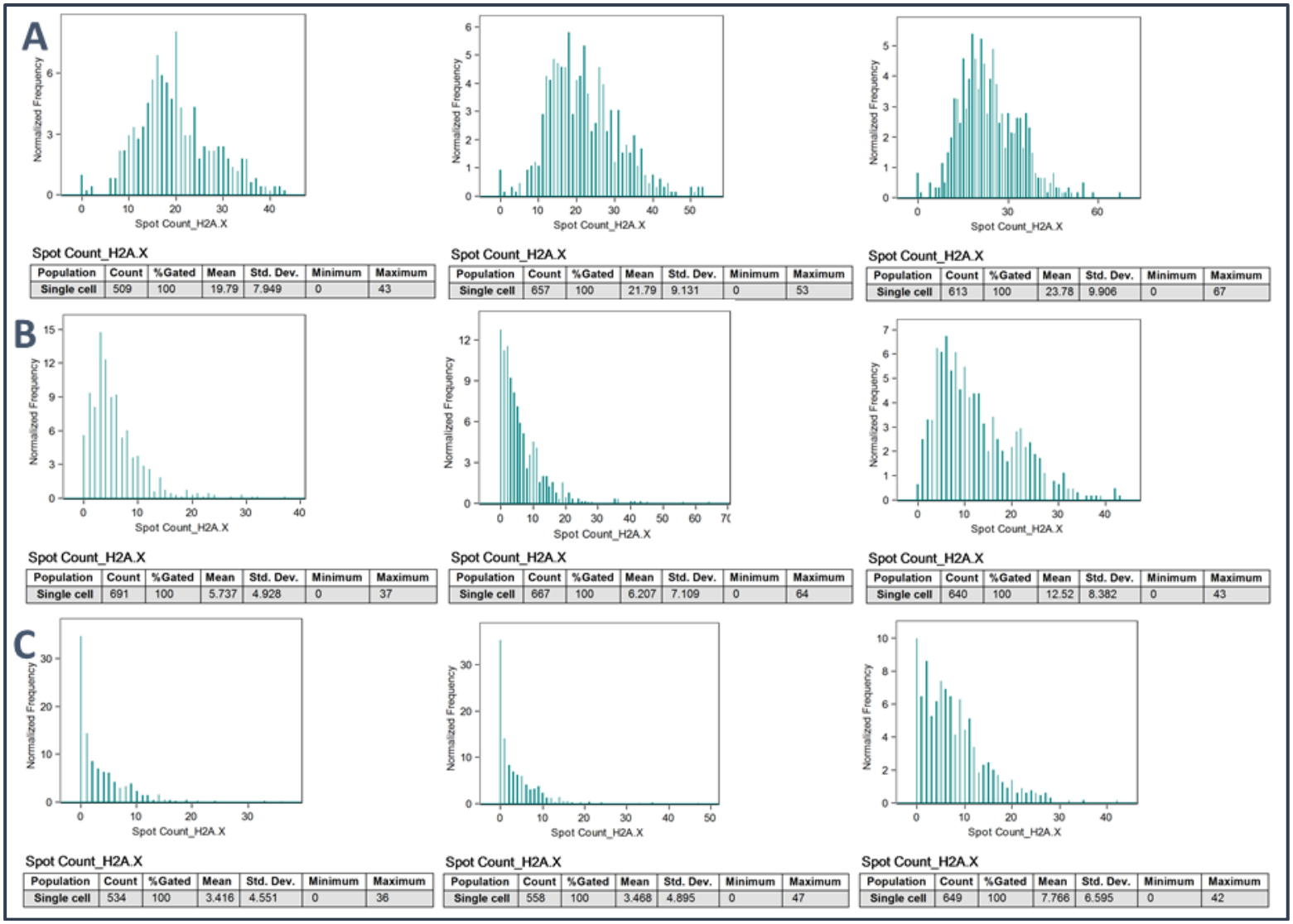


Figure J.1: HEK raw nuclear foci counts following exposure to 2 Gy ionising radiation (IR) at various time points (n=3) A. 1 hour post IR exposure; **B.** 4 hours post IR exposure; **C.** 24 hours post IR exposure. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.

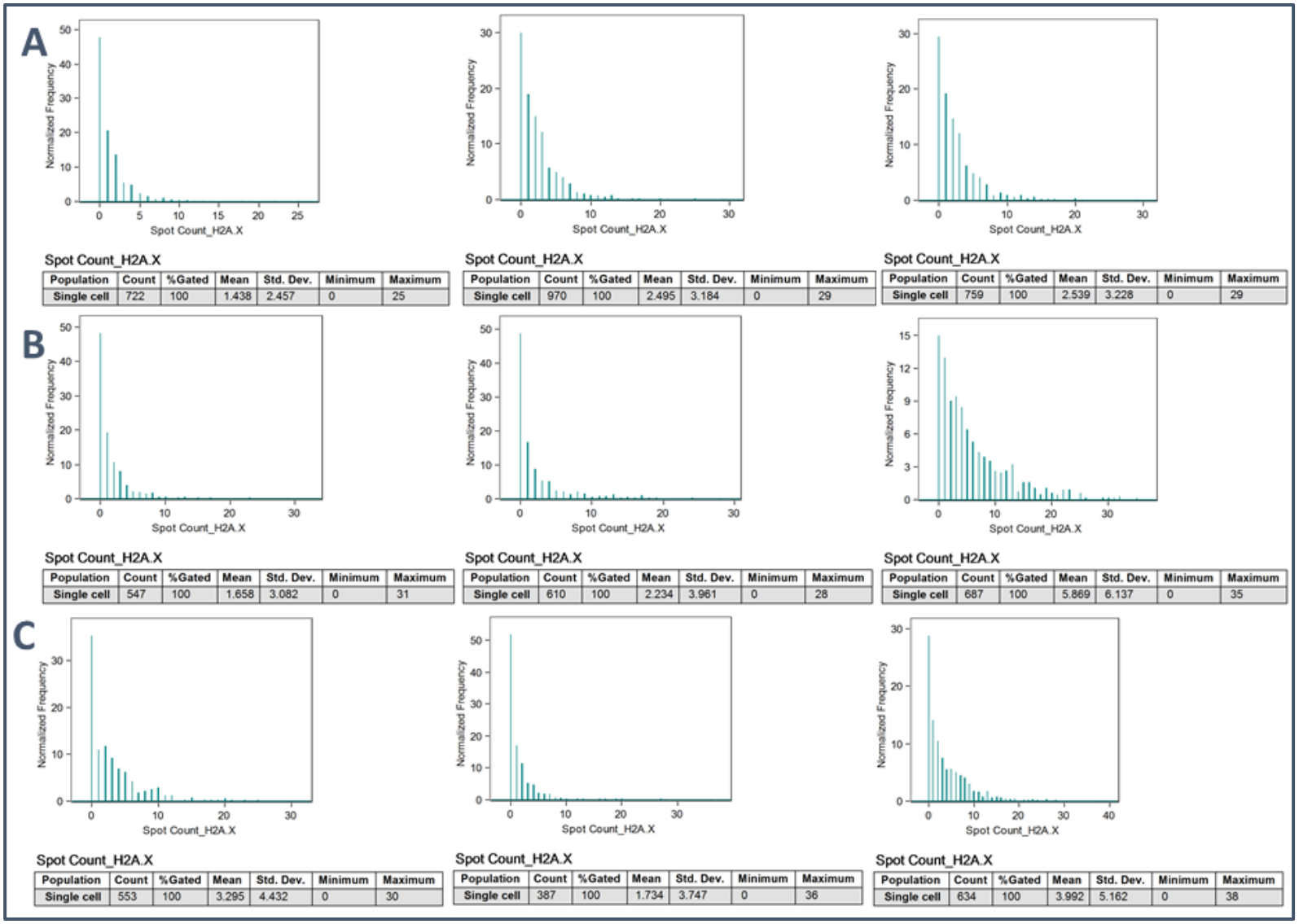


Figure J.2 HEK293 raw nuclear foci count plots following exposure to sham irradiation at various time points (n=3) A. 1 hour post sham irradiation; **B.** 4 hours post sham irradiation; **C.** 24 hours post sham irradiation. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.

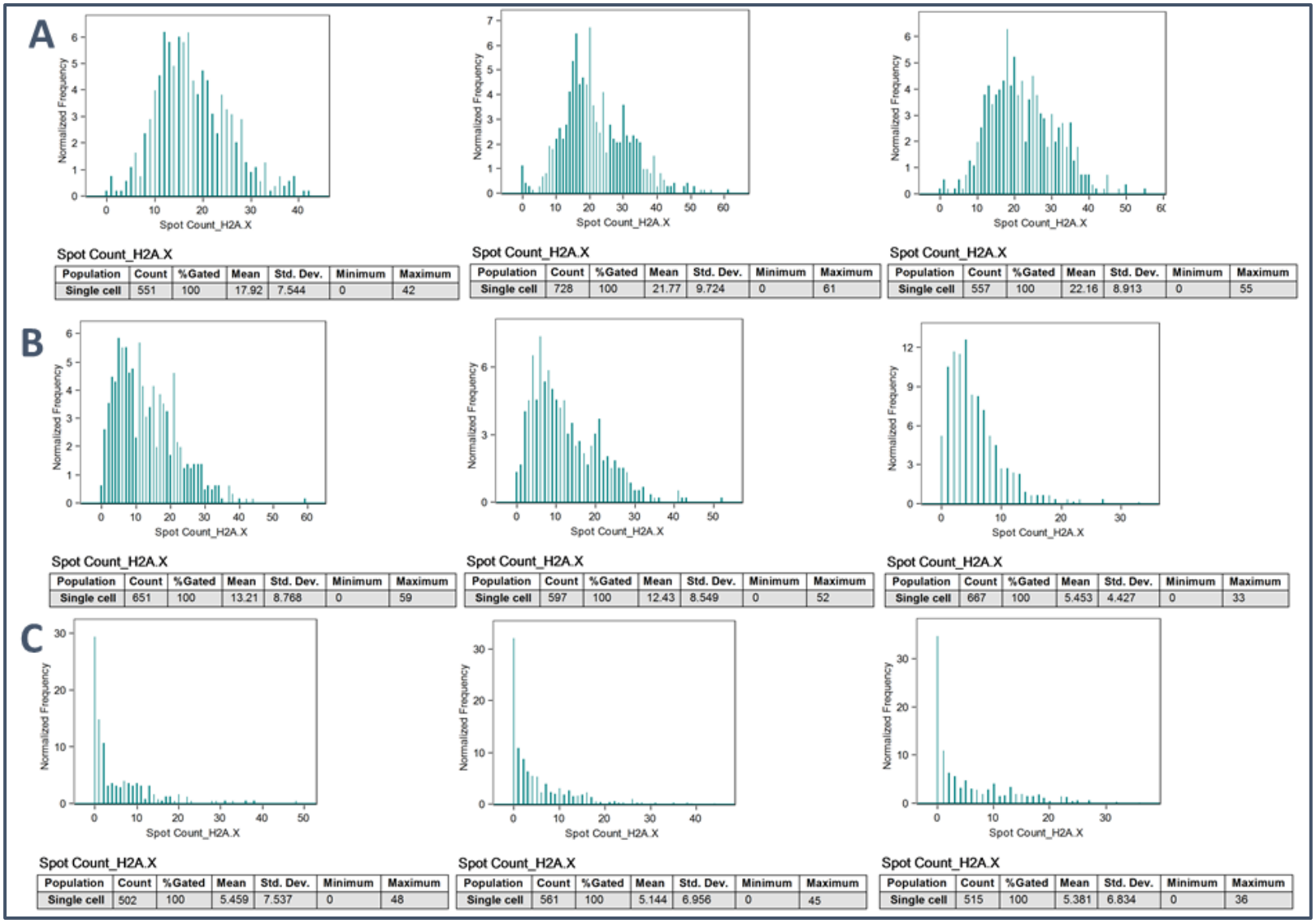


Figure J.3: PX330- (CRISPR sham) raw nuclear foci counts following exposure to 2 Gy ionising radiation (IR) at various time points (n=3) A. 1 hour post IR exposure; **B.** 4 hours post IR exposure; **C.** 24 hours post IR exposure. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.

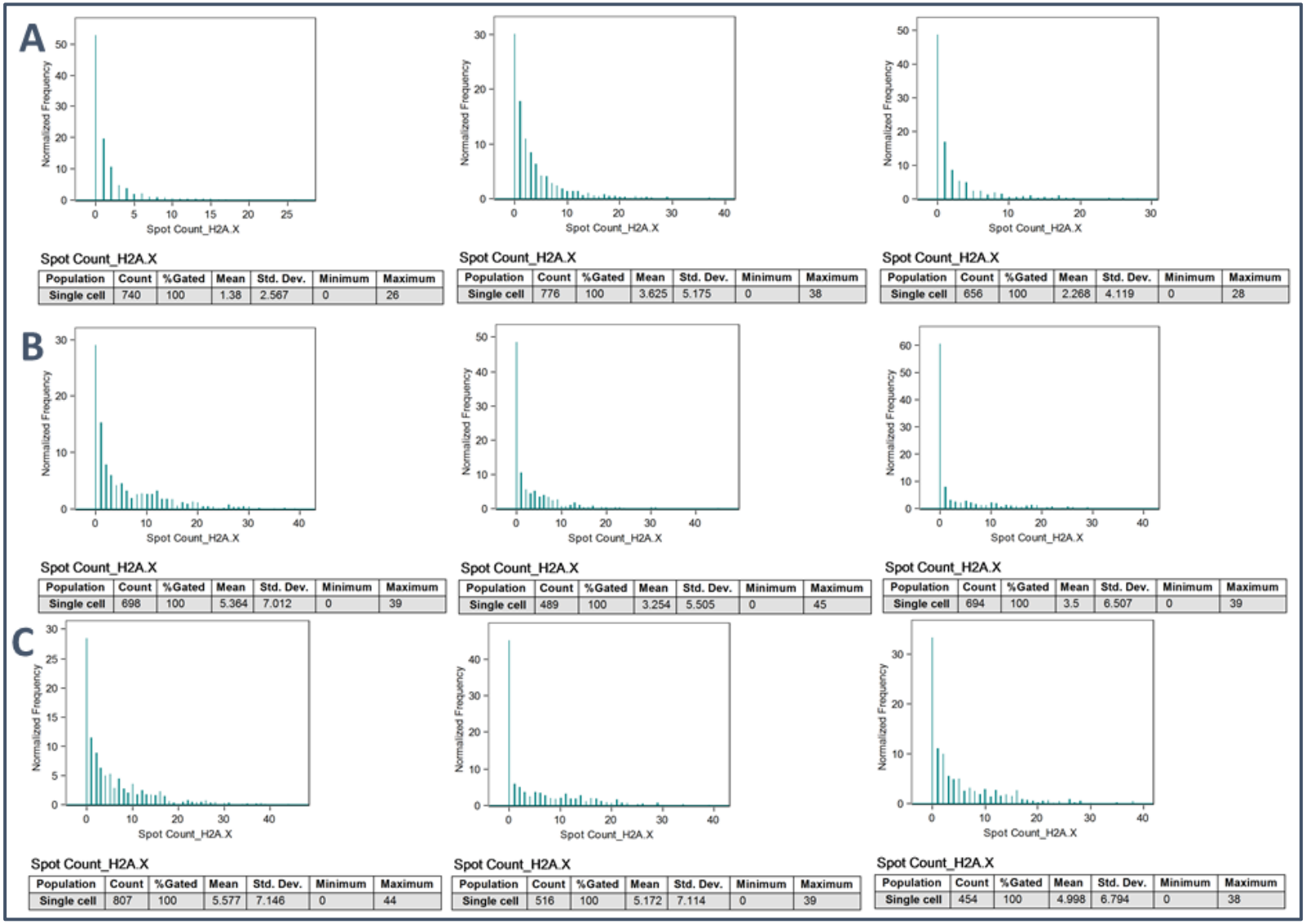


Figure J.4: PX330- (CRISPR sham) raw nuclear foci count plots following exposure to sham irradiation at various time points (n=3) A. 1 hour post sham irradiation; **B.** 4 hours post sham irradiation; **C.** 24 hours post sham irradiation. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot

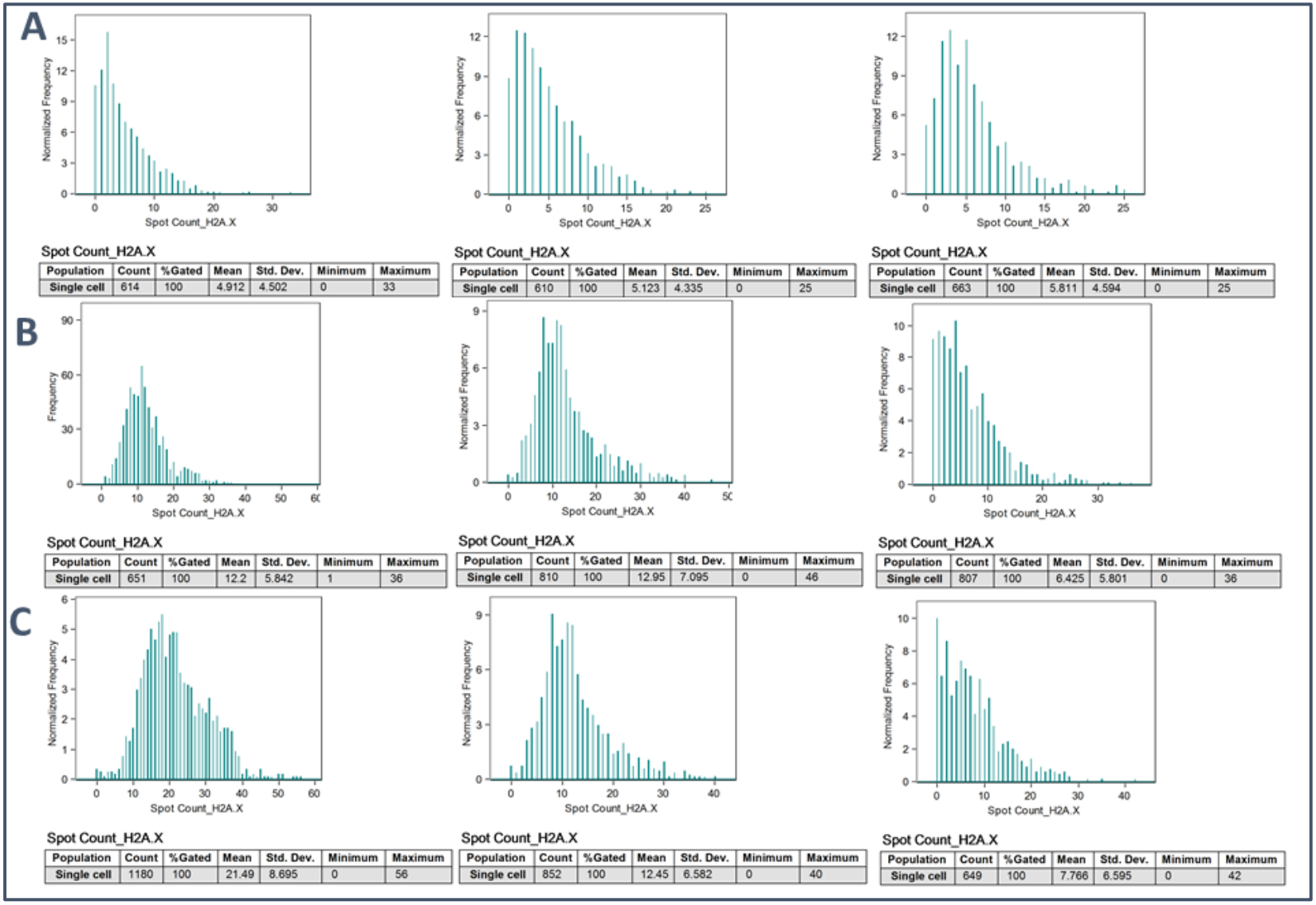
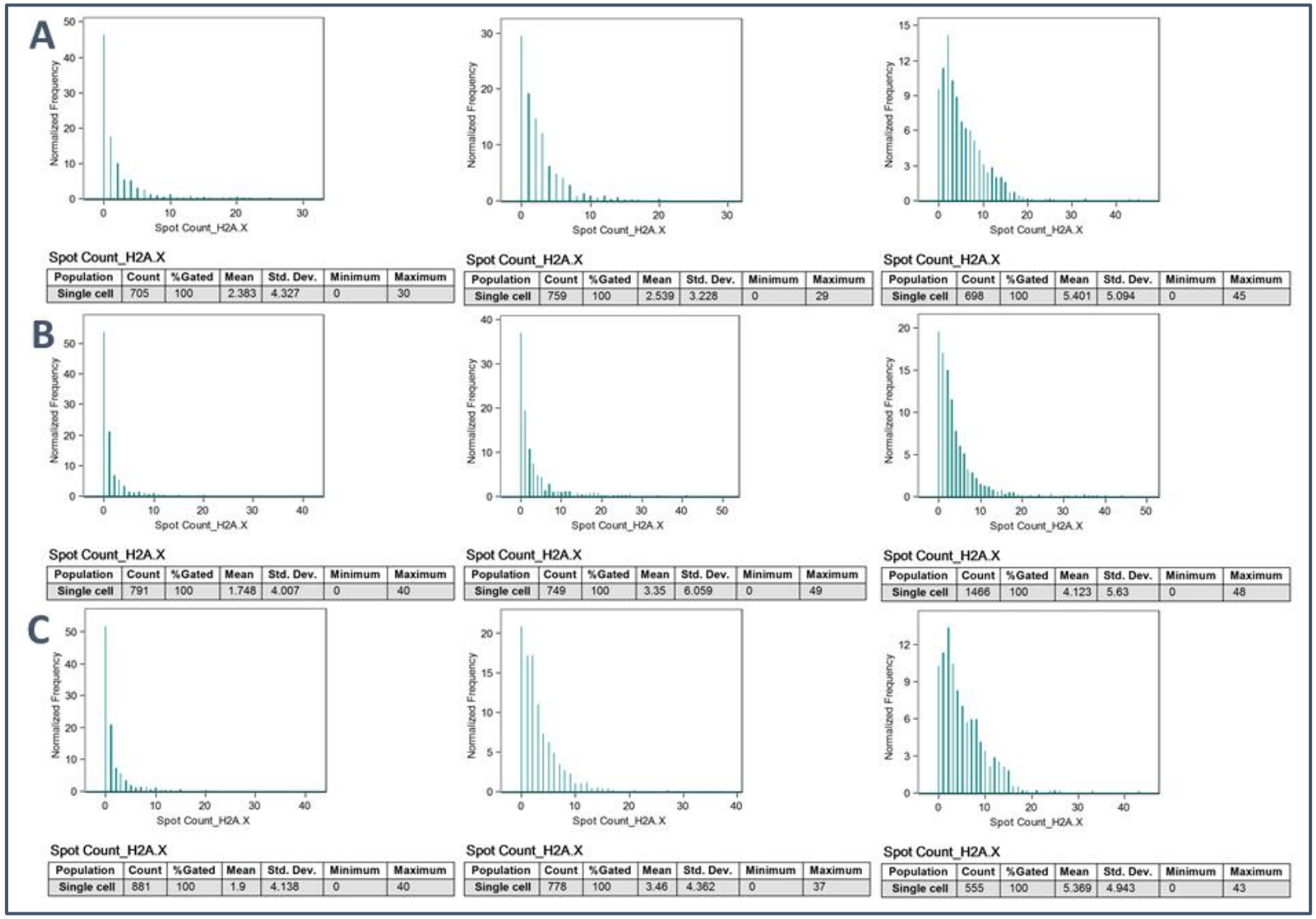


Figure J.5: e2-B1.15 (UIMC1 homozygous knockout) raw nuclear foci counts following exposure to 2 Gy ionising radiation (IR) at various time points (n=3) A. 1 hour post IR exposure; **B.** 4 hours post IR exposure; **C.** 24 hours post IR exposure. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	705	100	2.383	4.327	0	30

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	759	100	2.539	3.228	0	29

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	698	100	5.401	5.094	0	45

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	791	100	1.748	4.007	0	40

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	749	100	3.35	6.059	0	49

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	1466	100	4.123	5.63	0	48

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	881	100	1.9	4.138	0	40

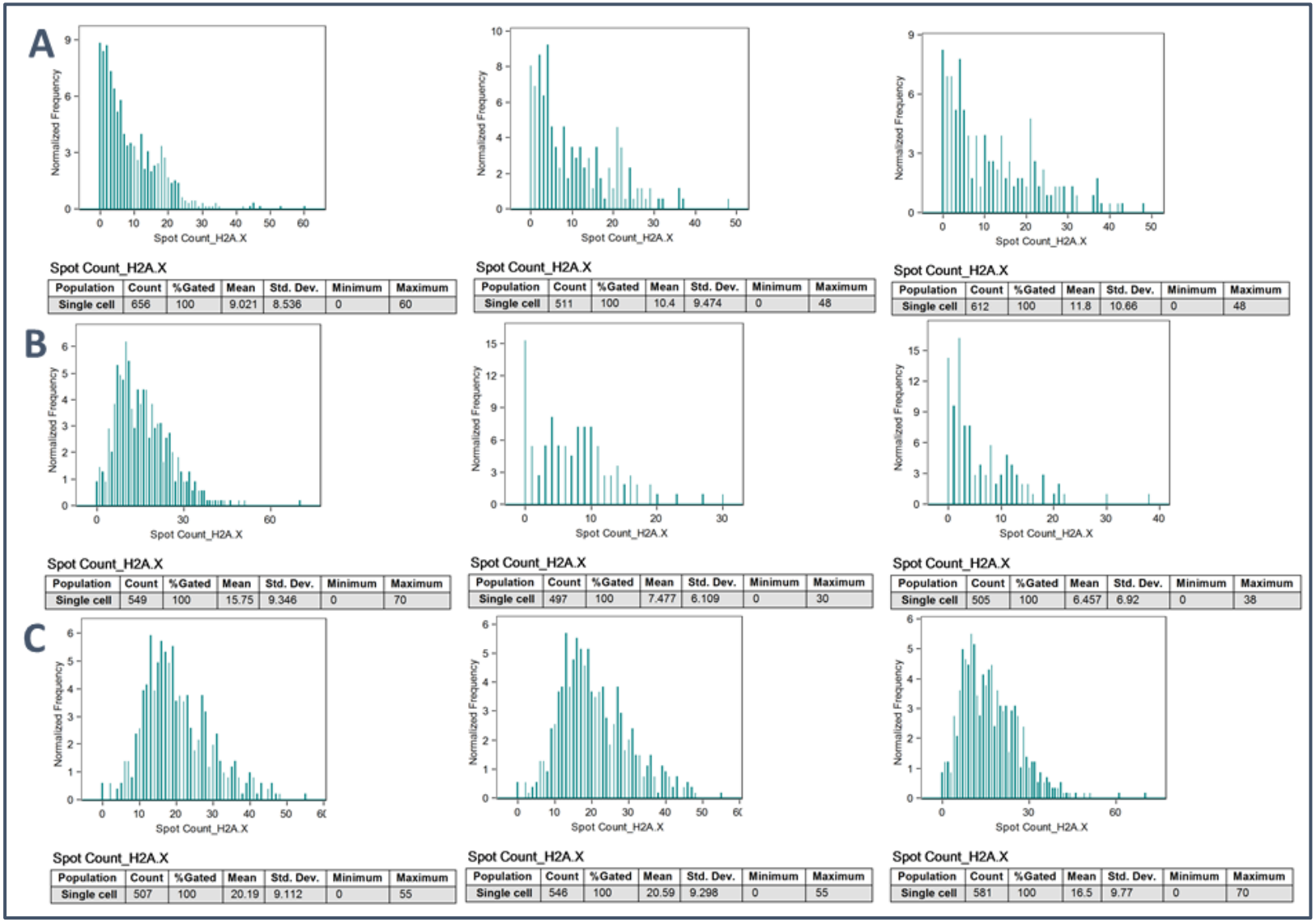
Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	778	100	3.46	4.362	0	37

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	555	100	5.369	4.943	0	43

Figure J.6: e2-B1.15 (UIMC1 homozygous knockout) raw nuclear foci count plots following exposure to sham irradiation at various time points (n=3) A. 1 hour post sham irradiation; **B.** 4 hours post sham irradiation; **C.** 24 hours post sham irradiation. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	656	100	9.021	8.536	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	511	100	10.4	9.474	0	48

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	612	100	11.8	10.66	0	48

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	549	100	15.75	9.346	0	70

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	497	100	7.477	6.109	0	30

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	505	100	6.457	6.92	0	38

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	507	100	20.19	9.112	0	55

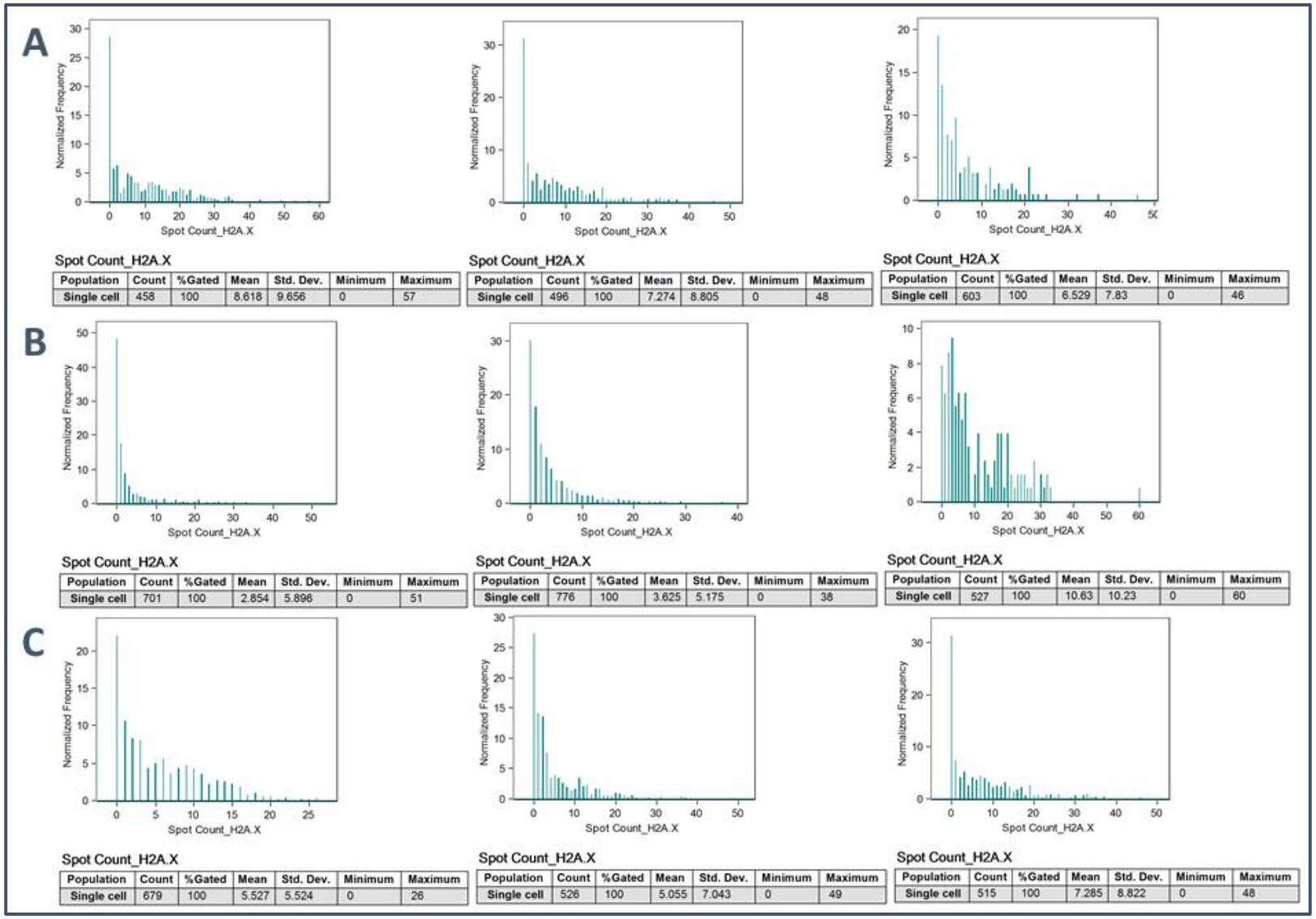
Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	546	100	20.59	9.298	0	55

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	581	100	16.5	9.77	0	70

Figure J.7: e2-B3.1 (UIMC1 heterozygous knockout) raw nuclear foci counts following exposure to 2 Gy ionising radiation (IR) at various time points (n=3) A. 1 hour post IR exposure; **B.** 4 hours post IR exposure; **C.** 24 hours post IR exposure. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	458	100	8.618	9.656	0	57

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	496	100	7.274	8.805	0	48

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	603	100	6.529	7.83	0	46

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	701	100	2.854	5.896	0	51

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	776	100	3.625	5.175	0	38

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	527	100	10.63	10.23	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	679	100	5.527	5.524	0	26

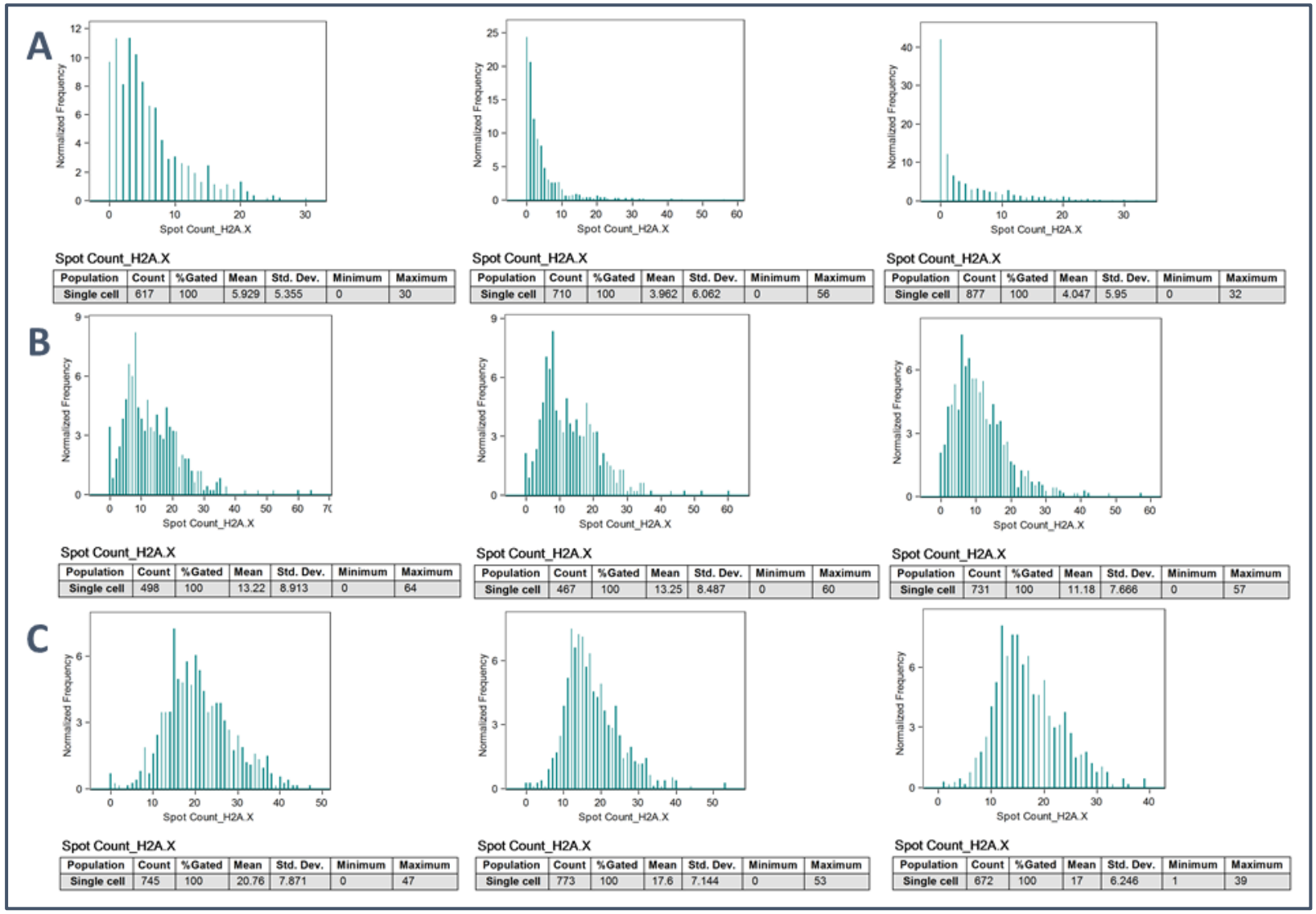
Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	526	100	5.055	7.043	0	49

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	515	100	7.285	8.822	0	48

Figure J.8: e2-B3.1 (UIMC1 heterozygous knockout) raw nuclear foci count plots following exposure to sham irradiation at various time points (n=3) A. 1 hour post sham irradiation; **B.** 4 hours post sham irradiation; **C.** 24 hours post sham irradiation. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	617	100	5.929	5.355	0	30

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	710	100	3.962	6.062	0	56

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	877	100	4.047	5.95	0	32

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	498	100	13.22	8.913	0	64

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	467	100	13.25	8.487	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	731	100	11.18	7.666	0	57

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	745	100	20.76	7.871	0	47

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	773	100	17.6	7.144	0	53

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	672	100	17	6.246	1	39

Figure J.9: e13-KO1 (UIMC1 with 1 AA deletion in ZFN) raw nuclear foci counts following exposure to 2 Gy ionising radiation (IR) at various time points (n=3) A. 1 hour post IR exposure; **B.** 4 hours post IR exposure; **C.** 24 hours post IR exposure. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot

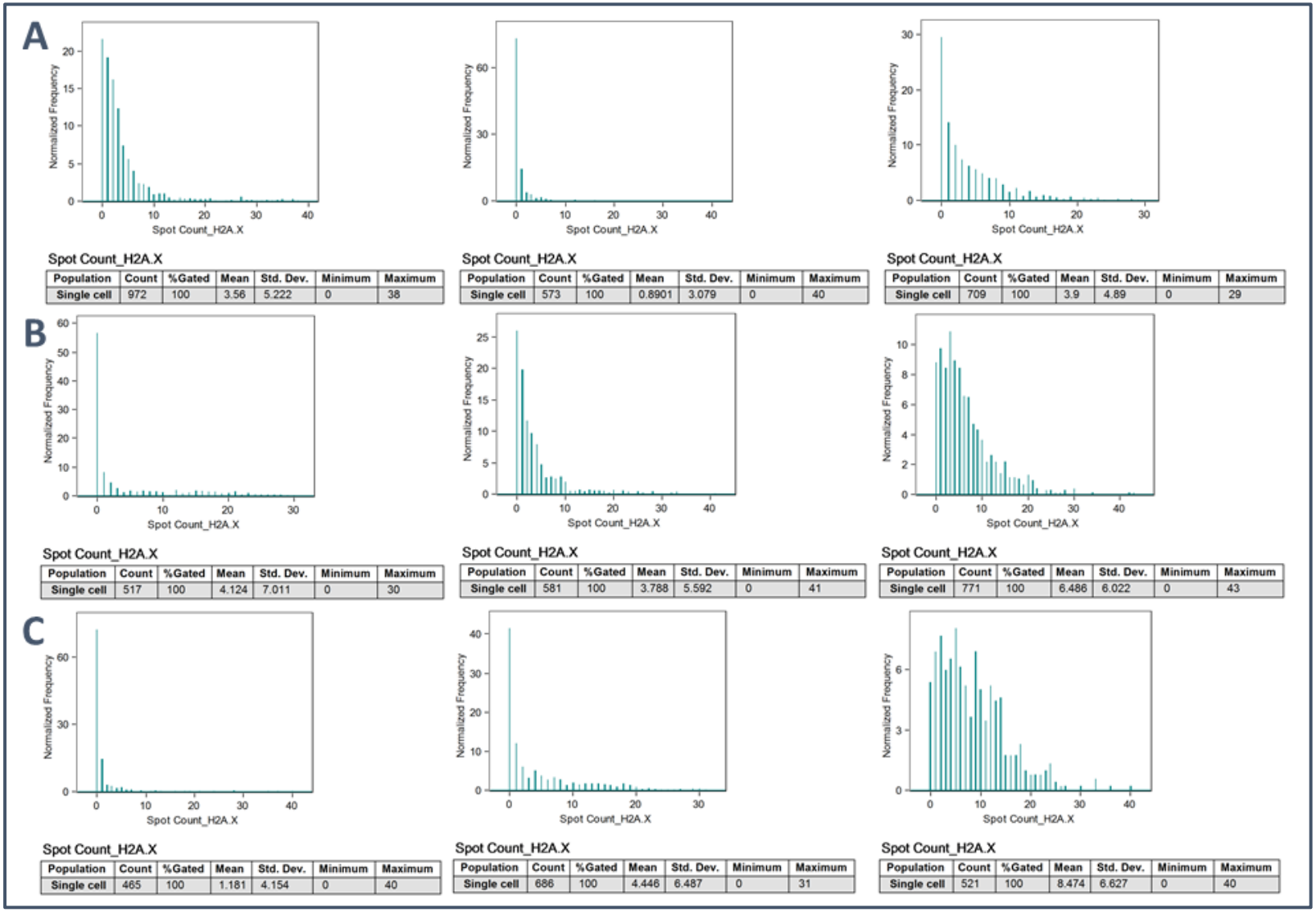
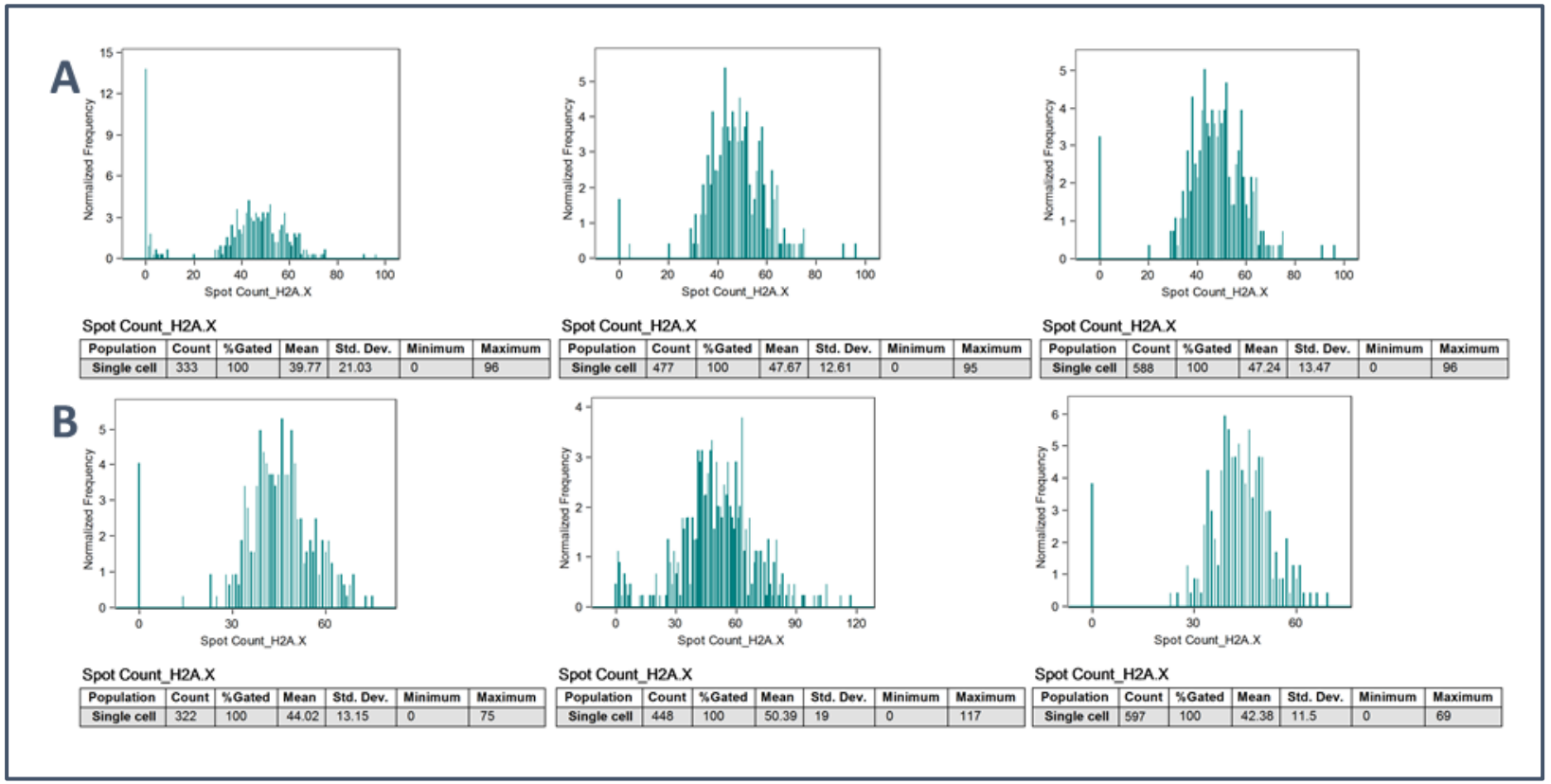


Figure J.10: e13-KO1 (UIMC1 with 1 AA deletion in ZFN) raw nuclear foci count plots following exposure to sham irradiation at various time points (n=3) A. 1 hour post sham irradiation; **B.** 4 hours post sham irradiation; **C.** 24 hours post sham irradiation. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot.



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	333	100	39.77	21.03	0	96

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	477	100	47.67	12.61	0	95

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	588	100	47.24	13.47	0	96

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	322	100	44.02	13.15	0	75

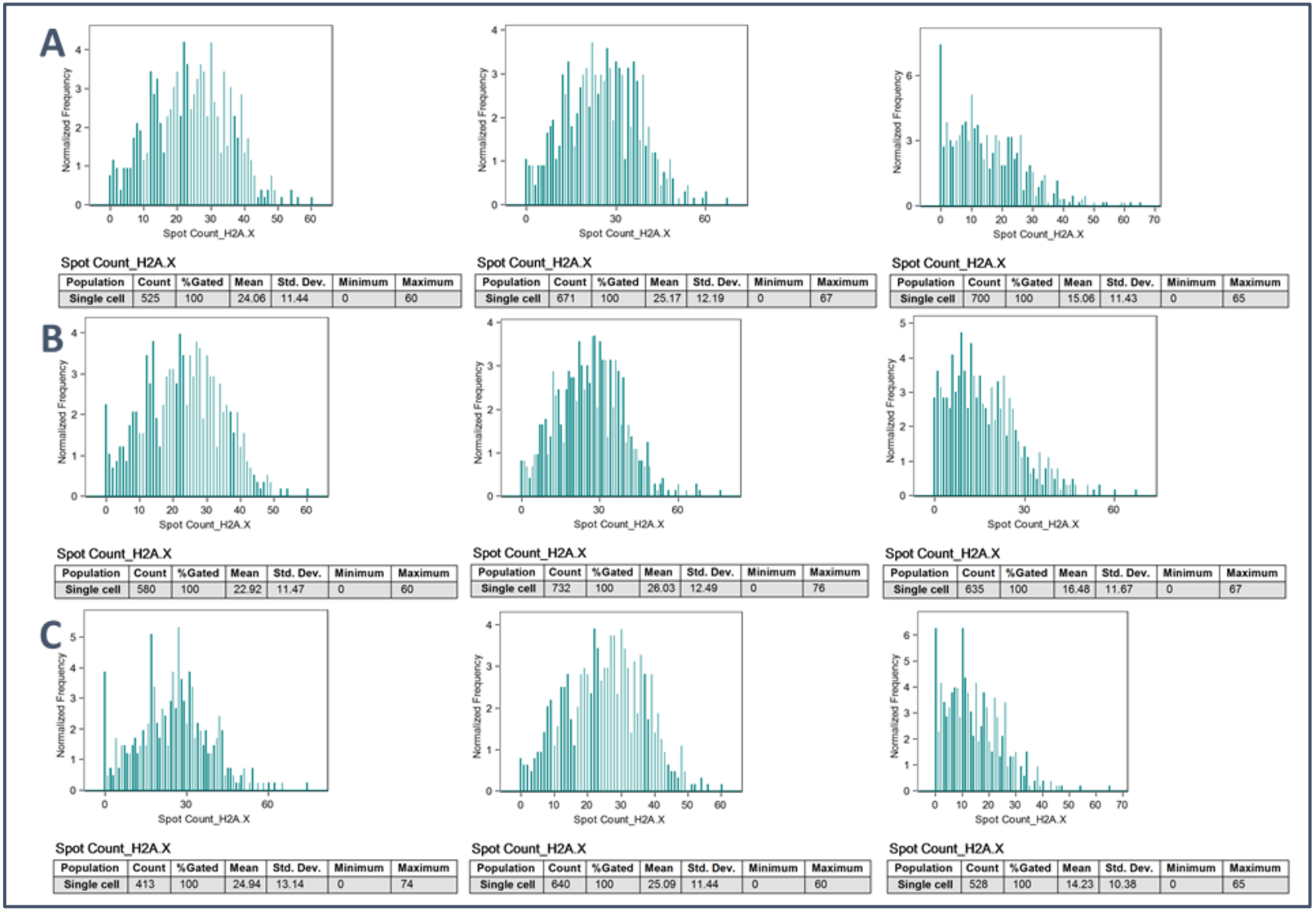
Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	448	100	50.39	19	0	117

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	597	100	42.38	11.5	0	69

Figure J.11: Cells treated with 250µM doxorubicin hydrochloride for 24 hours to induce DNA double stranded breaks (n=3). DNA damage repair quantified through γH2AX analysis, raw nuclear foci plots **A.** HEK293 cells; **B.** PX330- (CRISPR sham) cells. Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot



Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	525	100	24.06	11.44	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	671	100	25.17	12.19	0	67

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	700	100	15.06	11.43	0	65

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	580	100	22.92	11.47	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	732	100	26.03	12.49	0	76

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	635	100	16.48	11.67	0	67

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	413	100	24.94	13.14	0	74

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	640	100	25.09	11.44	0	60

Spot Count_H2A.X

Population	Count	%Gated	Mean	Std. Dev.	Minimum	Maximum
Single cell	528	100	14.23	10.38	0	65

Figure J.12: Cells treated with 250µM doxorubicin hydrochloride for 24 hours to induce DNA double stranded breaks (n=3). DNA damage repair quantified through γH2AX analysis, raw nuclear foci plots **A.** e2-B1.15 (UIMC1 homozygous knockout); **B.** e2-B3.1 (UIMC1 heterozygous knockout); **C.** e13-KO1 (UIMC1 with 1 AA deletion in ZFN). Single cell population used to count number of spots observed. Count indicates the number of cells counted in this population. Mean, standard deviation, minimum and maximum number of foci observed indicated below each plot