UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Identification of rare gene variants in South African breast cancer families through next generation sequencing

by

Juliet Lewie Dionne Mentoor

Submitted in fulfilment of the requirements for the degree of

Philosophiae Doctor in Human Genetics

at the Faculty of Health Sciences,

Department of Genetics,

University of Pretoria, Pretoria

July 2017

Supervisor: Prof Elizabeth J van Rensburg

Co-Supervisor: Prof Fourie Joubert

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration.

Student number: 04378067

I, Juliet Mentoor, declare that this dissertation/thesis entitled, **Identification of rare gene variants in South African breast cancer families through next generation sequencing**, which I hereby submit for the degree Philosphiae Doctor in Human Genetics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Juliet

31 – Jul - 2017

Juliet L.D Mentoor

Date

Summary.

Breast cancer (BC) has become the leading cancer amongst women in South Africa. The overall life time risk for developing this disease is one in 12 (National Cancer registry, 2000-2011). A strong family history (≥3 affected) is an important factor for inherited predisposition to BC that accounts for approximately 10% of cases worldwide. Mutations in several high- and moderate risk breast cancer genes have been associated with familial BC and includes *BRCA1, BRCA2, TP53, PALB2,* and *CHEK2*. Individuals that carry germline mutations in *BRCA1* and *BRCA2* possess an 80% lifetime risk for BC. Mutations in *BRCA1* and *BRCA2* are responsible for 29% and 25% of familial BC worldwide. In South Africa *BRCA1* mutations account for 19% and *BRCA2* for 47% of familial breast cancer. Mutations associated with a moderate risk for BC account for ~1% of cases. This data suggests that ~30% of South African BC families are not characterised by pathogenic mutations in known breast/ovarian (BC/OVC) genes.

The purpose of the present study was to identify gene variants that may predispose to breast cancer. Next generation sequencing was performed to investigate the germline DNA of high-risk BC/OVC families that have previously tested negative for premature truncating mutations in *BRCA1*/2, *PALB2* and *RAD51C*. Paired-end whole exome sequencing was performed with nine index cases, selected from six families with a strong background for BC/OVC. This resulted in the discovery of an average of 26 000 coding variants in index cases. Gene prioritisation strategies were incorporated to filter all exome variants and identify high-priority genes for further analysis. After sequence verification, three high-priority genes were selected for further analysis.

The three genes coded for; a novel putative tumour suppressor (TCHP) that is pro-apoptotic; the XPF-endonuclease homolog, EME2; and a POLQ like helicase enzyme (HELQ). Prioritised genes were screened in a total of 61 high-risk families and cohorts of patients with BC or OVC without a family history for their disease. Two potentially damaging variants (stop-gain & in-frame amino acid deletion) were identified in *TCHP*, four (frameshift, nonsense & two in-frame deletions) in *EME2* and one frameshift mutation in *HELQ* in high-risk families and cases that were without a family history for BC/OVC.

The analyses performed in the last section of this project was aimed at identifying other potential genes of interest by making use of a list of 516 well recognised and putative DNA repair genes. Through this approach, one additional truncating mutation in *POLN* (p.Q837SfsX7) was highlighted as a potential gene of interest for future investigation.

Despite the key roles that the high-priority genes play in their respective processes, the present study could not verify that the potential loss of function variants discovered make an appreciable contribution towards BC/OVC susceptibility in our setting. Further investigation is necessary to validate their involvement in breast/ovarian cancer predisposition.

Acknowledgements.

Herewith I would like to extend my sincere thanks and gratitude to the following individuals for their assistance during the course of this project.

1. To my supervisor, Prof Elizabeth J van Rensburg and co-supervisor Prof Fourie Joubert for their leadership, insight and guidance towards the completion of this study.
2. I thank the following funding bodies for their financial assistance: The Cancer Association of South Africa (CANSA), South African Medical Research Council (MRC) & the University of Pretoria, Institutional research themes (IRT).
3. Celmari Dorfling for her assistance.
4. All my fellow colleagues at the Cancer Genetics laboratory.
5. My dear friends (particularly Dr. Kerry Reid) for their support.
6. I would like to thank my family for their support and love. I also thank my late sister, grandmother and uncle for their spiritual guidance.
7. Lastly, I praise my heavenly father for his strength and blessings.

Table of Contents.

Page

Page

Abbreviations.

%	Percentage
µg	Microgram
µl	Microliter
A	
ADP	Adenosine diphosphate
A-GVGD	Align Grantham variation Grantham deviation
ATM	ATM serine/threonine kinase
ATP	Adenosine triphosphate
B	
BC	Breast cancer
BGI	Beijing Genomics Institute
bp	Base pair
BRC	Familial breast cancer patient
BRCA1	Breast cancer type 1 susceptibility protein
BRCA2	Breast cancer type 2 susceptibility protein
C	
CHEK2	Checkpoint kinase 2
Chr	Chromosome
CK	Cytokine
COSMIC	Catalogue of Somatic Mutations in Cancer
D	
Db	Databases
del	Deletion
DNA	Deoxyribonucleic acid
dsDNA	Double stranded DNA
DSB	Double strand breaks
E	
EA	European-American
EDTA	Ethylenediaminetetraacetic acid
ER	Estrogen receptor (+): positive, (-): negative
ERBB2	Receptor tyrosine-protein kinase erbB-2
ESP6500	Exome sequencing project6500
EtOH	Ethanol

EUR	European
Ex	Exon
ExAC	Exome Aggregation Consortium
F	
F	Forward
FRR	Familial relative risk
FS	Frameshift
FTP	File transfer protocol
G	
G2SBC	Genes-to-Systems Breast Cancer
GATK	Genome analysis toolkit
GATKv2.4.9	GATK version 2.4.9
GI	GenInfo Identifier
GWAS	Genome-wide association study
H	
hBC	Hereditary breast cancer
HER2	Human epidermal growth factor receptor 2
HG37	Human genome issue 37
HGVS	Human genome variation society
HSF	Human splice finder
I	
ICGC	International Cancer Genome Consortium
IF’s	Intermediate filaments
IF-indel	Inframe
Indel	Insertion-deletion
ins	Insertion
K	
K16	Keratin 16
K8/18	Keratin 8/18
L	
LJB	Predictions of missense variants by X. **L**iu, X. **J**ian, and E. **B**oerwinkle (Liu, et al. 2011)

M	
M	Molar
mAb	Monoclonal antibody
MAF	Minor allele frequency
MgCl$_2$	Magnesium chloride
mM	Milimolar
N	
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
NMD	Nonsense mediated decay
nsSNV	Non synonymous single nucleotide variant
^{o}C	Degrees Celsius
O	
OR	Odds ratio
OVC	Ovarian cancer
OVW	White ovarian cancer patient
P	
PALB2	Partner and localizer of BRCA2
PARP1	Poly(ADP-ribose) polymerase 1
PCR	Polymerase chain reaction
PE	Paired-end
PI3K	Phosphoinositide 3-kinase
PPV	Putative pathogenic variant
PR	Progesterone receptor
PTT	Protein truncation test
PTV	Premature truncation variant
Q	
QC	Quality control
R	
R	Reverse
RefSeq	Reference sequence
RNA	Ribonucleic acid
rpm	Revolutions per minute
RR	Relative risk
RSA	Republic of South Africa

S

SC	Splice-site
SG	Stop-gain
SIFT	Sorting tolerant from intolerant
SNP	Single nucleotide polymorphism
SSCP	Single strand conformation polymorphism
ssDNA	Single stranded DNA

T

Ta	Annealing temperature
Tann	PCR annealing temperature
TBE	Tris/borate/EDTA

U

UK	United Kingdom

V

V	Volts
VCF	Variant call format

W

WES	Whole exome sequencing

X

χ2	Chi-square

Y

y	Years

List of tables.

Chapter 1: Literature review

Chapter 3: Results of whole exome sequencing quality analysis and variant identification

Chapter 4: *TCHP* sequence variants in South African breast and ovarian cancer families

Chapter 5: *EME2* gene variants in high-risk South African breast and ovarian cancer families

Chapter 6: *HELQ* gene variants in high-risk non-*BRCA1/2* South African families

Chapter 7: Germline sequence variants in DNA repair genes of South African breast/ovarian cancer families with and without *BRCA* mutations

List of figures.

Chapter 1: Literature review

Chapter 2: Materials and Methods

Chapter 3: Results of whole exome sequencing quality analysis and variant identification

Chapter 4: *TCHP* sequence variants in South African breast and ovarian cancer families

Chapter 5: *EME2* gene variants in high-risk South African breast and ovarian cancer families

Chapter 6: *HELQ* gene variants in high-risk non-*BRCA1/2* South African families

Chapter 1:

Literature Review

1.1. Introduction

Breast cancer (BC) has become the most commonly diagnosed cancer in women, worldwide, with a reported incidence rate of 25.2% (Figure 1.1). GLOBOCAN estimates for 2012 show that BC is the leading cause of cancer-related deaths as it has a mortality rate of 14.7% (Ferlay, et al. 2015). In South Africa, BC is ranked as the most diagnosed cancer in women accounting for ~21% of all cancers. When arranged according to population group the incidence rates range from ~20% in white, ~21% in black to ~26% in coloured women (National Cancer registry, 2011). Overall, South African women have a life-time risk of one in 29 for developing BC before the age of 74 years. The risk levels vary in each population group, ranging from one in 12 in white, one in 18 in coloured and one in 50 in black women (National Cancer registry, 2011). The National Cancer Registry (NCR) uses a pathology-based method for data collection and is merely the tip of the iceberg. The data may still support previous suggestions that breast cancer diagnosis among South Africans has steadily increased and has become more common in non-white African individuals (van der Merwe, et al. 2012; Vorobiof, et al. 2001).

The distinct difference in BC global incidence and mortality rates seen among various population groups has become less significant (Jemal, et al. 2010). While BC incidence rates are high in European and American countries a rapid increase in cancer cases (including breast, lung, colorectal and prostate) has also been reported in less developed and socio-economically developing countries (Jemal, et al. 2010). The decline in BC related deaths within western countries in general could possibly be due to improved treatment strategies in healthcare systems (Autier, et al. 2011). The increased incidence rates in Africa could be ascribed to changes in lifestyle factors and improved detection practices (Jemal, et al. 2012), the latter of which could lead to increased awareness among immediate family members (J van Rensburg, et al. 2007; van der Merwe, et al. 2012).

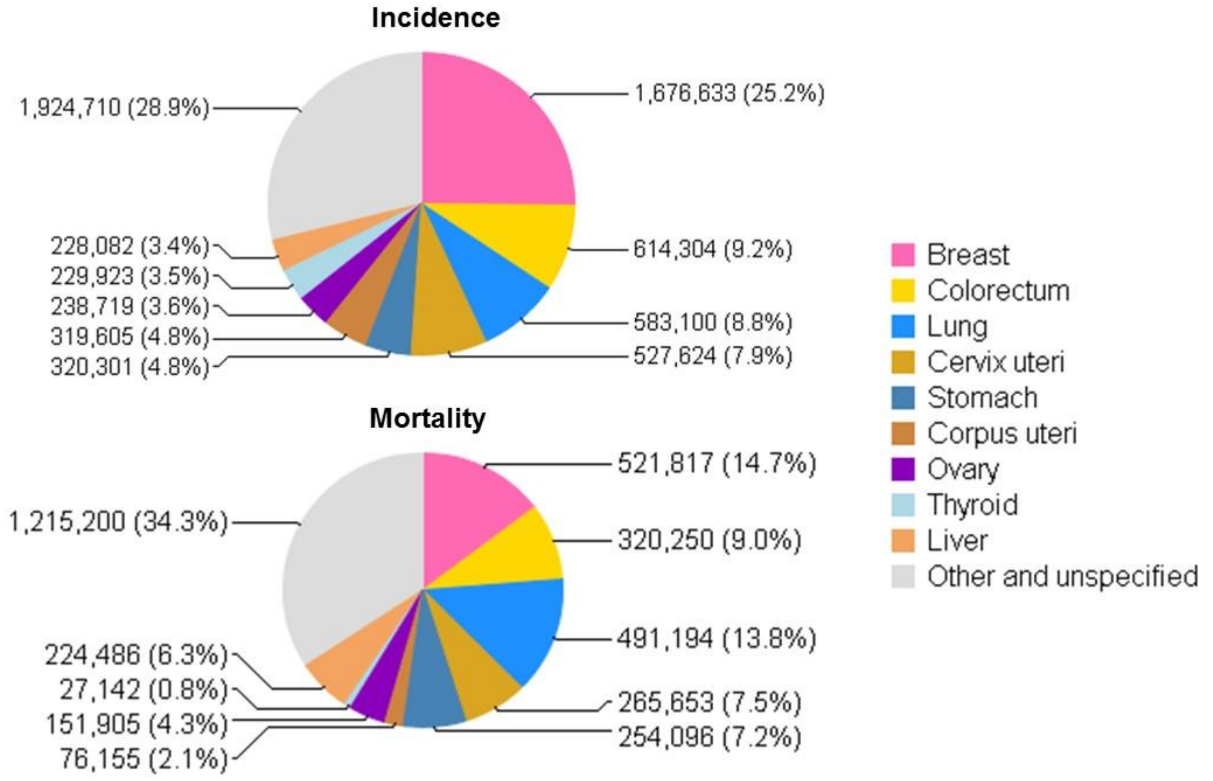


Figure 1.1: GLOBOCAN cancer incidence and mortality statistics among women during 2012.

The incidence and mortality rates (%) of BC is higher than any other cancer-related illness, worldwide. Breast and ovarian cancers (BC/OVC) are diagnosed at higher frequencies than all other non-communicable diseases (Ferlay, et al. 2015).

1.2. Histopathological and molecular classification of breast cancer

BC is a complex disease that is defined by various clinical and histological morphologies, molecular profiles and therapeutic responses (The Cancer Genome Atlas Network 2012; Usary, et al. 2013). The histological profiles of breast cancers differ greatly (Prat and Perou 2011). Various biological subtypes of BC exist and differ with regards to their molecular alterations, cellular composition, response to therapy and survival rates. These different histological BC profiles originate from the metastatic progression of distinct cell types (Prat and Perou 2011; Sotiriou and Pusztai 2009).

The classification of breast tumours have long been guided by the presence/absence of three pathological markers i.e. the oestrogen receptor (ER), progesterone receptor (PR) or the epidermal growth factor receptor 2 (ERBB2 aka HER2). Clinical parameters such as age, the involvement of axillary lymph nodes (Figure 1.2), tumour size and the level of cell differentiation (i.e. histologic grade) are incorporated as well (Olsson, et al. 2013). Histological grading is a method of classifying tumours and assesses the level of cellular abnormality (Tough, et al. 1969). Tumours are generally graded from 1 to 4, grade 1 cancers are composed of well-differentiated cells (low grade) and grade 4 contains un-differentiated cells (high grade). This method of classifying tumours has great prognostic benefits and is still applied (Weigelt, et al.).

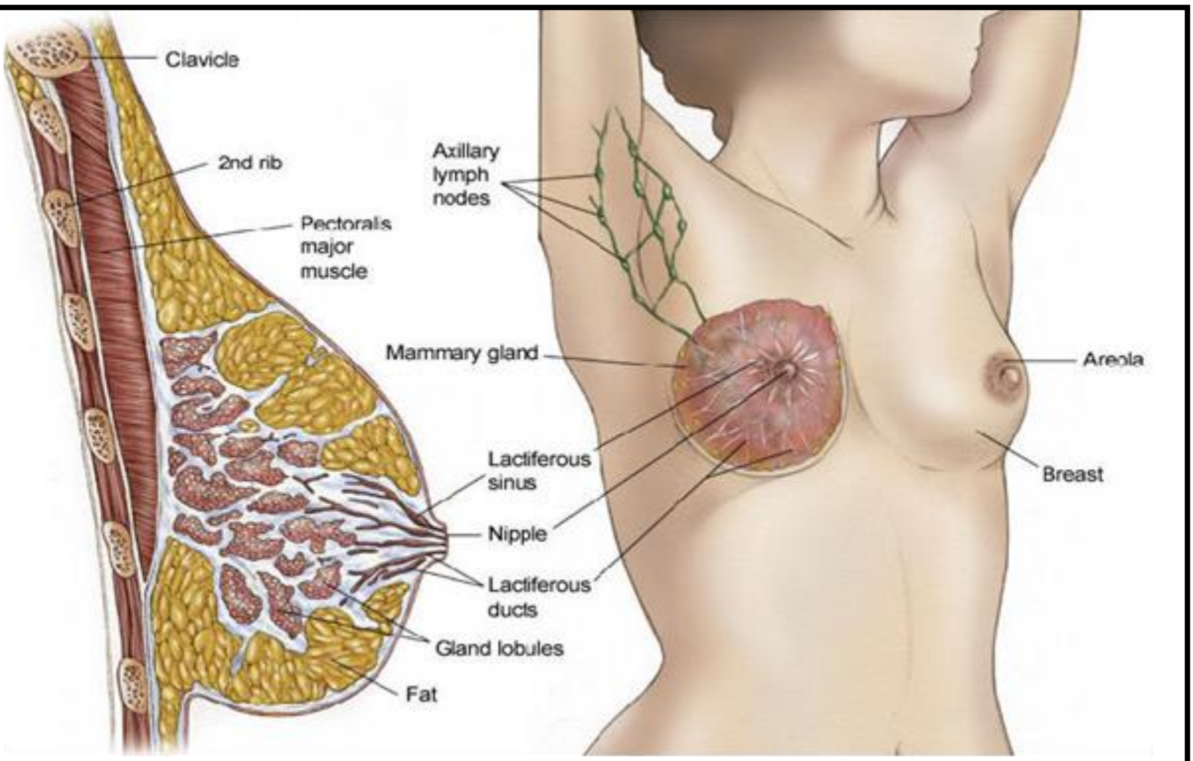


Figure 1.2: Cross sectional diagram displaying the anatomical structure of the female breast.

The image was acquired from the “Human diagram” webpage (http://diagrampic.com/female-anatomy-diagram/female-anatomy-breast/) and illustrates the different locations of primary tumour development.

The use of information provided by clinical and pathological factors are not sufficient to make an accurate prognosis. Molecular techniques such as gene expression profiling of cancer tumours have aided in the discovery of six main intrinsic subtypes used to classify BC tumours. These include luminal A, luminal B, HER2 enriched, Claudin-low, basal-like, normal breast-like and the triple-negative phenotype (Perou, et al. 2000). The use of advanced sequencing technologies have revealed ten novel breast cancer subtypes.

These subtypes are classified based on the single nucleotide polymorphisms, copy number variants and the gene expression profiles in breast cancers (Curtis, et al. 2012).

1.2.1. Breast cancer subgroups

Molecular subtypes are based on the expression pattern of luminal and basal cytokeratin (CK) proteins (a.k.a. CK7/8/18/19 and CK5/6/14, respectively) (Prat and Perou 2011). The subtypes also differ with regards to the status of three hormone receptor markers (ER, PR, HER2), incidence rates, survival and response to treatment (Prat and Perou 2011; Usary, et al. 2013).

The luminal tumour subgroup is characterised by overexpression of genes that are produced in normal luminal epithelial cells (duct, Figure 1.2), for example cytokeratin proteins CK8/18 and ERBB2, and are often ER/PR-positive (Eroles, et al. 2012; Prat and Perou 2011). Luminal (also known as ER$^+$) cancers are further divided into subgroups A or B. Subtype A expresses the highest levels of ER-α and proteins regulated by oestrogen. Overall, BC patients with the luminal A subtype have a longer disease-free survival time compared to the basal-like and HER2$^+$ patients. Subtype B express moderate levels of hormone receptors and are primarily comprised of abnormal/cancerous undifferentiated cells (i.e. high grade) (Prat, et al. 2013; Sotiriou and Pusztai 2009). The gene expression profiles of both luminal A & B breast tumours are diverse and they contain the highest number of significantly mutated genes. Luminal/ER$^+$ is a cluster of cancers that are highly heterogeneous with regards to patient outcome (The Cancer Genome Atlas Network 2012).

The epidermal growth factor receptor 2 (HER2) is a proto-oncogene and its activation plays an important role in forming some of the most aggressive forms of BC (Montemurro, et al. 2013). HER2/neu amplified tumours express low levels of ER and all proteins that are co-expressed with ER (Eroles, et al. 2012). Tumours that are referred to as “normal breast-like” possess gene expression profiles that are distinctive of basal epithelial and adipose (fat, Figure 1.2) cells as well as low levels of luminal cell gene expression (Bosch, et al. 2010).

Basal-like breast cancers show increased levels of CK5/6 & CK17, epidermal growth factor receptor (EGFR), laminin and fatty acid binding protein 7 (FABP7) expression. The status of five biomarkers is used as a more specific means of defining basal-like BC i.e. ER$^-$, PR$^-$, HER2$^-$, EGFR$^+$ & CK5/6$^+$ (Badve, et al. 2011). Basal-like tumours are universally referred to as triple negative tumours (i.e. TNBC) as the immune-phenotype of these subtypes share certain features such as the lack of ER’s (Badve, et al. 2011). However, the two molecular groups (TNBC and basal-like BC’s) are not equivalent. The term “basal-like” describes tumours that not only have increased gene expression levels in breast epithelial cells that are classified through microarray expression profiling (Sandhu, et al. 2010). On average these tumours are larger in size and consists mainly of poorly differentiated cells (i.e. high grade) (Badve, et al. 2011). Claudin-low tumour types are similar to basal-like cancers as they share the absence of HER2 and luminal gene clusters (Eroles, et al. 2012). Intrinsic claudin-low subtypes possess high migratory and invasive (mesenchymal) properties and low luminal or epithelial cell differentiation. These properties lead to the transition of epithelial-to-mesenchymal (EMT) cells which contributes to metastases (Taube, et al. 2010). Claudin-low cancers are a molecular subtype of aggressive triple negative breast cancers and have a poor prognosis (Prat, et al. 2010; Prat and Perou 2011; Taube, et al. 2010).

TNBC's are described by the absence of hormonal receptors ER, PR and HER2 which was determined by immunohistochemistry assays (Bosch, et al. 2010). TNBC's also have a high rate of cell mitosis and tumour necrosis. This cancer subgroup is highly invasive and form ductal carcinomas as well as other histological types such as medullary and metaplastic carcinomas, which is generally seen in patients (Prat and Perou 2011). Mammographic diagnosis and prognosis is quite challenging as the size of the primary tumour in TNBC cases is usually indirectly proportional to the chances of survival and tumours grow rapidly (Foulkes, et al. 2010). TNBC's are very difficult to treat as they don't express targets that can be inhibited to improve a patient's chance of survival and therefore have a very poor prognosis (Bosch, et al. 2010).

Histopathological and molecular classification of intrinsic tumour groups is a powerful tool used to determine the clinical outcome of patients. In order to accurately determine a patient's prognosis both genomic and clinical data must be included (Eroles, et al. 2012). Acquiring a correct prognosis will assist in making distinct predictions regarding the sensitivity of the different molecular classes to various treatment options. These recommendations will result in treatment strategies that are specifically designed for patients, minimising the risk of relapse and achieving maximum benefit (Olsson, et al. 2013).

1.3. Hallmarks of breast cancer development

It has been suggested that the occurrence of ten biological events form part of an organisational framework that explains the complexities of tumour development (Floor, et al. 2012). This includes; self-sufficient cell proliferation, insensitivity to anti-growth signals, evading apoptosis, angiogenesis, replicative immortality, metastasis, deregulated cellular metabolism, genome instability and immune evasion (Figure 1.3) (Emmert-Streib, et al. 2014; Hanahan and Weinberg 2011). Hallmarks of cancer development are defined as capabilities that allow cancer cells to survive, proliferate and disseminate (Floor, et al. 2012). These cascade events are brought on by tumour-promoting inflammation and increased genomic instability in otherwise healthy/normal cells (Hanahan and Weinberg 2011). The most important factor is the development of genomic instability and together, these events enable tumorigenesis (Floor, et al. 2012; Negrini, et al. 2010). The loss of genome integrity increases cellular diversity and promotes tumour progression by indirectly supporting the occurrence of many other hallmark events (Floor, et al. 2012; Hanahan and Weinberg 2011).

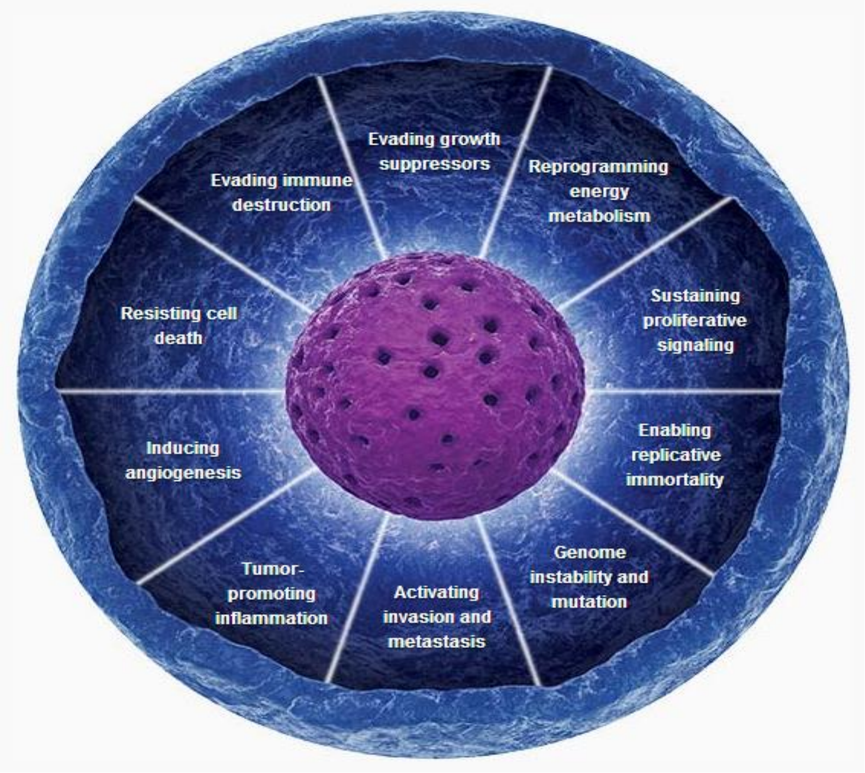

Evading growth suppressors
Reprogramming energy metabolism
Sustaining proliferative signaling
Enabling replicative immortality
Genome instability and mutation
Activating invasion and metastasis
Tumor-promoting inflammation
Inducing angiogenesis
Resisting cell death
Evading immune destruction

Figure 1.3: Hallmark biological events that explain the organisational processes that drive tumour progression.

There are a total of ten hallmark capabilities that must be acquired for cancer development/tumorigenesis. The loss of genomic stability initiates a cascade of downstream events that promote tumour progression (with permission from Hanahan and Weinberg 2011).

BC is characterised by various biological characteristics including molecular markers, gene expression profiles and genomic alterations (Kwei, et al. 2010). Improved understanding of the mechanisms that contribute to oncogenesis will assist in the development of improved treatment measures (Eccles, et al. 2013). Systematic re-sequencing of cancer genomes has assisted in identifying distinct patterns of genomic rearrangements that are used to characterise BC tumours.

1.3.1. Genes involved in the maintenance of genomic integrity

There are two classes of proteins that act together to regulate cellular proliferation, i.e. “caretakers” and “gatekeepers” (Hanahan and Weinberg 2011). Gatekeeper proteins function as part of a system of checks that monitors the level of cell division and cell death. These proteins are in place to lessen the effects of genome/tissue damage (Thiagalingam 2015). Caretaker proteins are responsible for maintaining genome integrity (Yao and Dai 2014). Genomic instability leads to the accumulation of DNA mutations that affect either genome gatekeeper or caretaker genes and make cells more sensitive to any additional mutations (Kwei, et al. 2010). Gatekeeper genes generally code for proteins that function in cell cycle arrest and cell death (Vogelstein, et al. 2013). Typical caretaker genes include genes that function in DNA repair pathways and in the different phases of cell mitosis (Negrini, et al. 2010). Genomic instability brought on by mutated caretaker genes renders the cell selectively vulnerable to both exogenous and endogenous mutagens (Smith, et al. 2010).

Cancer malignant neoplasia is aided by the mutation of two gene classes, oncogenes and tumour suppressor (TS) genes (Stephens, et al. 2012). Proto-oncogenes (normal/non-mutated oncogenes) code for proteins that drive cell growth and division and function in the gap (G) and segregation (S) phases of the cell cycle (Chow 2010). Growth factor receptors and proteins that initiate DNA replication are examples of such genes. Proto-oncogenes acquire mutations and become oncogenes that result in the activation of irregular cell proliferation and contribute to tumour growth. Oncogenes induce genome instability by activating growth signalling pathways which places the cell under replicative stress (Negrini, et al. 2010; Osborne, et al. 2004).

TS proteins restrict cell growth and promotes programmed cell death (apoptosis) (Delbridge, et al. 2012). Mutations in TS genes have a loss of function effect on cellular mechanisms that inhibit persistent cell division, ultimately resulting in malignant cell growth. The TS gene *TP53* and genes that code for DNA damage recognition proteins are examples which have proved to be the most mutated genes in cancer tumours (Chow 2010; Negrini, et al. 2010). Tumour suppressors and oncogenes provide a link between cell cycle regulation and control as well as the formation of tumours and ultimately to cancer development (Chow 2010). TS genes are significantly more mutated than oncogenes (Negrini, et al. 2010) and as a result, TS genes have proved to be most useful when diagnosing BC (Osborne, et al. 2004).

A number of variants are known to result in the inactivation of tumour suppressors and are acquired either somatically or are inherited. These variants have the potential to affect pathways that are important for tumorigenesis (Delbridge, et al. 2012). In 1971 Alfred Knudson proposed the “two-hit" hypothesis used to explain the genetic predisposition for cancer. This hypothesis states that mutations that are acquired affect one gene allele (e.g. germline mutations). The deletion of a second allele (e.g. somatic mutation) within the same gene location has the potential to cause the malignant phenotype of BC (Weitzel, et al. 2011).

1.3.2. Mutations that contribute to tumour cell growth

Cancer development occurs through the contribution of both germ line and somatic variants. Germ line mutations associated with cancer predisposition are generally located in genome regions that code for proteins implicated in the repair of DNA alterations and cell cycle checkpoint control (Helleday, et al. 2007). Somatic mutations are implicated in tumour development (Nehrt, et al. 2012). These variants stem are acquired from endo- or exogenous mutagens (Negrini, et al. 2010; Witsch, et al. 2010). The clonal proliferation of cancer cells leads to increased genomic instability (Negrini, et al. 2010). The majority of gene variants are “passengers” as they are biologically inert and not causally implicated in the development of cancer tumours. Gene mutations which drive cell growth are known as “driver” mutations (Vogelstein, et al. 2013). Cancer genes carry the most significant number of driver variants. Driver mutations consist of different combinations of genetic changes that are causally implicated in the development of cancer (Stephens, et al. 2012).

Gene deletions and amplifications have proved to be rarer than single nucleotide alterations (Vogelstein, et al. 2013). Genomic rearrangements that underlie BC tumorigenesis result from distinct mechanisms including tandem duplications, deletions, inversions and rearrangements in amplified genome regions (Stephens, et al. 2009). Various copy number variants have also been identified in the genomic profile of BC tumours (either gain or loss of gene dosage) (Kwei, et al. 2010) with 50% of BC mutations being located in protein coding regions (Stephens, et al. 2009). However, mutations in non-coding regions of the genome (i.e. promoters, enhancers, or negative regulatory elements) can also affect the regular expression of proteins that are required for normal cell cycling. Collectively, these events transform healthy cells to cancerous cells. (Ziebarth, et al. 2012)

1.4. Breast cancer risk factors.

Multiple factors contribute to the development of BC. The chances of developing BC increases with age. Reproductive factors such as early start of menarche (<12yrs), late onset of menopause (>55yrs) and pregnancy is said to modestly increase the risk for BC (Yang, et al. 2011). Factors such as geographic location and socioeconomic status are known to influence the likelihood of survival (Jardines, et al. 2011). Differences in the incidence of BC in women of higher socioeconomic background may be due to variable contributing factors, such as dietary fat intake, which has proved to be a significant external risk factor (Patterson, et al. 2010).

Other factors that may contribute to the risk for developing BC include, radiation exposure, viral infection (e.g. Epstein-Barr (EBV)) and exogenous hormone use (Ellsworth, et al. 2015; Mena, et al. 2009). The involvement of some are more definitive than others (Lawson 2009). Ultimately, the risk for developing BC is assessed by taking into account these multiple contributing factors including family history, personal clinical information and an individuals' risk, that is based on results ascertained from genetic testing (Mai, et al. 2011; Mealiffe, et al. 2010).

1.4.1. Environmental influence on breast cancer risk

BC is a multifactorial disease and its development occurs as a result of one or more environmental factors interacting with multiple genes (Eroles, et al. 2012; Prat and Perou 2011). The complexity of BC cannot be explained by each element separately but by the product of the relative risk of these factors as described by a multiplicative model (Thomas 2010). The degree of environmental influence is determined by different features such as the length of or age at exposure. The effect of environmental risk factors is directly proportional to the duration of exposure e.g. exposure to pollutant/chemical agents (Landau-Ossondo, et al. 2009).

Pesticides and/or industrial chemicals are examples of endocrine disrupting chemicals which are highly carcinogenic. A growing amount of evidence is now available which proves that these environmental chemicals have an effect on cancer development (Ferro 2012). Household pesticides do not pose a significant threat to cancer risk (Farooq, et al. 2010), whereas, prolonged use of agricultural biocides have a detrimental effect on ERs and cause DNA damage (Sonnenschein and Soto 2010). Endocrine disruption has severe physiological effects on tissue development. These exogenous agents interfere with endogenous hormones (Mnif, et al. 2011; Sonnenschein and Soto 2010). Studies have shown that endocrine disruption plays a significant role in BC development (Ferro 2012; Landau-Ossondo, et al. 2009; Shakeel, et al. 2010).

Carcinogens such as heavy metals, alcohol, tobacco smoke, reactive oxygen and androgen hormones are additional examples of exogenous and endogenous factors that each contribute to the risk for BC development (Johnson, et al. 2011). Heavy metals and reactive oxygen are examples of DNA damaging agents (Mena, et al. 2009). The hormone androgen has an important physiological effect on women as it is the precursor molecules to oestrogen (Lin, et al. 2009). The role that androgen plays in BC development remains unclear. Dihydrotestosterone (DHT), is a non-aromatizable androgen metabolite that stimulates irregular cell proliferation. *In vitro* studies have shown that DHT exposure results in cancer cells that are enriched with oestrogen and androgen receptors (Lin, et al. 2009; Need, et al. 2012b). Exploring the involvement of environmental risk factors could help in the discovery of novel preventative measures that may decrease the incidence of BC (Potter 2011).

1.4.2. Hereditary risk factors

Family history is known to be the strongest risk factor for BC. Approximately 5% to 10% of BC cases are caused by hereditary factors that is characterised by a high incidence of the disease among family members (Apostolou and Fostira 2013). Individuals that are part of BC families have a 1.8 – 3.9 times greater risk for developing BC than the general population risk of ~12% (Collins and Politopoulos 2011; DeSantis, et al. 2014). Familial relative risk (FRR) is the risk for disease that the relative of an affected individual possesses in comparison to that of the general population (Mavaddat, et al. 2010a).

FRR decreases as the age of disease onset in the affected relative increases. The risk for disease is also greater for index cases if the number of affected family members is high. In these instances the range of risk extends further to include more distant relatives as well. Lastly, the pathological subtype of tumours significantly affect the FRR for BC. The relatives of patients with ER^- and ER^+ BC are at higher risk for disease than the general population (Mavaddat, et al. 2010b).

Cancer research has greatly advanced in the last 40-plus years. Research is mainly aimed at identifying genes that, when mutated, significantly increases the risk for BC disease (Fletcher and Houlston 2010). The identification of cancer predisposition genes has led to a better understanding of the mechanisms that are involved breast cancer development (Rahman 2014).

1.5. Breast cancer susceptibility

Currently there are three classes of gene variants that are associated with high to low risk for BC (Figure 1.4). These variants are either located in high or moderate risk genes or are low risk alleles (Korde and Shiovitz 2013). The majority of high-penetrant BC predisposing mutations are found in genes that code for DNA caretakers and are the master regulators of the genome (e.g. *BRCA1, BRCA2, TP53, PTEN,* etc.) (Shuen and Foulkes 2011). Pathogenic variants in these genes are associated with an increased risk for BC that is 30–80% higher than the lifetime risk for the general population (Collins and Politopoulos 2011). Moderate penetrant mutations, confer a modest amount of risk for BC (15-40%, Figure 1.4). These mutant alleles are more frequent than high-risk alleles (frequency of 0.05 – 0.5) (Collins and Politopoulos 2011; Mavaddat, et al. 2010a). The most common mutant alleles that potentially also contribute to cancer development have a per-allele risk ratio that ranges from approximately one to 1.4-fold (Bogdanova, et al. 2013). Low penetrance BC susceptibility alleles (also known as low-risk, Figure 1.4) act collectively and confers a small increase in disease risk (Collins and Politopoulos 2011; Smith, et al. 2008).

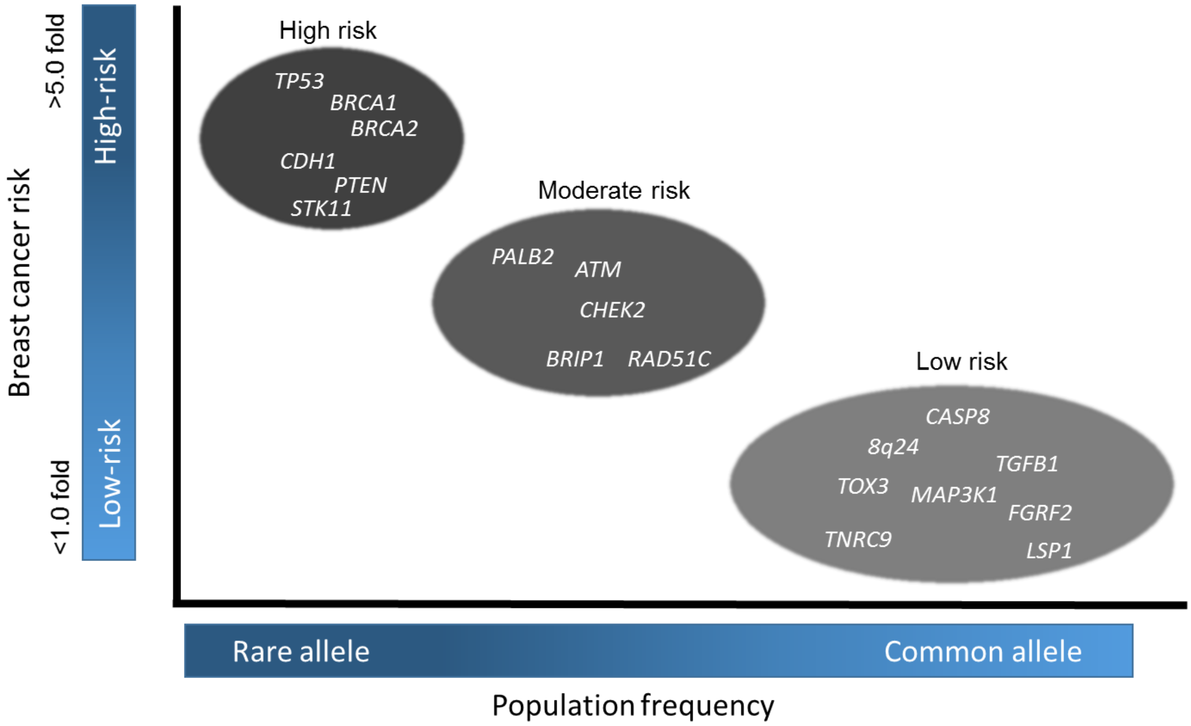


Figure 1.4: Known breast cancer susceptibility genes and associated risk.

The relative risk and allele frequencies of rare variants in high-, moderate- risk genes and low risk alleles (adapted from Weitzel, et al. 2011).

1.5.1. High penetrance mutations in breast cancer genes

Insights into the mechanisms that underlie cancer susceptibility has led to the identification and localisation of many cancer predisposition genes. There are ~30 known genes that significantly increase the risk for BC when mutated (Collins and Politopoulos 2011). This number is steadily rising as more studies are identifying novel candidate cancer genes that are involved in genome regulation (Apostolou and Fostira 2013; Latif, et al. 2010). High penetrance mutations in BC predisposition genes are associated with the most intermediate risk for disease (Apostolou and Fostira 2013; Fletcher and Houlston 2010). Table 1.1 lists the six main high-risk BC genes, three of which include the tumour suppressor gene (*TP53*) and BC early onset 1 & 2 (*BRCA1* and *BRCA2*). The discovery and localisation of *TP53*, *BRCA1* and *BRCA2* took place during 1979, 1990 and 1994 respectively with the use of linkage and positional cloning analysis (Hammer, et al. 2011; Miki, et al. 1994; Wooster, et al. 1995).

Table 1.1: High-risk breast cancer genes and associated syndromes

Gene	Protein function	Lifetime BC risk	Associated syndrome
BRCA1 **&** ***BRCA2***	Growth suppression and DNA damage response (late G1 cell phase) – BRCA1 RAD51 mediated homologous recombination repair of double strand (ds)DNA breaks – BRCA2	40 – 80% BC 11-40% OVC by age 70[a,c]	Hereditary breast/ovarian cancer
TP53	Mediates anti-proliferative processes in response to general cell stresses, including DNA damage	49% BC by age 60[b,c]	Li-Fraumeni syndrome (LFS)
PTEN	Inhibits the phosphatidylinositol 3' kinase mediated cell growth pathway	25-50% by age 50[c]	PTEN hamartoma tumour syndromes (PHTS) i.e. Cowden syndrome (CS)
CDH1	Maintain normal architecture and function of epithelial tissues	42% lobular BC by age 80[c,d]	Hereditary diffuse gastric cancer (HDGC)
STK11	Regulates cell polarity, proliferation and arrest	32% by age 60[c,e]	Peutz-Jeghers syndrome (PJS)

[a] (Petrucelli, et al. 2013), [b] (Masciari, et al. 2012), [c] (Korde and Shiovitz 2013), [d] (Hansford, et al. 2015), [e] (Resta, et al. 2013)

Hereditary cancer syndromes are also associated with a high lifetime risk for BC relative to the general population. The inherited variants are located in genes that predispose for multiple primary tumours including BC, brain, bone and skin (Apostolou and Fostira 2013). Cowden syndrome is an example of a hereditary disease with a strong BC component (Table 1.1). The disease develops as a result of loss-of-function mutations in the tumour suppressor, phosphatase and tensin homologue (PTEN). Recessive germline mutations in *PTEN* causes Cowden syndrome and these patients are at risk for BC (Farooq, et al. 2012; Ghoussaini, et al. 2013). *TP53* variants are linked to Li–Fraumeni syndrome-associated (LFS) tumours. Individuals with biallelic mutations develop LFS and patients diagnosed with this syndrome are predisposed to various cancers, including breast cancer (Ruijs, et al. 2010).

Fanconi's Anemia (FA) is a recessive disorder that is associated with BC susceptibility caused by mutations in the FA/DNA-repair pathway (Moldovan and D'Andrea 2009). This pathway is responsible for recognising and repairing damaged DNA regions. Patients with homozygous inactivating mutations in the components of the FA pathway develop various clinical abnormalities and is susceptible to early-onset BC (D'Andrea 2010; Moldovan and D'Andrea 2009). Peutz–Jeghers is a condition that is characterised by various malignancies and is caused by biallelic mutations in the tumour suppressor serine/threonine kinase 11 (STK11, also known as liver kinase B1 (LKB1)) (Resta, et al. 2013; Yajima, et al. 2013). The loss of STK11's kinase activity potentially disrupts normal cell proliferation and p53-dependant apoptosis which are pathways that when down-regulated, are associated with a significant risk for cancer (Beggs, et al. 2010; Floor, et al. 2012; Resta, et al. 2013; Yajima, et al. 2013).

1.5.1.1. *BRCA1* and *BRCA2* genes

BRCA proteins are involved in the same DNA repair pathway i.e. homologous recombination (Grabarz, et al. 2012). This mechanism repairs DNA by accurately resynthesizing damaged sequences from the intact homologous sister chromatid (Helleday, et al. 2007). Both BRCA proteins play a vital role in recognising and initiating DNA repair by assembling all the elements that are involved in this multidimensional pathway. BRCA1 is more directly involved in homologous repair, checkpoint control, spindle regulation and transcriptional regulation. BRCA2 functions as a regulator of RAD51 which plays a critical role in homologous DNA repair (Grabarz, et al. 2012).

The frequency of *BRCA1/2* mutations within the global population range between 1/800 – 1/1000 (Joosse 2012; Mavaddat, et al. 2010a; Shuen and Foulkes 2011). These mutations account for ~50-60% and ~30-40% of BC disease respectively. *BRCA* mutations account for 24% of breast cancer families (Kast, et al. 2016). The pathology and molecular biology of tumours that develop from carriers of *BRCA1* and *BRCA2* mutations differ (Mavaddat, et al. 2010a). Carriers of *BRCA1* & *BRCA2* mutations account for 11-40% of ovarian cancer cases

(Blay, et al. 2013; Petrucelli, et al. 2013; Zhang, et al. 2012). The prevalence of mutations in these high-risk genes vary greatly between different populations and geographical locations. Specific variants have been discovered among the population groups within Western Europe and America (Hall, et al. 2009; Janavičius 2010). In RSA *BRCA* mutations are found in different frequencies within various population groups (Schlebusch, et al. 2010; Sluiter and van Rensburg 2011; van der Merwe, et al. 2012). There are high-risk mutations in these specific genes that are uniquely responsible for hereditary BC susceptibility in certain ethnic groups (i.e. founder mutations) (Reeves, et al. 2004; Sluiter and van Rensburg 2011). Founder mutations account for a high number of *BRCA* variants. *BRCA1/2* founder mutations collectively account for ~60% of cancer in Afrikaner families (19% *BRCA1*, ~40% *BRCA2*) with a significant history of >3 affected individuals (J van Rensburg, et al. 2007; Loubser, et al. 2012).

1.5.1.2. The *TP53* gene

TP53 serves as a transcription factor that activates the expression of genes involved in the cell's response to stress. Genes that are expressed under the control of p53 code for proteins involved in DNA repair, cell cycle control and programmed cell death (i.e. apoptosis) (Hammer, et al. 2011). The primary role of the p53 signalling pathway is to inhibit the growth and survival of abnormal cells and the loss of TP53 function will lead to tumorigenesis (Olivier, et al. 2010). *TP53* mutations account for up to 17% of breast cancers, selected for family history and for only ±3% of sporadic cases. In 50% of the mutation carriers, tumours develop in the breast, soft tissue and bone (Olivier, et al. 2010). Significant loss of function mutations, like large deletions that affect the whole *TP53* gene, lead to the development of cancer phenotypes with a more aggressive pathology (Fernandez-Cuesta, et al. 2012; Ruijs, et al. 2010; Yang, et al. 2013). The significance of only a small number of single nucleotide (also known as missense) variants has successfully been validated (Leroy, et al. 2013).

According to the data generated by the World Health Organisation (WHO) – International agency for research on cancer (IARC), germ line variants in the *TP53* gene account for ~27% of breast tumours among different population cohorts in the world (Petitjean, et al. 2007). A significant number of additional mutated genes also moderately contribute to the global incidence of BC (Fletcher and Houlston 2010).

1.5.2. Moderate penetrant mutations in breast cancer genes

Mutations in a second class of BC predisposition genes are associated with a moderate penetrant disease risk. In contrast to high-risk mutations these are associated with a lower degree of BC predisposition of approximately two to three-fold higher than the general population risk (Foulkes 2008). Moderate penetrant variants do not always act alone in conferring increased risk. Carriers of moderate penetrant gene mutants may have a higher risk for BC if they also possess variants that act like polygenic alleles especially in instances where a strong family history is prominent (Shuen and Foulkes 2011). The proteins coded by this gene class play a more supportive role in DNA damage repair and interact with BRCA1 & BRCA2. The number of genes with moderate risk variants are growing, however, some of the most significant are *CHEK2, BRIP1, ATM, PALB2,* etc. (Table 1.2) (Korde and Shiovitz 2013). The discovery of such genes were all made possible with a combination of population-based or family-based studies (Filippini and Vega 2013). *RAD51C, MRE11, RAD50,* etc. are among a few other moderate–risk genes (Korde and Shiovitz 2013). The significance of their association with BC development remains an interesting topic for investigation (Rahman and Stratton 2008). RAD50 forms part of a complex of three proteins comprised of MRE11, RAD50 and NBS1 (MRN complex). The MRN-complex is critical in the maintenance of genomic integrity and plays a tumour suppressor role (Mavaddat, et al. 2010a).

Research performed in Europe and UK have determined that moderate penetrant alleles collectively account for only 2-3% of familial BC and contribute a <3% relative risk (Stratton and Rahman 2008). The genes are not commonly mutated and have an allele frequency ranging between 0.005-0.01 (Filippini and Vega 2013).

Table 1.2: Moderate-risk breast cancer genes and associated syndromes

Gene	Protein function	Lifetime BC risk	Associated syndrome
CHEK2	Stabilises p53 during cell cycle checkpoint control, DNA repair and cell death	25%[a]	Early onset BC and Hereditary breast & colorectal cancer (HBCC)
BRIP1	DEAH helicase that interacts with the BRCA1 C-terminal domain for DNA repair and checkpoint control	20%[b]	Fanconi's Anemia (FA)
ATM	Activates cell response to dsDNA breaks through phosphorylation of BRCA1, p53 and CHEK2	15%[b]	Ataxia-telangiectasia (AT)
PALB2	Interacts with BRCA2 during DNA repair	35%[c]	Fanconi's Anemia (FA)

a (Nevanlinna and Bartek 2006), b (Apostolou and Fostira 2013; Mavaddat, et al. 2010a), c (Antoniou, et al. 2014)

1.5.2.1. *CHEK2, PALB2, BRIP* and *ATM* genes

CHEK2 codes for a cell cycle checkpoint kinase and functions as a signal transducer for cell response to DNA damage. *CHEK2* was identified in 1999 as the first gene that is associated with a moderate risk for BC when mutated (Dong, et al. 2003). This serine threonine kinase protein phosphorylates p53 & BRCA1 and is responsible for cell cycle arrest at the point of DNA damage (Hollestelle, et al. 2010). Pathogenic *CHEK2* mutations have been discovered in Li–Fraumeni and Li–Fraumeni like families and studies have suggested that this gene may be associated with LFS (Nevanlinna and Bartek 2006; Xiang, et al. 2011). Hereditary breast and colorectal cancer is also linked to *CHEK2* gene mutations (Liu 2012; Xiang, et al. 2011). Population specific founder mutations have been discovered in many studies, confirming its association as a BC susceptibility gene (Walsh, et al. 2006; Xiang, et al. 2011). Variants in *CHEK2* confer a two-fold increase risk for BC to individuals without a family history and up to five-fold higher risk to those that are part of BC families. However, *CHEK2* mutant alleles do not completely segregate among BC families and increased risk is often associated with the presence of a combination of mutations in other genes (Hollestelle, et al. 2010).

The ataxia-telangiectasia mutated gene (*ATM*) is directly associated as the cause of the rare disorder ataxia-telangiectasia (A-T). A-T is characterised by progressive cerebellar ataxia, oculomotor apraxia and increased risk for BC (Lin, et al. 2014; Shuen and Foulkes 2011). *ATM* codes for a PI3K-related protein kinase and together with CHEK2, BRCA1 & TP53 plays a central role in orchestrating the response to DNA damage repair (Zhan, et al. 2010). ATM regulates checkpoint control by phosphorylating p53, BRCA1 & BRCA2. The absence of ATM causes aberrant cell-cycle progression and increased sensitivity to ionising radiation resulting in chromosomal breakage. A five-fold increased risk for BC is carried by family members of individuals with A-T (Apostolou and Fostira 2013; Broeks, et al. 2008; Fletcher, et al. 2010).

BRCA1 interacting protein C-terminal helicase 1 (BRIP1) mutations have been identified in FA patients and confer a two-fold risk for BC and OVC (Economopoulou, et al. 2013). BRIP1 is a helicase protein that interacts with BRCA1 during checkpoint control (Apostolou and Fostira 2013; Mavaddat, et al. 2010a). The discovery of the partner and localizer of BRCA2 (PALB2) and BRIP serves as proof that functional similarities among proteins are good leads towards identifying candidate BC genes (Casadei, et al. 2011; Filippini and Vega 2013; Hollestelle, et al. 2010). PALB2 and BRIP1 proteins both interact with BRCA1. PALB2 stabilises and localises BRCA2 during DNA repair (Economopoulou, et al. 2013; Hellebrand, et al. 2011).

1.5.3. Common genetic variants

Common variants also contribute to a smaller portion of the total amount of hereditary BC cases worldwide. Common alleles that are located in these genes are associated with a small increase in disease risk and act in a polygenic/multiplicative manner (Apostolou and Fostira 2013). Variant alleles that confer a low risk for BC development are mostly identified through genome-wide association studies (GWAS) (Long, et al. 2013; Reeves, et al. 2010; Sapkota, et al. 2013).

A number of low penetrant alleles have been identified in genes such as; *CASP8*, *TGFB1*, *FGRF2*, *TOX3*, *MAP3K1*, *LSP1*, *TNRC9* (Figure 1.5) and have been associated with BC and often confer risk for disease that is population specific (Broeks, et al. 2011; Filippini and Vega 2013). Technological and methodological advances have introduced whole genome DNA sequencing methods to increase the number of loci containing low penetrant gene mutations (Bonifaci, et al. 2008). It is challenging to prove an association between low penetrant polymorphism and disease as many variants are located in non-coding genome regions (Korde and Shiovitz 2013). However, it has been proposed that low-risk mutations primarily disrupt molecular pathways that play a role in cell communication and death (Bonifaci, et al. 2008). Low penetrant, common genetic variants may affect genes and lead to the activation cell proliferation rather than inactivating DNA repair (Korde and Shiovitz 2013). As a result, low-risk susceptibility loci are also used to predict the pathological subtype and underlying etiology of breast tumours (Broeks, et al. 2011).

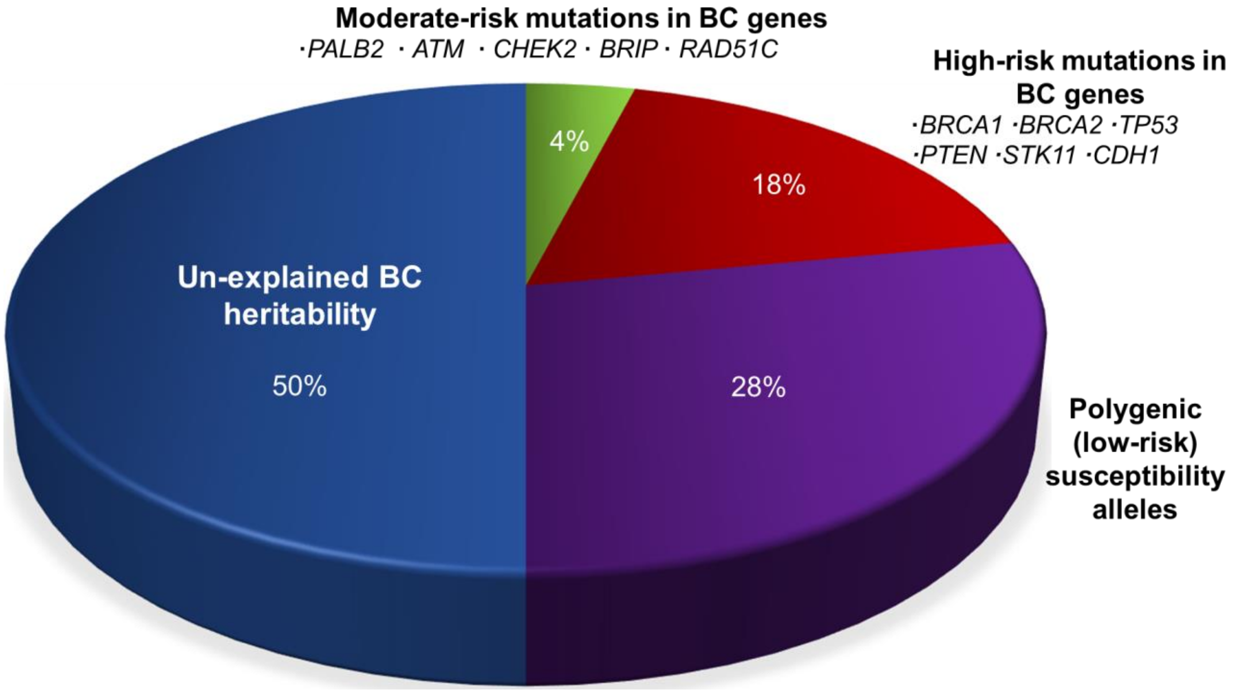


Figure 1.5: Mutations in genes associated with hereditary breast cancer.

Individuals who inherit high and moderate risk variant alleles are susceptible to hereditary breast cancer. Low-risk susceptibility alleles and variants that form part of the unexplained familial risk are often explained through the “polygenic” model (Couch, et al. 2014).

Recent use of next generation sequencing (NGS) suggests that a large portion of the missing heritability (Figure 1.5) seen in non-*BRCA* BC families may be associated with high and moderate susceptibility alleles (Gracia-Aznarez, et al. 2013). A high polygenic risk score may also explain the disease penetrance that is observed in certain BC families that do not carry mutations in known high or moderate penetrant genes (Lee, et al. 2014a; Muranen, et al. 2017).

1.5.4. The role of high and moderate breast cancer risk genes in DNA damage repair

Damaging mutations that lead to the inactivation of processes that maintain DNA integrity contribute towards the aetiology of cancer and is an important hallmark for the disease (Helleday, et al. 2007; Kwei, et al. 2010). The loss of either cell-cycle checkpoint and/or the occurrence of chromosomal instability creates a need for processes aimed at the recognition, signalling and repair of damaged DNA lesions (Filippini and Vega 2013). As illustrated in Figure 1.6, there are various types of DNA lesions that are caused by endogenous and exogenous agents. Replication forks are arrested by endogenous stresses including bulging

DNA regions and intrastrand crosslinked DNA that occur as a result of environmental stresses such as ultraviolet light exposure (Dexheimer 2013). Spontaneous hydrolysis and reactive oxygen species (ROS) molecules result in the chemical modification of DNA producing single stranded breaks, abasic sites and the deamination of nucleotides (Mehta and Haber 2014). DNA lesions such as double strand (ds) breaks and interstrand crosslinks (ICL) are caused by exposure to X-rays and radiation. These lesions are the most damaging as they generate genetic rearrangements that can result in DNA instability (Dexheimer 2013; Grabarz, et al. 2012; Mena, et al. 2009; Shuen and Foulkes 2011). Variants such as small deletions, insertions and mismatch mutations that cause premature protein truncation or nonsense mediated decay of RNA occur as a result of replication errors (Stephens, et al. 2009; Stephens, et al. 2012). DNA damage repair plays a tumour suppressive role as it is central to preventing endogenous replicative stress (Helleday, et al. 2007).

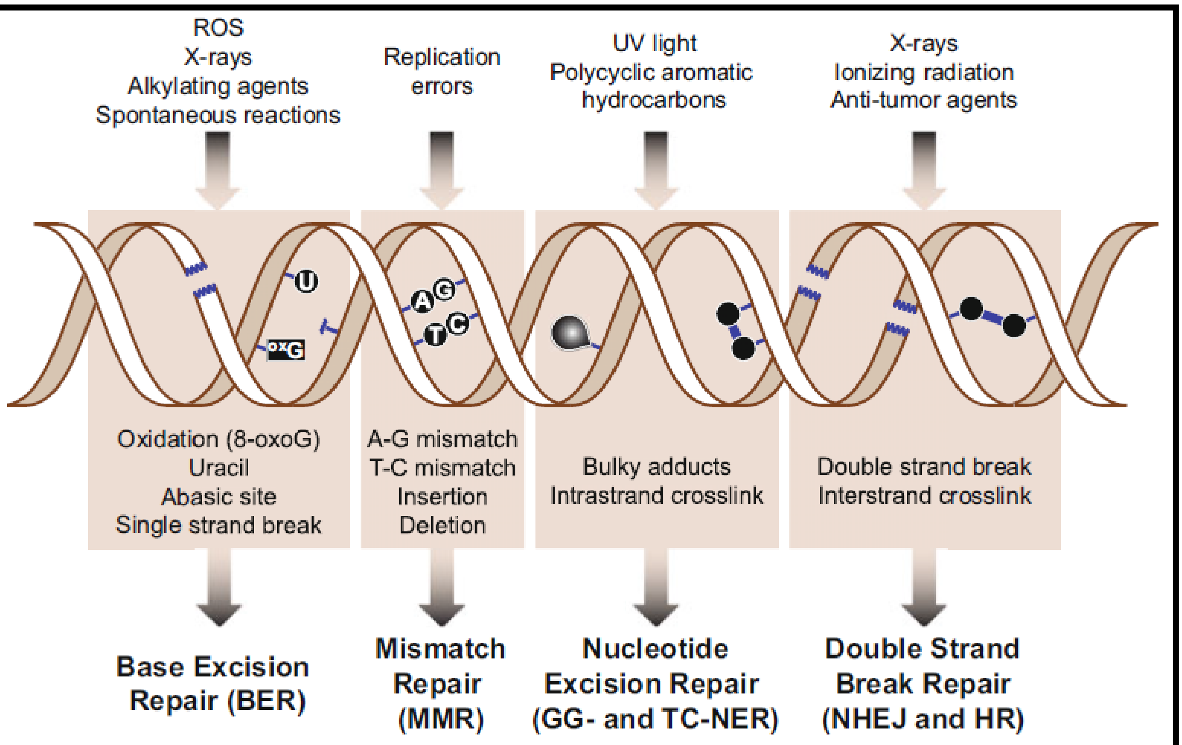


Figure 1.6: Four classes of DNA damage and corresponding DNA repair mechanisms.

Taken from Dexheimer 2013.

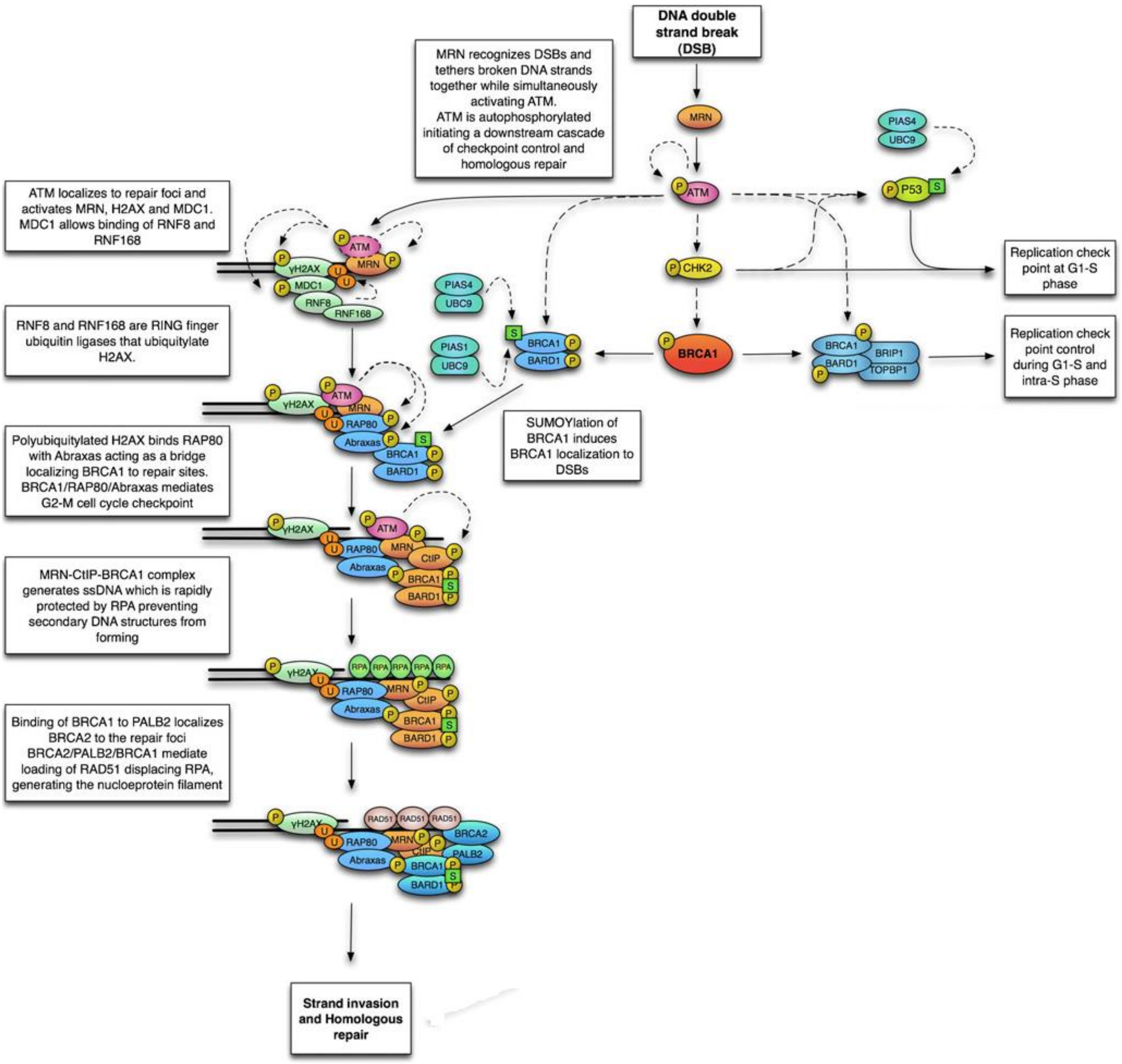
DNA damage signalling and repair mechanisms are comprised of a series of proteins that interact directly or indirectly with DNA structures (Milanowska, et al. 2011; Polo and Jackson

2011). These pathways are connected to each other at multiple points (e.g. recognition of DNA damage) (Shrivastav, et al. 2008). The process of DNA damage repair can be divided into five main categories, each containing different methods used to recognise and repair particular DNA lesions (Figure 1.6) (Dexheimer 2013). Base excision repair (BER) is the predominant mechanism used to repair small non-bulky DNA lesions without distorting the overall helical structure. BER is a multistep process with many entry points that are dependent on the type of damage that is encountered (Kim and Wilson 2012). This pathway makes use of lesion-specific DNA glycosylases that remove damaged base(s) through a base flipping mechanism. The excised base-site is then repaired by first replacing/hydrolysing the phophodiester backbone in preparation for DNA polymerase and subsequent ligase reactions (Kunz, et al. 2009). Mismatch repair (MMR, Figure 1.6) plays an important role in correcting miss-incorporated bases that have escaped the proofreading capability of polymerase enzymes, insertion/deletion loops and polymerase slippage during replication of repetitive sequences (Fukui 2010). MMR is highly conserved and consists of three principal steps. These include; recognition of mispaired bases, excision and degradation of the mismatched strand and synthesis or “filling” of DNA gaps caused by the excision step (Dexheimer 2013).

Nucleotide excision (NER) repair removes any bulky lesions that may disturb the DNA helix structure. BER shares many similarities with this process, however, NER requires up to 30 different proteins to perform a multistep mechanism (Cabelof 2012; Dexheimer 2013). Steps include the recognition of DNA damage, opening the helix and excision of short single strand fragments in regions that span the lesion. Synthesis repair and strand ligation is then used to restore the damaged strand (Dexheimer 2013). The DNA damage recognition complex, MRN (MRE11–RAD50–NBS1), plays an important role in detecting DNA damage and activating the ATM kinase activity. MRN binds to the broken ends of DNA strands and undergoes multiple conformational changes that increases the affinity of ATM to its substrates (Filippini and Vega 2013). MRN interacts with BRCA1 and co-synthesizes single-stranded DNA that is needed for homologous DNA repair (Figure 1.7). The repair process is activated on recognition of interstrand DNA crosslinks that prevent the replication or transcription of DNA (Shuen and Foulkes 2011).

The loss of members of the MRN complex predisposes individuals to disorders such as ataxia-telangiectasia-like disorder (MRE11^{-}), Nijmegen breakage syndrome (NBS1^{-}) and NBS-like disorder (RAD50^{-}) (Filippini and Vega 2013). *RAD50* germline variants predispose individuals to BC and this association has been validated in a number of case-control studies. Carriers of mutations in *RAD50* genes have approximately four-fold increased risk for BC and those heterozygous for *NBS1* variants carry a two to three-fold risk (Hollestelle, et al. 2010).

MRE11 gene variants have also been identified, however, its association with BC has not been conclusively verified (Shuen and Foulkes 2011).



DNA double strand break (DSB)

MRN recognizes DSBs and tethers broken DNA strands together while simultaneously activating ATM. ATM is autophosphorylated initiating a downstream cascade of checkpoint control and homologous repair

ATM localizes to repair foci and activates MRN, H2AX and MDC1. MDC1 allows binding of RNF8 and RNF168

RNF8 and RNF168 are RING finger ubiquitin ligases that ubiquitylate H2AX.

Polyubiquitylated H2AX binds RAP80 with Abraxas acting as a bridge localizing BRCA1 to repair sites. BRCA1/RAP80/Abraxas mediates G2-M cell cycle checkpoint

SUMOYlation of BRCA1 induces BRCA1 localization to DSBs

Replication check point at G1-S phase

Replication check point control during G1-S and intra-S phase

MRN-CtIP-BRCA1 complex generates ssDNA which is rapidly protected by RPA preventing secondary DNA structures from forming

Binding of BRCA1 to PALB2 localizes BRCA2 to the repair foci BRCA2/PALB2/BRCA1 mediate loading of RAD51 displacing RPA, generating the nucleoprotein filament

Strand invasion and Homologous repair

Figure 1.7: DNA damage recognition and response.

This figure illustrates the recognition of DNA damage (MRN complex) and the recruitment of proteins aimed at the repair of double strand DNA lesions. Phosphorylation, ubiquitylation and SUMOylation events are represented by dashed lines (with permission from Shuen and Foulkes 2011).

Two processes that assist in the repair of double strand breaks (DSB) include homologous recombination (HR) and non-homologous end joining (NHEJ) (Shrivastav, et al. 2008). DSB’s are used to generate genetic diversity during normal processes such as meiosis and rearrangement of immune genes. DSB's can be hazardous to the cells and if left unrepaired, induces cell death and is associated with cancer development (Grabarz, et al. 2012). As illustrated in Figure 1.7, HR is initiated by the presence of ssDNA 3' tails that are coated by the ssDNA binding protein, RPA (Shuen and Foulkes 2011). A BRCA2-PALB2 complex is formed, removes the RPA binding protein and loads the pivotal HR protein, RAD51 (Grabarz, et al. 2012).

Activation of the RAD51 protein promotes the exchange of DNA strands between sister chromatids (Holthausen, et al. 2010). HR must be controlled as it has a direct effect on genome stability (Grabarz, et al. 2012). NHEJ differs from HR as it is not dependent on sharing homologous sequences with intact chromosome partners (Shrivastav, et al. 2008). Alternative end-joining (A-EJ) is another repair pathway used to correct DNA strand breaks. A-EJ is similar to HR but requires a smaller number of homologous sequence and is a less controlled pathway which makes it highly mutagenic (Grabarz, et al. 2012; Polo and Jackson 2011). Individuals that carry defects in the DNA damage repair (DDR) process are susceptible to tumour development (Grabarz, et al. 2012).

1.5.4.1. The Fanconi Anemia pathway

The genetic disease, FA, is associated with mutations in genes that comprise the FA pathway. The main function of the FA pathway is to coordinate distinct pathways that resolve interstrand cross-linked DNA regions (ICL) (Filippini and Vega 2013; Kee and D'Andrea 2010). ICL's cause lesions in the DNA double strand helix and block the progression of cellular replication and transcription (Deans and West 2011). This pathway contains elements of three classic DNA repair processes including NER, the error-prone translesion synthesis (TLS) and HR to remove cross-linked DNA (Kim and D'Andrea 2012). There are 15 genes that are associated with FA including *FANCA–C*, *D1* (*BRCA2*), *D2*, *E–G*, *I*, *J* (*BRIP1*), *L* (*PHF9*), *M* (*Hef*), *N* (*PALB2*), *RAD51C* (*FANCO*) and *SLX4* (*FANCP*) (Hucl and Gallmeier 2011).

The FA pathway remains dormant and is only activated during the cell cycle synthesis phase (Garner and Smogorzewska 2011). During the cell cycle S-phase or upon recognition of DNA damage a kinase protein, such as ATR kinase, phosphorylates the FA core proteins. This core complex consists of eight proteins i.e. A, B, C, E, F, G, L, M and is stabilised by other weaker sub-complexes for e.g. partial interactions with additional FA-associated proteins such as FAAP24 and FAAP100 (Moldovan and D'Andrea 2009; Shuen and Foulkes 2011). Together, these proteins assemble into a multi-subunit ubiquitin ligase complex (Hucl and Gallmeier 2011). FAAP24 directly recognises cross-linked DNA and interacts with FANCM to open DNA strand replication forks and recruits the FA core complex to ubiquitinate FANCD2 and FANCI (Moldovan and D'Andrea 2009). Unique DNA repair structures (i.e. foci) are formed under the direction of the ubiquitinated FANCD2 & FANCI (Hollestelle, et al. 2010). To link the FA pathway with HR, FANCD2 co-localises with FANCD1, FANCN, FANCJ and RAD51 (Figure 1.8) (Shuen and Foulkes 2011). Specific DNA structures with FANCD2-I, FANCD1 (BRCA2), RAD51 and PCNA are vital towards promoting homologous DNA repair (Kee and D'Andrea 2010). The FANCJ-BRIP1 complex has a 5’ – 3’ helicase activity that localises to DNA repair structures associated with BRCA2 and RPA and remodels DNA structures, regulating HR and other repair processes (Moldovan and D'Andrea 2009).

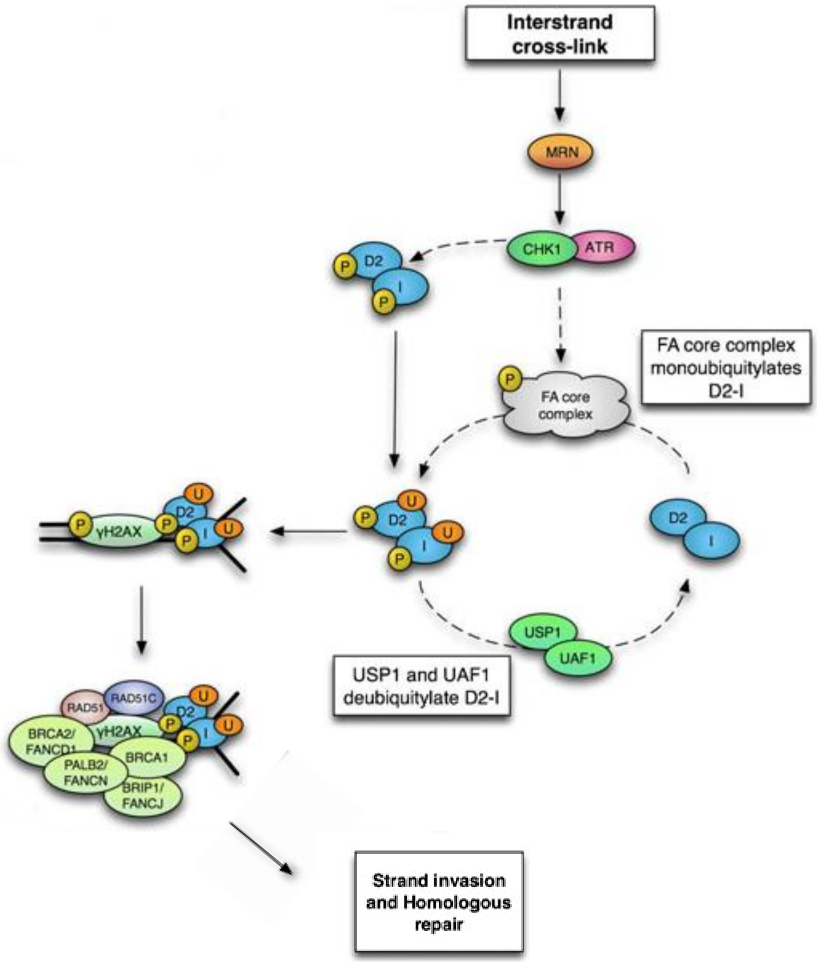


Figure 1.8: Fanconi Anemia pathway

This figure illustrates the recruitment of the FA core complex towards the repair of interstrand crosslinks. The FA pathway generated dsDNA breaks that ultimately leads to homologous DNA repair. Adapted from Shuen and Foulkes 2011.(Shuen and Foulkes 2011)

FA is characterised by aplastic anemia, congenital defects and bone marrow failure (D'Andrea 2010; Hucl and Gallmeier 2011). However, heterozygous mutations in components of the FA pathway; i.e. *FANCD1* (*BRCA2*) *FANCJ* (*BRIP1*), *FANCM* (*Hef*), *FANCN* (*PALB2*) and *FANCO* (*RAD51C*) confers a risk for breast and ovarian cancer (Filippini and Vega 2013).

1.6. The “missing” hereditability of familial breast cancer

Understanding the molecular roles of BRCA1/2 in the DNA repair mechanism has helped identify mutations in other BC genes that form part of the same pathway (Shuen and Foulkes 2011). As illustrated in Figure 1.5, mutations in known cancer predisposing genes collectively account for ~50% of global familial cases ((Casadei, et al. 2011; Filippini and Vega 2013; Gracia-Aznarez, et al. 2013; Njiaju and Olopade 2012; Osher, et al. 2012; Thompson, et al. 2012).

Many familial BC/OVC cases do not carry high-risk mutations in disease susceptibility genes identified thus far, i.e. *BRCAx* families (Collins and Politopoulos 2011). Ongoing research studies are aimed at identifying reasons for this missing heritability (Shuen and Foulkes 2011). Next generation technologies are at the forefront of discovering additional genetic variation that may account for the etiology of *BRCAx* cancer families (Gracia-Aznarez, et al. 2013; Thompson, et al. 2012). Whole Exome Sequencing (WES) is a next generation sequencing method that has been used to identify rare genetic variants in coding regions that are associated with high-risk for breast cancer (Park, et al. 2012; Robinson, et al. 2011). As common variants have a smaller effect on disease risk the identification of rare gene variants may account for the missing heritability in high-risk *BRCAx* families (Collins and Politopoulos 2011).

1.7. Next-generation sequencing methods used to discover novel breast cancer genes

1.7.1. Recent advances in DNA sequencing

Traditional sequencing methods have advanced in recent years to facilitate automation and increase throughput (Moorthie, et al. 2011). These “second generation” sequencing methods are partially replacing gold standard techniques such as traditional Sanger sequencing, especially for research purposes (Laura 2010). Second generation sequencing is classified by clonal amplification of DNA fragments prepared from isolated samples. Sufficient amplification is required for the purpose of signal detection during base calling (Moorthie, et al. 2011). NGS technologies have been made available through multiple platforms, each with different underlying chemistries. One of these processes, the cyclic reversible termination (CRT) method, resembles Sanger sequencing (sequencing-by-synthesis) and has been commercialised by Illumina (Genome analyser & HiSeq systems). CRT makes use of clonally amplified single stranded template DNA and detects fluorescently labelled nucleotide bases as they are incorporated into the growing DNA strand (Chen, et al. 2013). Sequencing by ligation (SBL) is another cyclic method but differs from CRT as sequences are obtained by hybridising fluorescently-labelled nucleotide base probes to DNA templates (Metzker 2010). Pyro-sequencing is a bioluminescence method that measures the amount of light emitted from bases that are incorporated into complementary sequence positions (Kircher and Kelso 2010).

The costs per base of massively parallel sequencing (MPS) has significantly decreased from previous expensive rates, however, sequencing the whole genome is still not a viable approach for most clinical applications (Ernani and LeProust 2009; Laura 2010). Consequently, recent studies are employing a more targeted approach of interrogating the human genome i.e. WES (Need, et al. 2012a). WES targets the coding region (exome) making the majority of the variants that are identified more interpretable (Robinson, et al. 2011). There are three general techniques used to select the desired coding regions for sequencing, namely; PCR-based, molecular inversion probe ligation sequencing (MIPS) & oligonucleotide hybridization-based methods (Hedges, et al. 2011; Moorthie, et al. 2011).

The performance and utility of different sequencing techniques are assessed by a number of matrices including, raw sequence quality, system errors, read depth and read coverage, etc. These factors are used to support the certainty of variant calls (Moorthie, et al. 2011). Data analysis of MPS results can be divided into three main phases; primary analysis, which is performed on an automated machine and includes base calling (DePristo, et al. 2011). The secondary phase includes read alignment and variant calling followed by the biological interpretation of gene variants that forms part of the tertiary phase (Metzker 2010). Continued assessment of these technologies have resulted in the development of a more streamlined approach for variant identification (Challis, et al. 2012; DePristo, et al. 2011).

1.7.2. High-throughput data analysis and variant discovery

NGS data processing deals with large volumes of sequences that use multiple bioinformatics tools, require sufficient computational support and the most accurate genomics databases (Moorthie, et al. 2011). The amount of variants identified through whole genome sequencing (WGS, ~five million) outweighs that of targeted approaches such as WES (i.e. ~12 000 variants) (Pabinger, et al. 2013; Thompson, et al. 2012). Whole exome sequencing still produces large amounts of data, however, recent advances in bioinformatic tools has made the analysis of WES data more manageable making it the more favourable sequencing strategy (Bamshad, et al. 2011; Laura 2010).

High-throughput data analysis is still being optimised and poses many challenges (Ding, et al. 2010; Pop and Salzberg 2008). A plethora of tools are carefully considered in the workflow and are specific to the setup of each experiment. These tools must be easily accessible, constantly updated and should support standard data formats (Pabinger, et al. 2013). Bioinformatics analysis is comprised of five distinct steps i.e.; quality assessment, read alignment, variant identification, - annotation and – visualisation (Van der Auwera, et al. 2013). Quality assessment is the first step taken towards obtaining corrected, analysis-ready data (García-Alcalde, et al. 2012). Raw reads are trimmed to remove sequence artefacts, poor quality reads as well as adapter and primer contamination (Patel and Jain 2012). Reads are then aligned to the most updated version of the human reference genome (Li and Durbin 2009). It is recommended that paired-end (rather than single-ended) DNA libraries are made during initial sample preparation. This will incorporate bi-directional sequences that will reduce the amount of miss-aligned reads and the possibility of making ambiguous variant calls (Li and Homer 2010). Multiple duplicate reads and those with many mismatches are then discarded before variant discovery (Li and Durbin 2009).

Variant identification tools are divided into four categories, i.e.: germline & somatic mutation callers or copy number and structural variant identifiers. Mutation discovery tools are either freely accessible web-based tools or commercial MPS analysis suites (Fischer, et al. 2012; Ji 2012). Variant annotation then follows and predicts the structural and/or functional effect of mutations. This is also achieved by using either offline or web-based applications (Chang and Wang 2012; Pabinger, et al. 2013; Wang, et al. 2010). Offline tools are not dependent on access to internet services but require a certain degree of computational and programming skills. Web-based programs are more user-friendly and can be used by molecular biologists (Goecks, et al. 2010; Paszkiewicz and Studholme 2012). Lastly, NGS visualisation is used to display sequence data and is useful for the interpretation of results (Pabinger, et al. 2013; Sanborn, et al. 2011).

Data obtained through NGS make use of two main means of analysis (Pabinger, et al. 2013). Analytical pipelines have a defined set of steps that are based on built-in algorithms and uses raw input data to systematically generate variants (e.g. Genome Analysis Toolkit) (Van der Auwera, et al. 2013). Workflow management systems are comprised of analytical steps that can be modified for data manipulation (e.g. Galaxy) (Goecks, et al. 2010). The challenge of applying high-throughput sequencing methods for disease gene discovery lies in implementing an efficient and comprehensive framework to narrow down the number of variants (Li, et al. 2012). Various approaches are being researched to prioritise for the most deleterious variant in genes and identify the causal mutation (Bodmer and Tomlinson 2010; DePristo, et al. 2011; Li, et al. 2012; Tranchevent, et al. 2011).

1.7.3. Evaluation of candidate genes through variant filtration and prioritisation

NGS can be applied for various reasons namely; i) identifying causative variants in Mendelian disorders (i.e. germline mutations), ii) investigating candidate genes that are linked to diseases for later functional studies, iii) identifying driver and passenger mutations as somatic variants in cancer tissues (Li, et al. 2012; Robinson, et al. 2011; Vogelstein, et al. 2013). Whole genome and exome sequencing methods both produce a great number of variants. The amount of variants are reduced for further investigation by filtering variants according to selection criteria that are biologically meaningful to a study (Feng, et al. 2011).

To discover variants that confer a high risk for disease, some of the following selection processes can be followed. Choosing individuals from specific ethnic and/or locations could lead to the discovery of founder mutations that have higher allele frequencies within such closed population groups (Chong, et al. 2012). Patients with a strong family history and/or early disease-onset are selected in order to identify rare variants with a detectable increase in relative risk (Feng, et al. 2011). Variants suitable for individual assessment are characteristically found in higher numbers among index cases and relatives (Bodmer and Tomlinson 2010).

Selecting variants that may confer a high-risk predisposition to cancer can be achieved by filtering variants based on mutation type, variant call quality and allele frequency (Stitziel, et al. 2011). Making use of public databases such as the 1000 Genomes project, the Exome Aggregation Consortium (ExAC) and the National Heart, Lung, and Blood Institute Exome Sequencing Project (NHLBI-ESP), may assist in excluding alleles that are more common in the general population (Consortium 2010; Fu, et al. 2013; Lek, et al. 2016). Dominant, high-risk variants that are deleterious to protein function have allele frequencies that range between 0.1 – 1% (Bodian, et al. 2014). Assessing the effect that mutations have on protein function is determined by predictive protein analysis programs and is limited by the validity of functional data (Thomas and Bonchev 2010). The most significant deleterious variants are those that result in protein truncation (splice-site disruption, stop-gain, insertion-deletions) or copy number alterations and subsequent nonsense mediated decay (Johansson, et al. 2012; MacArthur, et al. 2014). Missense variants affect protein function by altering protein-protein interactions and variations in promoter regions may affect gene expression. The clinical and epidemiological consequences of single nucleotide polymorphisms (SNP) are more difficult to validate and the majority are located in non-coding genome regions (i.e. introns or non-coding exon) (Bodmer and Tomlinson 2010). Meta-analysis studies are used to investigate the role that SNP's play in cancer-risk (Bodian, et al. 2014; Shen, et al. 2011).

Variant filtration is followed by gene prioritisation; this identifies the most probable and biologically relevant candidate for further experimental investigation (Moreau and Tranchevent 2012). Gene prioritisation is aimed at producing the most probable candidates that may be linked to essential molecular pathways and the maintenance of cellular integrity (Walsh, et al. 2011). Mutations in promising candidate genes are then verified to rule out the possibility of investigating false positive variants. Thereafter, variants are evaluated to determine if they segregate with the disease in high-risk BC/OVC families or in larger cohorts of patients without a family history for the disease (Pabinger, et al. 2013). In both instances variants should be higher in cases than controls. Screening large population cohorts forms an association study that requires the participation of many individuals for increased statistical power (Bodmer and Tomlinson 2010). Functional assay studies are then used to determine the clinical implication of gene variants on protein function (Pabinger, et al. 2013). Assessing the functional effect that variants have on protein stability and function forms another layer of candidate gene validation and can be performed *in silico* (Vazquez, et al. 2012).

The continued use of genome and exome sequence technologies have resulted in more streamlined strategies of mining high throughput data. This is very important for the discovery of disease-causing variants (Challis, et al. 2012; Koboldt, et al. 2013; Pabinger, et al. 2013; Stitziel, et al. 2011).

1.8. Research study motivation and project aim

Currently, mutations in known high and moderate BC susceptibility genes do not account for all familial cases of this disease. It is estimated that the underlying cause of ~40-50% of familial BC cases are unknown (Blay, et al. 2013; Janavičius 2010). In SA, similar numbers are seen as depicted by research conducted at various local institutions including the cancer research laboratory at the University of Pretoria. Approximately 30% of high-risk BC families do not carry mutations in *BRCA* genes (J van Rensburg, et al. 2007; Loubser, et al. 2012; Reeves, et al. 2004; Schlebusch, et al. 2010; van der Merwe, et al. 2012; Vorobiof, et al. 2001). Variants with moderate penetrance account for 1% - 2% of early-onset breast cancer (Dorfling, et al. ; Sluiter, et al. 2009). The absence of mutations in known BC predisposition genes in high-risk BC cases, suggests the presence of novel genes with variants that may predispose individuals to this disease. Mutations in such genes may display a similar inheritance pattern as known high- to moderate-risk BC susceptibility genes.

High throughput whole exome sequencing may assist in identifying additional BC risk genes. This DNA analysis method has advanced greatly in recent years and has become a more plausible strategy to rapidly identify BC predisposing mutations. Novel rare disease risk loci with high penetrance may account for the numbers of unexplained *BRCAx* families with multiple affected individuals in each generation.

1.8.1. Project aim

To discover and characterise novel gene variants through whole exome capture sequencing of SA patients that do not carry pathogenic *BRCA*-gene mutations from families with a strong history of the disease.

1.8.1.1. Objectives

I. Select individuals (negative for *BRCA1* and *BRCA2* mutations) from SA families with a history of BC/OVC (≥ three cases). Submit DNA samples to the Beijing Genomics Institute for whole exome capture sequencing (WECS) on the Illumina platform.

   a. Generate high quality sequence reads for variant discovery from raw sequence data through systematic data mining.

II. Identify and annotate gene variants with the use of tools from the Genome Analysis Toolkit (GATK).

   a. Obtain a list of prioritised functionally deleterious variants through a hierarchical set of steps using *in silico* analysis, literature mining and computational systems biology methods.

III. Verify variants in genes of interest through Sanger sequencing and generate a refined list of high priority genes for further analysis.

IV. To validate candidate genes towards the discovery of novel high-penetrant BC predisposition genes in population cohort of high risk BC/OVC families in S.A.

V. Determine the functional implication of validated novel genes with the use of computational systems biology.

Chapter 2:

Materials and Methods

2.1. Patient selection

2.1.1. Ethics

Ethical approval for this study was granted by the University of Pretoria, Faculty of Health Sciences, Ethics Review Board (173/2012, Annexure 1). All DNA samples were collected after each individual had given written consent. All index cases included in the study previously screened negative for *BRCA1/2* mutations, using SSCP/Heteroduplex analysis and PTT. Furthermore, these samples also screened negative (with Sanger sequencing) for pathogenic variants in *PALB2* and *RAD51C*.

2.1.2. Selection for whole exome sequencing

Whole exome sequencing (WES) was performed on a total of nine South African patients from six high-risk families (≥ 3 breast and/or ovarian cancer cases, Annexure 2). These comprised of one person for each of three families and two persons from each of three families.

2.1.3. Selection for targeted screening of candidate genes

Once promising candidate genes were identified with WES, 61 samples from 56 additional high-risk (*BRCAx*) families were analysed in an effort to determine their potential role in breast cancer susceptibility. These families comprised 10 (~17%) with a history of breast and ovarian cancer, two (~3%) with ovarian cancer only and the remaining were breast cancer families. All of the families are of European descent with 43 families that have Afrikaner ancestry, five with Ashkenazi Jewish ancestry and seven families that originate from the United Kingdom.

2.2. Sample preparation and WE sequencing

The Cancer Genetics Research Group, University of Pretoria have previously extracted blood DNA samples from all the participants in this study using a slightly adapted version of the method described by Johns and colleagues (Johns Jr and Paulus-Thomas 1989). Approximately 3µg of the extracted sample was sent to the Beijing Genomics Institute (BGI, North China) where DNA was sheared and paired-end DNA libraries prepared with 150bp insert sizes. The Agilent SureSelect Human All exon version 4 (51Mb) (Santa Clara, CA, USA) sequence capture method was used to enrich for all known exome regions within the human genome. Illumina sequencing with the HiSeq 2000 genome analyzer IIx (Illumina, San Diego, CA, USA) then followed in order to generate paired-end reads. Raw sequence data, received from the institute, was processed for variant discovery.

2.3. Data processing

Scripts compiled for whole exome sequence data analysis purposes have been described in Annexure 3. As illustrated in Figure 2.1, raw sequence base call quality scores were reviewed with the Java-based tool, FastQC (Babraham Institute, Cambridge, UK). Reads were trimmed with the FastX read processing tool to filter bases with PHRED Scaled Quality scores <30. Trimmed paired-end sequence reads were mapped to the human reference genome sequence (hg19, build 37) downloaded though NCBI (National Centre for Biotechnology Information) with the use of the Burrows-Wheeler Alignment tool (BWA, (Li and Durbin 2009)). Duplicate reads were removed with Picard (http://picard.sourceforge.net/) followed by local realignment and quality score recalibration with software from the Genome Analysis Toolkit (GATKvs2.8, (Van der Auwera, et al. 2013)). Target enrichment efficiency assessment then followed using the CLC Genomics workbench software package (CLCBiov6, Aarhus, Denmark). Sequence coverage statistics were collected from mapped reads and expressed as the total number of reads that map uniquely to targeted intervals. Variant calling was performed with the Unified Genotyper and parameters were set to identify confident variant calls with a PHRED-scaled score of 50 and maximum coverage of up to 1000X.

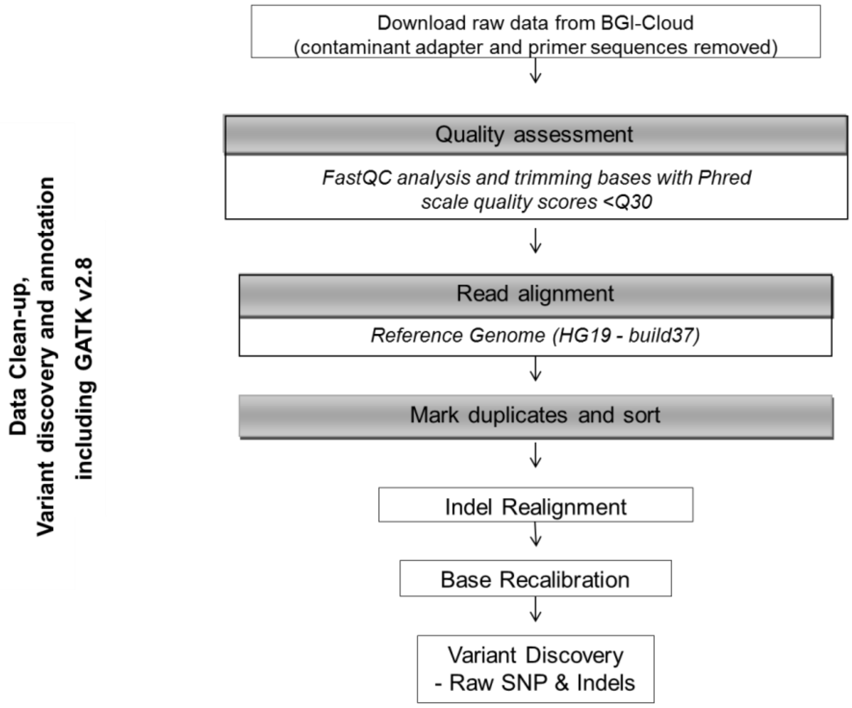

Download raw data from BGI-Cloud
(contaminant adapter and primer sequences removed)
Quality assessment
FastQC analysis and trimming bases with Phred scale quality scores <Q30
Read alignment
Reference Genome (HG19 - build37)
Mark duplicates and sort
Indel Realignment
Base Recalibration
Variant Discovery
- Raw SNP & Indels
Data Clean-up,
Variant discovery and annotation
including GATK v2.8

Figure 2.1: Outline of the data processing workflow.

2.4. Variant filtration and annotation

To generate a set of highly accurate base calls, variant recalibration (Figure 2.2) was performed by the variant recalibrator two-step process. These tools make use of the variant.vcf file as an input. The variant recalibrator generated an adaptive error model based on true gene variants obtained from HapMap 3.3 and polymorphic sites from the Omni 2.5M SNP chip array, for humans. Variants from our dataset were compared to the known true variants and those with lower quality scores than the internally determined quality score threshold (VQSLOD) were “flagged” for removal. Variant records were manually filtered as well to label mutations that pass the following criteria: i.e. located in regions of acceptable read coverage (≥10X), possess an allele count >0.15, quality by depth score <2, variant call quality ≥30 and genotype likelihood scores that favour heterozygous (AB) genotypes. The output file generated was then annotated with ANNOVAR (Wang, et al. 2010).

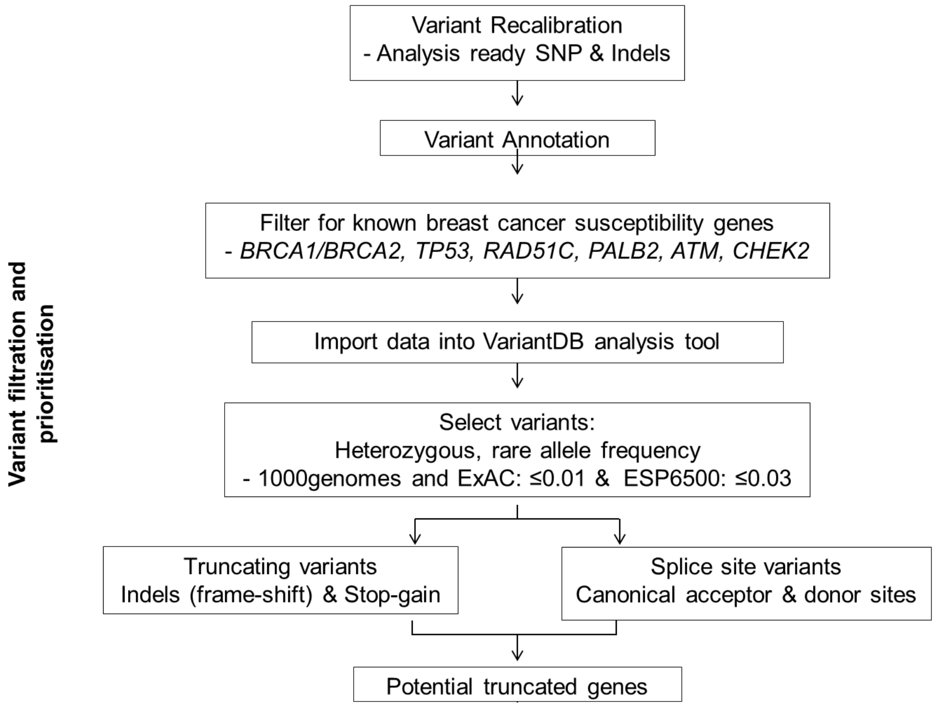

Variant filtration and prioritisation
Variant Recalibration
- Analysis ready SNP & Indels
Variant Annotation
Filter for known breast cancer susceptibility genes
- *BRCA1/BRCA2, TP53, RAD51C, PALB2, ATM, CHEK2*
Import data into VariantDB analysis tool
Select variants:
Heterozygous, rare allele frequency
- 1000genomes and ExAC: ≤0.01 & ESP6500: ≤0.03
Truncating variants
Indels (frame-shift) & Stop-gain
Splice site variants
Canonical acceptor & donor sites
Potential truncated genes

Figure 2.2: Outline of the variant filtration and prioritisation process.

2.5. Computational approach for candidate gene prioritisation

Manually filtered variant calibrated files were imported to the VariantDB analysis tool (Figure 2.2) (Vandeweyer, et al. 2014). Gene mutations were categorized according to their genomic location and predicted effect on the protein-coding region. Those identified in intronic, intergenic, and untranslated regions were removed. Genes that contain deleterious mutations (i.e. frameshift indels, canonical splice-site and nonsense mutations) were prioritised for further consideration (Figure 2.2). Deleterious mutations were further triaged to include variants that were novel (i.e. mutations absent from dSNP and/or 1000genomes databases), known disease-associated variants (OMIM, HGMD, candidate gene association studies), variants present in various cancer genome databases with minor allele frequencies (MAF) ≤0.01 (dbSNP, ExAC and 1000genomes), ≤0.03 (ESP6500) and mutants that have previously been implicated in cancer gene association studies (Figure 2.2). The importance of non-synonymous variants was determined by the number of functional predictive algorithms indicating a damaging/pathogenic variant effect. Functional prediction scores derived from four data sources were incorporated during variant annotation (Chang and Wang 2012; Liu, et al. 2011). Predictions were extracted from the dbNSFP database and variants were classified as follows: LJB SIFT>0.95 were d(eleterious), LJB PolyPhen2>0.15 & >0.85 were either potentially (P) or probably (D) damaging, LJB GERP++_RS>4.4 and LJB phred CADD>20 were classified as pathogenic (Kircher, et al. 2014). Missense variants were included if predicted pathogenic by all four *in silico* tools.

Genes were prioritised by manually filtering for components of molecular pathways such as the regulation of genomic stability (i.e. DNA damage recognition & repair) and cellular integrity (e.g. proliferation and apoptosis). In cases where two individuals in a family were used for exome sequencing we classified high-priority gene variants as mutations that were shared by both cases (Figure 2.3). Variants that qualified as biologically interesting but were not shared by two relatives were included as well. Our gene lists were cross-referenced with somatic variant data derived from breast cancer resources including COSMIC, cBioPortal & the International Cancer Genome Consortium (ICGC) data portal and the ClinVar database (Forbes, et al. 2011; Gao, et al. 2013; Landrum, et al. 2016). High-priority gene variants were selected for further analysis.

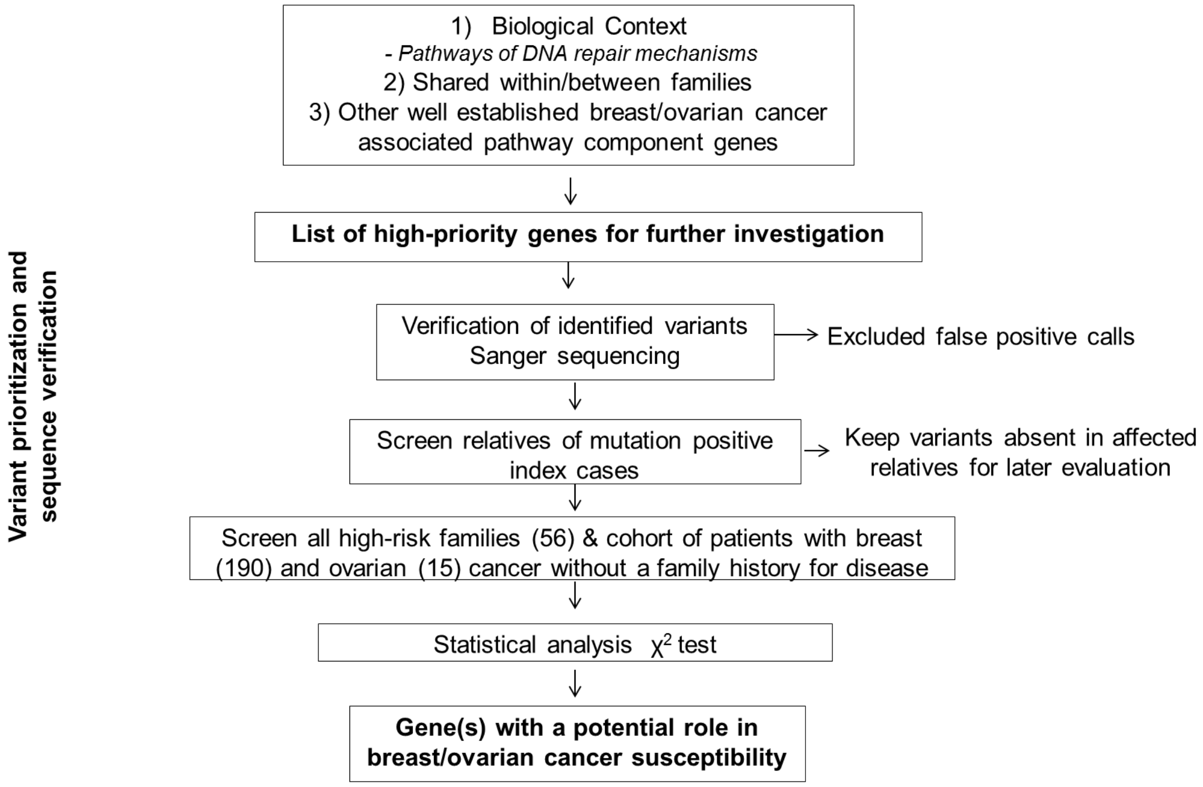

Variant prioritization and sequence verification
1) Biological Context
- Pathways of DNA repair mechanisms
2) Shared within/between families
3) Other well established breast/ovarian cancer associated pathway component genes
List of high-priority genes for further investigation
Verification of identified variants Sanger sequencing
Excluded false positive calls
Screen relatives of mutation positive index cases
Keep variants absent in affected relatives for later evaluation
Screen all high-risk families (56) & cohort of patients with breast (190) and ovarian (15) cancer without a family history for disease
Statistical analysis χ2 test
Gene(s) with a potential role in breast/ovarian cancer susceptibility

Figure 2.3: Variant prioritization pipeline and sequence verification.

2.6. Sanger sequencing

2.6.1. Confirmation of WES sequence variants

All variants classified as functionally deleterious after candidate gene prioritisation, were validated with standard Sanger sequencing (Figure 2.3). This was conducted with DNA from BC/OVC patients and unaffected control samples. First the specific variants, identified through WES analysis in index cases, were verified. To confirm that variants of interest were not sequence artefacts; target regions were confirmed through bi-directional sequencing. Samples from the additional individuals from the set of six families were also selected to validate mutations. Mutations identified through whole exome sequencing were labelled as true positive variants if found in the index cases sequenced as well as their affected family members (where DNA was available).

2.6.2. Investigating the potential role of high-priority candidate genes in breast cancer susceptibility

The genes containing true positive variants were further verified. Uni-directional sequencing was utilized to screen for germline mutations in the entire coding region of prioritised genes in patients from 56 high-risk BC/OVC families.

2.6.3. Amplification and sequencing of target regions

2.6.3.1. Amplification and PCR clean-up

Primer pairs flanking intron-exon boundaries were designed with Primer3 Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Primer synthesis was achieved at Invitrogen (Life Technologies, CA, USA). Target exon regions were amplified using genomic DNA (25ng) with the Bioline high fidelity enzyme (Bioline Inc, MA, USA). Each 20µl PCR reaction contained $MgCl_2$ ranging between [1-1.5mM], 1 x reaction buffer, [0.025mM] dNTP, 0.5U enzyme and [0.2µM] forward/reverse primer (Invitrogen, CA, USA). Thermal cycling protocols were conducted with annealing temperatures (Tann) optimised according to the calculated temperature (Ta) for each primer pair ((%GC x 0.41) + 34.9). An initial denaturation step at 94^{o}C for 3 min was followed by 35 cycles of amplification that was comprised of dsDNA denaturation at 94^{o}C for 1 min, primer annealing for 1 min and extension for 3 min at 72^{o}C. Aliquots of each reaction (i.e.5µl) was loaded onto a 1.5% agarose gel containing 1 x syber safe (Thermo Scientific Inc., MA, USA) and resolved in 1 x TBE (Tris Borate EDTA, ph8) at 80V, for 25 to 35 min. Fragment sizes were verified against the 0.3µg of a 100bp DNA marker (Thermo Scientific Inc). Successfully amplified PCR products were purified with Illustra ExoProStar 1-Step PCR and Sequence Reaction Clean-up kit (GE Healthcare Life Sciences, UK). Each 7.5µl reaction mixture contained 5µl amplicon and 2µl Illustra™ ExoProStar™ 1-step that was incubated at 37^{o}C for 15 minutes and 80^{o}C thereafter. The purified product was used for downstream automated Sanger sequencing.

2.6.3.2. Cycle sequencing

A 1:3 dilution was made of purified PCR products in H_2O and after splitting this dilution 1µl of a 20µM primer solution was added. Three microliters of the DNA-primer mixture was added to 7µl of a BigDye v3.1 dilution for sequencing (ABI Big Dye terminator sequencing kit, Applied Biosystems, CA, USA). Utilising the BigDye terminator v3.1 chemistry; 25 rounds of cycle sequencing was performed including denaturation at 96^{o}C for 10 seconds, annealing at primer specific temperatures for five seconds and extension at 60^{o}C for four minutes.

Precipitation of each sequencing reaction was achieved in a 100µl final volume comprised of 0.09M sodium acetate (pH4.75) and 65% EtOH. Precipitation mixtures were incubated at room temperature for 16 minutes and then centrifuged at 12 500rpm (Boeco model U-320 centrifuge, Hettich 1460 rotor) for 10 minutes. DNA pellets were washed twice with 70% EtOH (12 500rpm centrifugation for 10min each). Pellets were vacuum dried for five minutes and submitted for sequence analysis on the ABI 3130 system (Applied Biosystems, CA, USA). To verify the presence of variants of interest sequence traces were visualised with the Applied Biosystems Sequence Scanner (version2). Sequence variants were described according to HGVS recommendations (Dunnen and Antonarakis 2000).

2.7. *In silico* prediction of pathogenic variants

The potential pathogenic effect of all missense variants identified with Sanger sequencing was assessed similar to the WES variants. We included predictions made by the Align-GVGD (Grantham variance-Grantham difference) algorithm that uses manually generated sequence alignments (Mathe, et al. 2006). Manual alignments were generated with T_Coffee (Di Tommaso, et al. 2011) and curated alignments were used by SIFT and PolyPhen-2. Concordance between the three methods have proven to give accurate predictions as A-GVGD compensates on specificity whereas SIFT and PolyPhen-2 have better sensitivity (Hicks, et al. 2011).

Intronic and synonymous variants were evaluated with the web-based application, human splice finder vs3.0 (HSFvs3.0) to investigate their effect on splicing signals. HSFvs3.0 incorporates up to three software approaches i.e. Human Splice Finder matrices, MaxEntScan & NNSplice (Desmet, et al. 2009; Reese, et al. 1997; Yeo and Burge 2004). Matrices from the ESEFinder and the RESCUE-ESE method form part of the web application to classify variants in candidate enhancer and silencing recognition sequences. Mutations located up to 2bp in the intron were considered as canonical splice site-disrupting variants. To assess the significance of intronic variants (>2bp from intron/exon boundaries) we made use of post hoc thresholds as determined by Whiley and colleagues (Whiley, et al. 2011).

Chapter 3:

Results of whole exome sequencing quality analysis and variant identification

3.1. Quality assessment of Illumina whole exome data

Whole exome sequencing on the HiSeq 2000 yielded 2–2.1 Giga base pairs of 2x90bp paired-end (PE) reads per index case. All quality assessment results have been provided in Table 3.1. FastX processing trimmed the size of mainly reverse reads to include base calls with PHRED scaled quality scores ≥Q30. The last five bases of reverse raw sequence reads (e.g. BRC_R) did not meet the defined standards (Figure 3.1). Sequences from three cases (cases from family 9 and 95) were trimmed to 84 base pairs (bp) and the remaining individuals' reads were only trimmed to 85bp. Post trimming the average sequence quality scores was Q38 for all nine individuals which indicated that the reads contained base calls that were close to 99.99% accurate. The total reads were reduced to 37 million on average with the index case from family 71 having the highest read count, closely followed by BRC 92-3. All the reads (including trimmed reads) were ~87bp in size. Final FastQC analysis results indicated that an average of ≥99% base call accuracy was achieved in forward and reverse paired-end sequence reads. All of the FastQC analysis modules that were incorporated verified that high base call accuracy and sequence quality was achieved for each run. These include measures of the per base sequence content, GC content and sequence duplication levels (Annexure 4). The base and GC content observed in data of index cases did not report any bias in the sequences (i.e. no adapters or cross species contamination).

Table 3.1: Whole exome sequencing coverage and variant call results

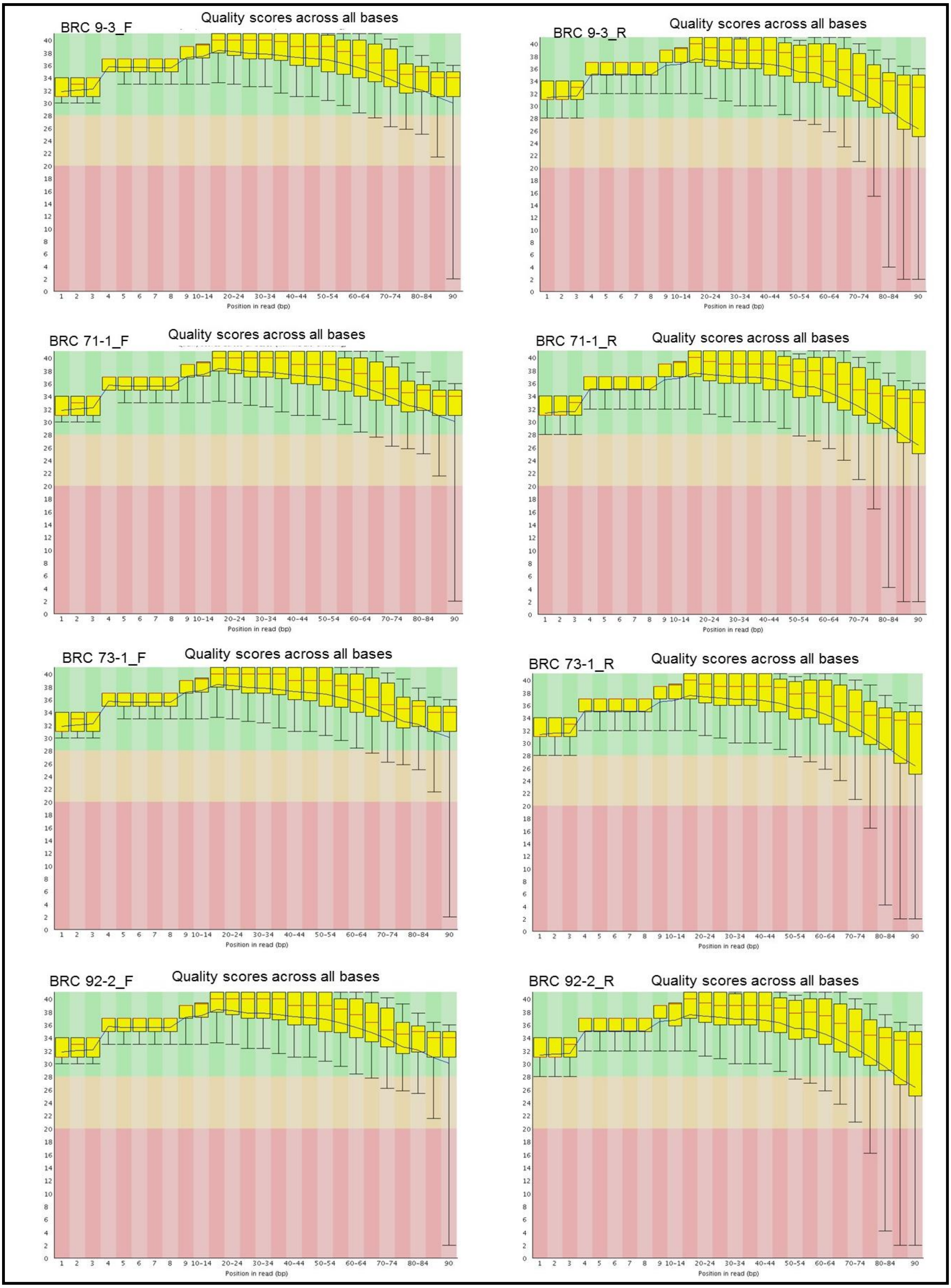
	Total reads	% Mapped reads[a]	Sequence duplication levels[b]	% Reads on target[c]	% Target bases 10 fold Coverage[c]	% Target bases 20 fold Coverage[c]	% Target bases 40 fold Coverage[c]	Mean coverage for target bases[c]	Ti/Tv ratio[d]	GC%[c]	Total variants (SNV: Indel)
BRC9-3	36,928,598	99.4	20.2	78.0	92.4	78.2	42.6	39.7	2.814	47.8	25 732:1280
BRC71-1	39,513,267	99.2	20.8	73.2	92.5	78.6	43.6	40.3	2.814	47.6	25 106:1270
BRC73-1	36,661,032	99.5	24.1	81.7	92.6	79.3	45.4	41.5	2.901	47.4	25 582:1291
BRC92-2	36,735,443	99.4	24.2	80.6	92.3	78.8	44.9	41.1	2.865	47.2	22 267:1269
BRC92-3	38,875,739	99.3	25.8	77.5	92.5	79.3	45.8	41.7	2.840	47.3	25 243:1280
BRC94-1	37,397,412	99.4	23.4	80.9	92.6	79.4	45.4	41.4	2.898	47.4	25 178:1262
BRC94-2	36,340,432	99.4	18.6	79.1	92.3	78.1	42.7	39.7	2.836	47.4	25 213:1241
BRC95-2	37,477,048	99.4	16.8	76.0	92.3	77.6	40.9	38.6	2.836	47.7	25 741:1240
BRC95-3	36,039,619	99.5	23.2	83.2	92.8	79.3	45.0	41.3	2.837	47.8	25 578:1263
Average	37,996,510	99.4	21.9	78.9	92.5	78.7	44.0	40.6	2.849	47.5	25 071:1266

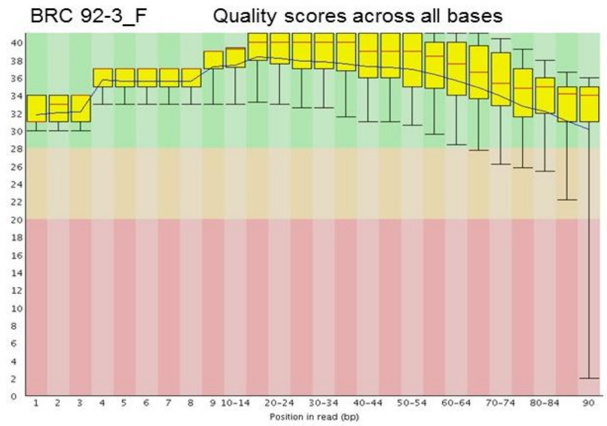
[a] Total mapped reads (assessment with the CLC Genomics workbench software v5) / total reads(Qualimap analysis) *100

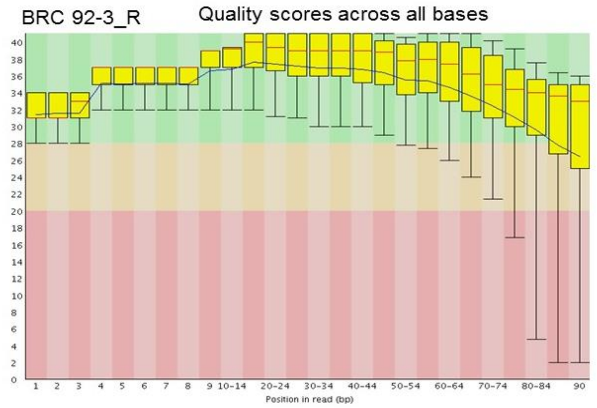
[b] Sequence duplication levels : (FastQC analysis Fwd + FastQC analysis Rev) / 2

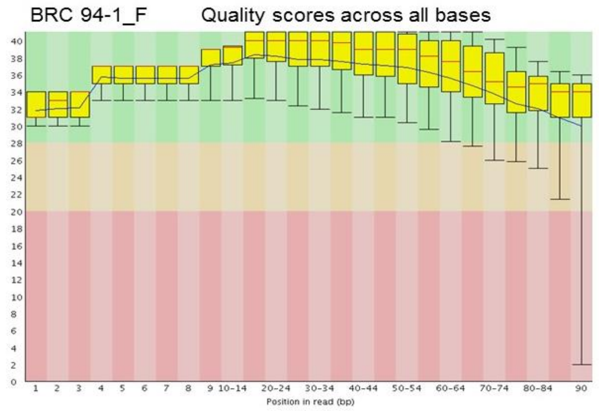
[c] CLC Genomics workbench v5 mapping report

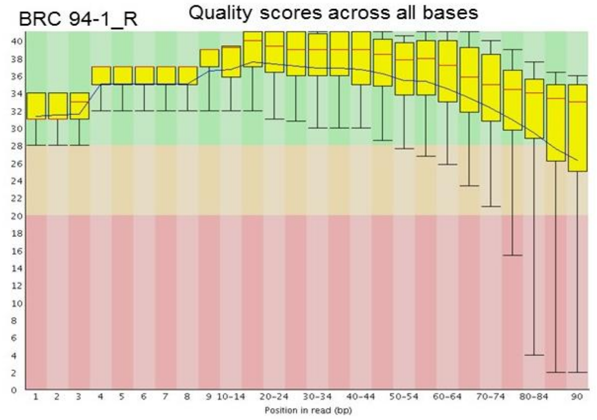
[d] According to VariantDB annotation

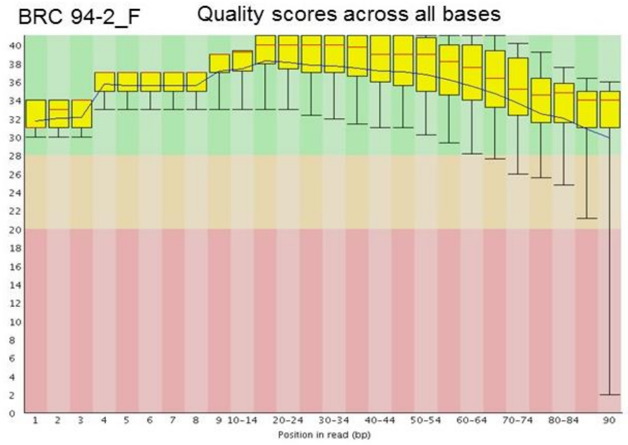

BRC 9-3_F
Quality scores across all bases
BRC 9-3_R
Quality scores across all bases
BRC 71-1_F
Quality scores across all bases
BRC 71-1_R
Quality scores across all bases
BRC 73-1_F
Quality scores across all bases
BRC 73-1_R
Quality scores across all bases
BRC 92-2_F
Quality scores across all bases
BRC 92-2_R
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)

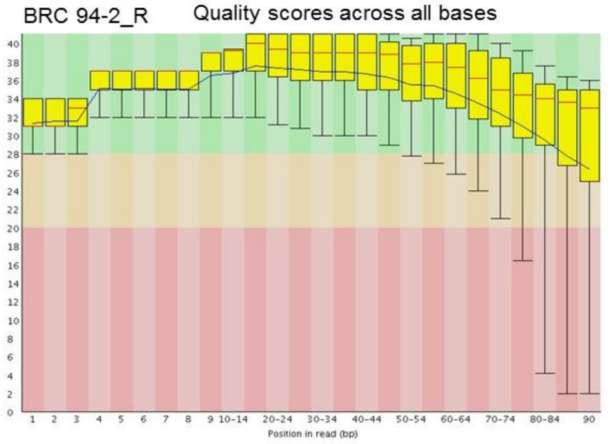

BRC 92-3_F
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)

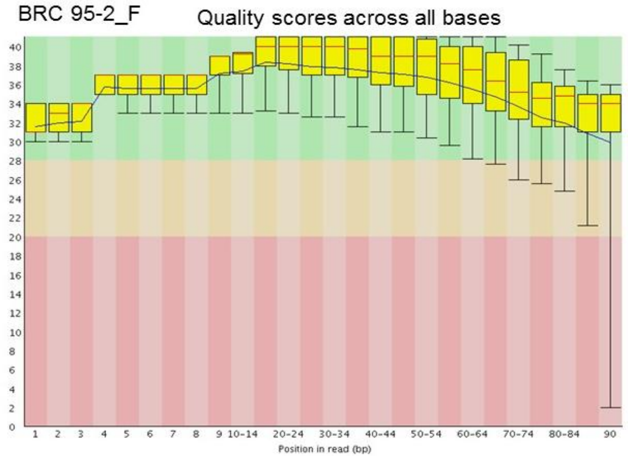

BRC 92-3_R
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)

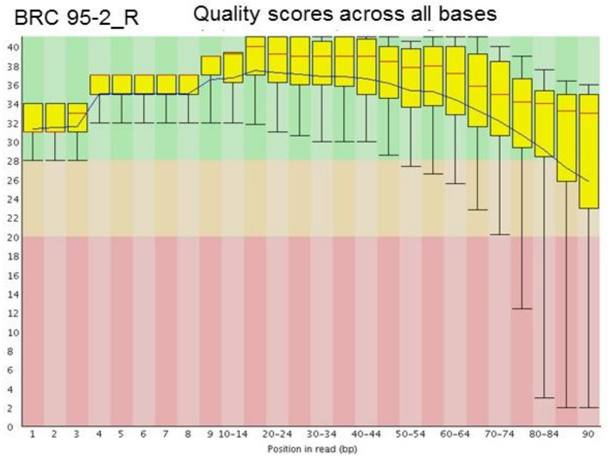

BRC 94-1_F
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)


BRC 94-1_R
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)


BRC 94-2_F
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)


BRC 94-2_R
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)


BRC 95-2_F
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)


BRC 95-2_R
Quality scores across all bases
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)

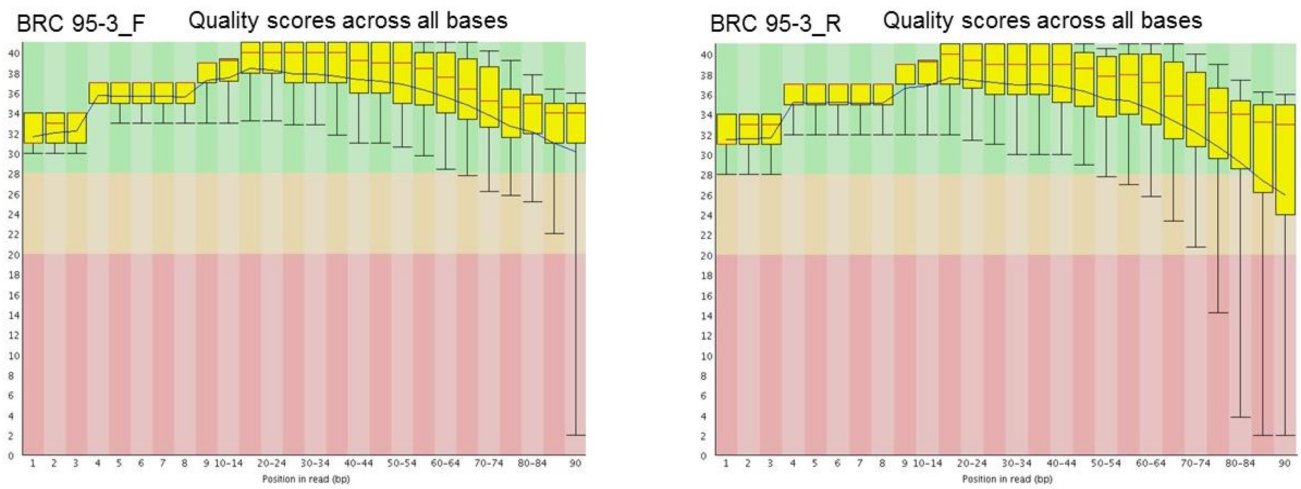


Figure 3.1: Representations of per_base_sequence quality.

Quality scores across the bases of forward (F) and reverse (R) paired reads according to Illumina 1.5 encoding was evaluated through FASTQC analysis (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Images of reverse reads before trimming have been provided.

3.2. Evaluating paired-end read mapping

Ninety-nine percent of the total reads for each individual mapped to the hg19 reference genome. Sequence coverage results for each individual is given in Table 3.1. Target intervals were covered by up to 83% of the sequence reads. Interestingly, BRC 71-1 sequences had the highest read count but less of these reads (73%) aligned to the enriched target regions. Individuals with a lower overall read count appeared to have more on-target coverage (incl. BRC 73-1, 92-2 and 95-3). The average target read depth for each dataset was 40X (Table 3.1). However, the mean depth of coverage for target bases in BRC 9-3, 94-2 and 95-2 was lower than this average. At a minimum read depth of 10X at least 92% of the target regions were covered and ~44% of these bases were covered with an average depth of 40X. Only 2% of the target bases were covered at a maximum depth of 100 fold.

The mean coverage appeared to follow a normal distribution with the target region base positions. Sequence reads mapped almost uniformly to target regions across all 24 chromosomes with the specificity deviating less than 20% between each chromosome (Figure 3.2).

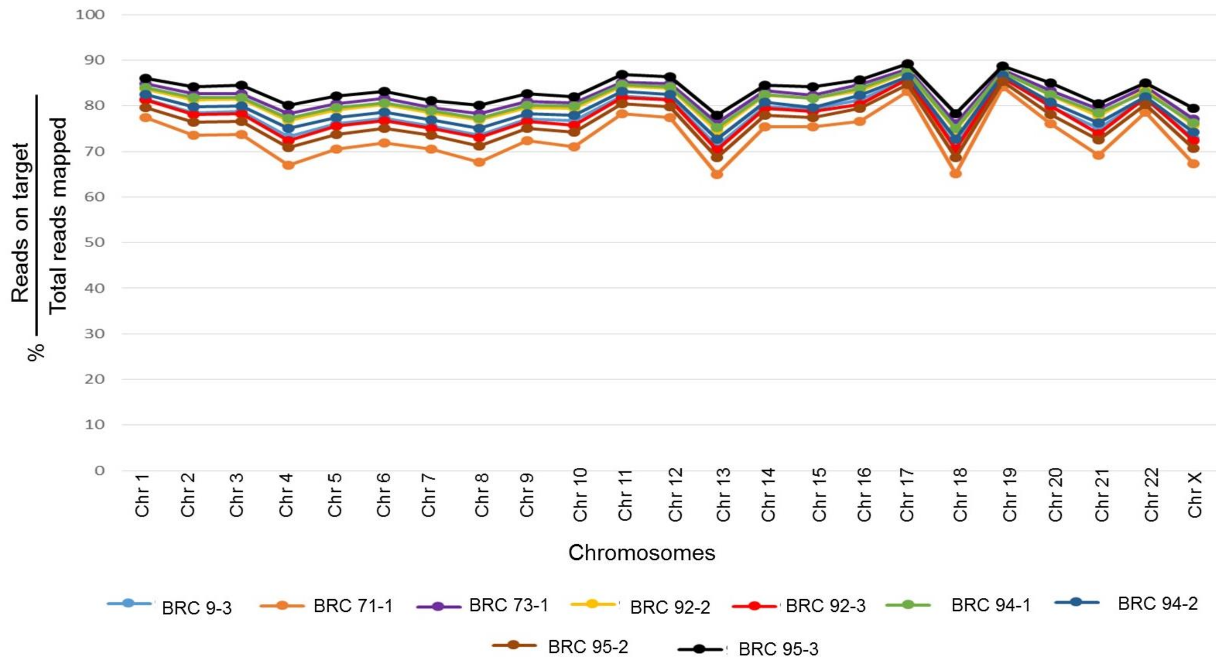


Figure 3.2: Percentage of reads that map to enriched target regions.

Exome capture sequencing yielded a total of 79% on-target reads for each index case.

The high quality sequence data that was generated for all index cases in turn produced high depth of coverage. Genetic variants were then identified from the germline sequence data of patients towards achieving the study aim.

3.3. Variant identification and filtration

The unified genotyper identified ~32 000 variants in each index case. Variant call files were submitted to the VariantDB web-based tool for annotation and comprehensive analysis. Raw variant calls were filtered based on allele count, depth and quality score thresholds which yielded an average of ~26 600 in each of the nine high-risk breast cancer patients. An average of ~26 000 variants (approx. 25 000 SNV and 1260 Indels) were detected in each case of which close to 18 800 were coding variants. No correlation was seen between the number of variants identified in each patient and their BC/OVC family history.

VariantDB parameters were set to select high quality reads that included; novel - and rare dbSNP v138 variants with allele frequencies equal to or below the threshold (MAF<0.01) for 1000 genomes, ExAC and ESP6500 projects variants. Variants of functional consequence were described as premature protein truncation (i.e. frameshift insertion-deletions and stop codon changes), as well as in-frame and canonical splice-site variants. The role of rare variants of functional significance were considered first priority.

3.3.1. Mutations in known breast/ovarian cancer genes

The variants of each index case was assessed for pathogenic germline mutations in well-known breast cancer predisposition genes such as *BRCA1, BRCA2, ATM, PALB2, PTEN, TP53, RAD51C and D.* One deleterious *BRCA1* nonsense mutation (c.5110C>T:p.R1704X) was detected in an individual (BRC 71-1) diagnosed with breast cancer at age 37 years. Previously these patients were screened for alterations in the *BRCA1/2* genes with the use of PTT and SSCP/heteroduplex analysis. Single base changes that evaded detection by heteroduplex analysis have become a more regular occurrence in whole exome sequencing studies (Snape, et al. 2012; Thompson, et al. 2012). These methods are known to have reduced sensitivity for detecting single base changes which could explain this stop-gain mutation evading detection (Gerhardus, et al. 2007). Sanger sequencing was used to verify the presence of this causal variant in the sister of BRC 71-1 (diagnosed with BC) as DNA was only available for this individual. The C>T change in codon 1704 will result in the loss of the BRCA1 protein's BRCT domain. This C-terminal region in BRCA1 regulates interactions with many other proteins that function in cell cycle checkpoint response to DNA damage e.g. BRIP (FANCJ) helicase, etc. (di Masi, et al. 2011; Nelson and Holt 2010). The nonsense variant has not been reported in the breast cancer information core database (www.research.nhgri.nih.gov/ projects/bic, (Szabo, et al. 2000)) but has been asserted as a clinically significant disease-causing variant associated with familial breast/ovarian cancer in ClinVar (Landrum, et al. 2014). The exome data of the remaining eight *BRCA1/2* negative individuals was then filtered for further analysis.

3.3.2. Variants identified in eight *BRCA1/2* negative index cases

To identify gene variants shared by patients within and between different families all the mutations of the remaining eight index cases were combined. On average approximately 18 000 coding variants were identified in each of the eight individuals. This was comprised of an average of 93 frameshift (FS), 190 in-frame indel (IF), 52 stop-gain (SG), 93 splice-region (SC), 8628 missense and 9725 synonymous mutations. Truncating and putatively truncating variants were analysed separately from missense coding variants. Silent and non-coding mutations were excluded from analysis. To identify putatively pathogenic missense mutations only rare, substitutions predicted as pathogenic by four *in silico* tools would be analysed. The total amino acid substitutions consisted of 121 unique novel missense variants and 343 that have been identified with minor allele frequencies below 1%. One hundred and twelve of the rare missense mutations have previously been detected at allele frequencies less than 0.5%. Only 10% of these were detected within or between two or more families. Nonsynonymous mutations with MAF as low as 0.2% (64) were also identified among the index cases but none of the missense mutations (MAF ≤ 1%) were predicted as pathogenic or damaging by any of the four *in silico* tools.

Variant prioritisation was then performed with frameshift and in-frame indels, nonsense variants and canonical splice-site mutations. The approach followed in this study was comprised of two levels to systematically select variants and generate a list of potential candidate genes for further validation. The first objective was to choose genes that contain truncating and putatively truncating (in-frame indels and splice site) variants in two or more of the families. Thereafter, family-specific variants were evaluated. The first level of variant prioritisation was based on the genetic architecture of high penetrant hereditary breast cancer. Variant alleles that confer a high-risk for this disease are significantly enriched for in affected cases (MacArthur, et al. 2014; Reva, et al. 2011). There were no rare gene variants that were significantly damaging and segregated in all eight individuals between the five *BRCAx* families. Table 3.2 shows all potentially deleterious variants (i.e. frameshift, in-frame insertion/deletions, stop gain & canonical splice site) with MAF less than 1% that are shared amongst the high-risk BC/OVC patients. The 41 deleterious variants (Table 3.2) that were discovered were shared within one or between two separate families (maximum four). It is due to similar observations that an alternative hypothesis exists which suggests that the missing heritability seen in many *BRCA1/2* negative patients may be specific to each family with a history for breast cancer (Wen, et al. 2014). The 39 genes listed are associated with well-established breast cancer-associated DNA repair mechanisms or code for proteins mainly involved in processes such as; cell cycle regulation, cell division, centrosome organisation, regulation of transcription, programmed cell death, immune response processes, cell adhesion and G-protein coupled protein signalling. Mutations that affect the last mentioned processes have previously been implicated to play a role in carcinogenesis by providing selective growth advantage (Economopoulou, et al. 2013; Lim and Kaldis 2013; O'Hayre, et al. 2014; Rotunno, et al. 2014; Signore, et al. 2013; Wang, et al. 2013).

All variants that were shared between BC/OVC families were cross-referenced to resources such as the somatic cancer databases (COSMIC, cBioportal & ICGC) and the ClinVar repository. None of the mutations were reported as clinically significant in the ClinVar database. Interestingly, of all the variants only *ASPN:* p.51_52del and *CELA1*: c.7_8del have previously been identified in breast/ovarian cancer tumours (Table 3.2).

Table 3.2: Truncating (frameshift indels & nonsense), in-frame and splice-site mutations between and within *BRCAx* families

Variant position[a]	Gene symbol[b]	Gene-Id[c]	Variant[d]	Predicted effect	#Families n=5 [e]	dbSNP [f]	1000g EUR [g]	ESP6500 EUR [g]	ExAC EUR [g]	Gene Description [h]
1_27699670	FCN3	NM_173452	c.316delC	p.L106fs	1	rs28357092	0.03	0.01780	0.0164	Human autoantigen
1_40773150	COL9A2	NM_001852	c.976C>T	p.Q326X	1	rs12077871	0.00	0.00105	0.0033	Human cartilage collagen fibril
1_41486083	*SLFNL1*	NM_144990	c.250G>T	p.E84X	1	rs148992629	0.012	0.0045	0.0080	Nucleotide binding protein
1_54605318	*CDCP2*	NM_201546	c.1224dupC	p.M409fs	3	rs200136238	-	-	-	Involved in cholesterol biosynthesis and electron transport
1_86820290	ODF2L	NM_001184766	c.1443dupT	p.K482_F483delinsX	1		0.00	0.00010	0.0006	
1_145415397	*HFE2*	NM_213653	c.217delG	p.G73fs	1		.	.	.	Involved in iron metabolism
2_48033792	*MSH6*	NM_000179	c.3611+2_3611 +5delTAAC		1	rs367548604	0.002	0.0013	0.0012	Contributes towards mismatch DNA-repair
2_120078770	*C2orf76*	NM_001017927	c.134_143del	p.D45fs	2		0.01	.	0.0068	
4_2074703	*POLN*	NM_181808	c.2509delC	p.Q837fs	1	rs3833632	0.002	0.0034	0.0027	Crosslink repair and homologous recombination
4_84368022	*HELQ*	NM_133636	c.1358delG	p.R453fs	1		-	-	-	Participates in DNA crosslink repair
4_113353522	*ALPK1*	NM_001253884	c.2586_2588del	p.862_863del	2	rs150577225	0.01	0.00850	0.0077	
5_156479568	*HAVCR1*	NM_012206	c.462_476del	p.154_159del	3		-	-	-	Regulate immune cell activity as host response to viral infection
5_156479568		NM_012206	c.599_601del	p.200_201del	3		-	-	-	
7_1586653	*TMEM184A*	NM_001097620	c.1176_1177ins GGC	p.S393delinsGS	3	rs112463195	-	0.0002	0.0001	Vesicle transport in exocrine and Sertoli cells
7_140064266	SLC37A3	NM_001287498	c.314_316del	p.105_106del	1		.	.	.	
7_92923950	*CCDC132*	NM_017667	c.1167+2T>G		2	rs75893203	-	-	0.0073	
8_16012594	*MSR1*	NM_002445	c.877C>T	p.R293X	1	rs41341748	0.012	0.0098	0.0100	Plays a role in cell apoptosis. possible genetic marker for Prostate cancer

Variant position[a]	Gene symbol[b]	Gene-Id[c]	Variant[d]	Predicted effect	#Families n=5 [e]	dbSNP [f]	1000g EUR [g]	ESP6500 EUR [g]	ExAC EUR [g]	Gene Description [h]
9_95237024	*ASPN*	NM_001193335	c.153_155del	p.51_52del	4	rs111419727*	-	-	-	Encodes a cartilage extracellular protein that is member of the small leucine-rich proteoglycan family
9_95237024			c.155_156insTGA	p.E52delinsDE	1		-	-	-	
9_107556793	*ABCA1*	NM_005502	c.5383-2->TT		2	rs3029409	-	-	-	Member of the superfamily of ATP-binding cassette (ABC) transporters
10_35625846	*SYT15*	NM_181519	c.652-2->C		2		-	-	-	A member of the Synaptotagmin (Syt) family of membrane trafficking proteins
11_21729952	*SKA3*	NM_145061	c.1120-2->TT		4	rs11446085	0.002	-	-	Regulates microtubule attachment to kinetochores during mitosis
11_59480952	OR10V1	NM_001005324	c.367C>T	p.Q123X	2	rs499037	0.01	0.012	0.01	Interact with odorant molecules to initiate a neuronal response for smell perception
11_65350815	EHBP1L1	NM_001099409	c.2673dupT	p.T891fs	1		.	.	.	
11_66455734	SPTBN2	NM_006946	c.6277_6279del	p.2093_2093del	1		.	.	0	
12_29936562	*TMTC1*	NM_001193451	c.123C>A	p.Y41X	2		-	-	-	
12_51740414	*CELA1*	NM_001971	c.7_8del	p.V3fs	2	rs370927847*	0.32	.	0	Serine proteases that hydrolyses elastin
12_110353234	*TCHP*	NM_001143852	c.1347G>A	p.W449X	1	rs143201598	-	0.0012	0.001	Putative tumour suppressor with negative regulation of cell growth Pro-apoptotic during cell stress
13_24242915	TNFRSF19	NM_001204459	c.528G>A	p.W176X	1		.	.	.	TNF-receptor superfamily member
15_68582570	*FEM1B*	NM_015322	c.874_875insA	p.E292fs	1		-	-	-	Induction & regulation of apoptosis Regulation of DNA damage checkpoint
15_70961675	*UACA*	NM_001008224	c.1309C>T	p.R437X	1		-	-	-	Endogenous regulators of the apoptosome
16_1825982	*EME2*	NM_001257370	c.964C>T	p.Q322X	1	rs61753375	0.016	0.015	0.01	Plays a role in DNA recombination & repair

Variant position[a]	Gene symbol[b]	Gene-Id[c]	Variant[d]	Predicted effect	#Families n=5 [e]	dbSNP [f]	1000g EUR [g]	ESP6500 EUR [g]	ExAC EUR [g]	Gene Description [h]
16_48261815	*ABCC11*	NM_033151	c.297G>A	p.W99X	1	rs145048685	-	0.001	0.001	Located in cell membrane
16_74709590	MLKL	NM_152649	c.1107_1110del	p.D369fs	1		0.00	0.057	0.004	Contributed to tumor necrosis factor (TNF)-induced necroptosis, a programmed cell death process
16_88496020	ZNF469	NM_001127464	c.2143_2145del	p.715_715del	1		.	.	.	
17_67039819	ABCA9	NM_080283	c.611C>A	p.S204X	1	rs145251776	0.00	0.0005	0.001	Transport molecules across extra- and intracellular membranes
18_32558479	*MAPRE2*	NM_001143827	c:-2_2delGAAT		1		-	-	-	Microtubule binding. Cell cycle division & proliferation
18_56205373	*ALPK2*	NM_052947	c.2045dupC	p.P682fs	2		-	-	-	Kinase that recognizes phosphorylation sites
20_44520237	*CTSA*	NM_000308	c.84_85insCTG	p.F28delinsFL	2	rs10582052	0.003	-	0.001	Protective protein that stabilizes beta-galactosidase and neuraminidase activity
22_29885598	*NEFH*	NM_021076	c.1970_1975del	p.657_659del	2	rs149571560	-	-	0.001	Comprise the axoskeleton and functionally maintain neuronal caliber
X_35937926	CXorf22	NM_152632	c.10C>T	p.Q4X	1	rs141633156	-	0.0001	0	

[a] Genomic position (chromosome_nucleotide site) of variant relative to the human reference genome GRCh37, i.e. hg19

[b] Official HGNC gene symbols

[c] RefSeq Gene identification number

[d] Variants were written according to the HGVS requirements in reference to Genbank reference sequences. Nucleotide and amino acid/codon changes have been indicated

[e] The number of families carrying the variant of interest

[f] Accession numbers of variants reported in the NCBI dbSNP database, build 38, have been provided (http://www.ncbi.nlm.nih.gov/projects/SNP/)

[g] 1000 Genomes browser, release date 16.10.2014 (http://browser.1000genomes.org). NHLBI Exome sequencing project (ESP) 6500, release version v.0.0.30, date 03.01.2014 (http://evs.gs.washington.edu/EVS/). The Exome Aggregation Consortium (ExAC), release version v.0.3.1 (http://exac.broadinstitute.org/). Minor allele frequencies (MAF) have been provided

[h] Gene ontology as described in the literature (where available) * Variants reported in cBioportal, COSMIC and ICGC databases in breast/ovarian cancers

Breast cancer susceptibility in the study group could also be explained by family-specific variants, which has previously been suggested by many research studies (Merdad, et al. 2015; Thompson, et al. 2012; Wen, et al. 2014). This became the final step in our prioritisation pipeline.

3.3.3. Family-specific variants

Mutations that had fulfilled the variant filtration criteria and were identified in either one or two index cases within one specific *BRCAx* family were selected. This included truncating (nonsense and frameshift) and potential splice site disrupting variants. It is known that not all premature truncating variants are pathogenic (Robinson, et al. 2014). For example; mutations that are located in the 3' gene terminus should be interpreted with care as they do not necessarily affect protein function (Borg, et al. 2010). Genes that play a role in DNA repair and cell cycle regulation processes were selected and putative pathogenic variants located in important protein domains were prioritised as these may be more damaging. Nine genes of interest with variants unique to four of the families were further prioritised, these are: *POLN, MSR1, TCHP, FEM1B, UACA, EME2, ABCC11, HELQ* and *MAPRE2.* Novel mutations were found in *MAPRE2, HELQ, FEM1B* and *UACA* genes. *POLN:*c.2509delC, *TCHP*:c.1347G>A and *ABCC11*:c.297G>A were three family-specific variants that have previously been reported in the dbSNP with allele frequencies ranging from 0.001 - 0.003 in the 1000 Genomes public database.

DNA polymerase ν (POLN) is an enzyme that has recently been under investigation for its contribution towards DNA template synthesis during the repair of crosslinked DNA lesions. Studies have shown that POLN is involved in homologous recombination and interacts with proteins of the fanconi anemia pathway (FANCD2-I and RAD51) as well (Marini, et al. 2003; Moldovan, et al. 2010). While mutations in the polymerase protein family have been associated with increased predisposition to various cancers (Lange, et al. 2011) it is known that POLN incorporates nucleotides with low fidelity during DNA template synthesis (Takata, et al. 2015). More evidence is needed to support the role that POLN plays in DNA repair in order to motivate further investigation. The macrophage scavenger receptor A protein has also been investigated for the role it plays in familial cancer syndromes but studies have indicated that germline mutations in this gene are not associated with increased cancer risk (Blute, et al. 2003; Neyen, et al. 2013; Sun, et al. 2006).

Rare *ABCC11* gene variants have been identified in studies that have investigated the genetic component of breast/ovarian cancer through whole exome sequencing (Snape, et al. 2012). However, the cellular function of this transmembrane protein (Ishikawa, et al. 2012) did not motivate further analysis of the variant. Lastly, *FEM1B* and *UACA* both code for proapoptotic proteins and may be investigated in future. Very little information is available to prove their role in inducing apoptosis (Moravcikova, et al. 2012; Subauste, et al. 2010). Of the nine gene variants, only four were considered good potential candidates for further investigation.

MAPRE2 expression produces the microtubule-associated, RP/EB family, member 2 protein (EB2) that regulates microtubule reorganisation during epithelial cell differentiation (Goldspink, et al. 2013). Microtubules are structural components of the cytoskeleton involved in cell division, cytoplasmic organisation, cell migration and chromosome segregation (Abiatari, et al. 2009). EB2 is one of three family proteins that influences microtubule dynamics and stability (Arens, et al. 2013). EB1 and EB3 promote cell growth and EB2 mainly maintains dynamic microtubule formation (Goldspink, et al. 2013). *TCHP* encodes the trichoplein, keratin filament binding protein (Cerqua, et al. 2010). TCHP has recently been studied as a novel putative tumour suppressor protein primarily because of the role it plays in inhibiting uncontrolled cell growth (Kim, et al. 2010). The potential protein truncation variant (c.1347G>A) may affect the structural integrity of the expressed protein.

EME2, codes for the essential meiotic structure-specific endonuclease 2 subunit which forms a heterodimeric complex with MUS81 and functions as a structure selective endonuclease (Pepe and West 2013). The complex is responsible for resolving stalled replication forks by cleaving 3’ flap/fork nicked dsDNA towards maintaining genomic integrity (Amangyeld, et al. 2014). *HELQ* codes for a DNA helicase enzyme that facilitates in DNA crosslink repair and cellular recovery from replication stress (Guler, et al. 2012; Tafel, et al. 2011). HELQ localises to damaged replication forks and unwinds DNA helices in order to facilitate DNA repair and replication restart (Tafel, et al. 2011). This protein associates with RAD51 paralogues (RAD51B/C/D and XRCC2) and participates in interstrand cross-linked (ICL) DNA repair (Adelman, et al. 2013).

3.4. Sanger sequence confirmation

High-priority gene mutations were highlighted for Sanger sequence variant verification and included; *MAPRE2* (c:-2_2delGAAT)*, TCHP* (c.1347G>A)*, HELQ* (c.1358delG) and *EME2* (c.964C>T). Index cases and their affected relatives (where DNA was available) were screened for each specific variant. Up to nine control samples from unaffected individuals were screened as well. Primer pairs and optimized PCR conditions are indicated in Table 3.3. Amplified fragments were sequenced bi-directionally at the same conditions as PCR reactions.

Variants specific to each family were then evaluated. The *MAPRE2* variant (c:-2_2delGAAT) segregated within index cases from family 95 and affected relatives and was not detected in any of the controls. This protein can be expressed as one of four alternate isoforms differing with regards to their 5'UTR or alternate start codon positions (www.ncbi.nlm.nih.gov/gene/10982 (2015)). Due to the multiple variant isoforms (i.e. four), the biological effect of one absent transcript may not be significant enough to predispose for disease, therefore this gene was not included for further analysis.

The *TCHP* c.1347G>A mutation was verified in one of the individuals that had undergone exome sequencing from family 92 (III-6) and an additional first degree relative (III-5). This may suggest the variant may be segregating with the disease in this family. The fact that *TCHP*:p.W449X was absent in their cousin (III-4) places some doubt on the last mentioned theory. Further screening was needed to investigate the involvement of this gene in breast cancer predisposition.

The *HELQ* c.1358delG variant was discovered in one of the two index cases from family 95 (III-2) and was absent in the other two family members that were analysed. It may be possible that this *HELQ*:p.E522DfsX4 positive individual may have a mutation of paternal origin but information of the BC family history was only available for the maternal family line. The clinical significance of this variant has not been described in any clinical repository and no population information was available at the time of the current study.

Table 3.3: Primers and amplification conditions used for variant verification with Sanger Sequencing

Gene	**Variant location (cDNA)**	**Exon/Site**	**Primer Name**	**Primers**	**PCR: Tann**	**PCR $[MgCl_2]$**
EME2	c.964C>T	7	EME2_7F	AAGGCTTCTCTCTGTCCCCA	60ºC	1,0mM
			EME2_7R	GAGCAGGCAAAGGCATGAGA		
HELQ ***(HEL308)***	c.1358delG	4	HELQ_4BF	GTTTCTTTGTTGAAGAATATGCTGG	53ºC	1,5mn
			HELQ_4BR	ACTAAGTAGATCCACACAATAACCA		
MAPRE2 ***(EB2)***	c:-2_2delGAAT	5'UTR/Ex1	MAPRE2_1F	GTCAGACGCAGCACCTACTT	60ºC	1,5mM
			MAPRE2_1R	AAGCATAGCTCTACCCACGC		
TCHP ***(MITOSTATIN)***	c.1347G>A	12	TCHP_12F	CGCAGGCTTACCTTCCTCAT	58ºC	1,0mM
			TCHP_12R	GAGTCTGTCTTGCCGGTGTT		

Lastly, the single nucleotide variant identified in *EME2* (c.964C>T) was verified. This will introduce a premature termination codon in the protein. Based on EME2's involvement in DNA repair it may be tempting to suggest that such a rare variant allele may account for additional unexplained high-risk breast/ovarian cancer cases. *EME2* c.964C>T was identified in the one index case selected for WES from family 9. Sequence analysis was performed and it was verified that BRC 9-3 and four of her five first degree relatives available carried this putatively damaging mutation. The individual that was negative for the variant (IV-1, her daughter) had not been diagnosed with breast cancer at the time of sample collection.

3.5. Final list of candidate genes

In-depth literature searches were used to prioritise rare protein-truncating variants. Sufficient literature and experimental evidence was available for *TCHP, MAPRE2, HELQ* and *EME2*. After Sanger sequencing only three of the probable genes of interest were included for further investigation. *TCHP, EME2*, and *HELQ* code for proteins that are primarily involved in processes such as DNA double strand break repair and apoptosis (Cerqua, et al. 2010; Pepe and West 2014; Takata, et al. 2013). Several studies have shown that these particular processes could play a causative role in development cancer development (Vecchione, et al. 2008; Wang, et al. 2013; Yao and Dai 2014).

The present study supports the hypothesis that the unknown hereditary predisposition seen amongst *BRCAx* families may be explained by exploring the role of rare gene variants specific to each family affected by the disease (Wen, et al. 2014). Potentially damaging mutations in *TCHP*, *EME2*, and *HELQ* were discovered in three individuals from separate high-risk BC/OVC families. The genes were then screened in other hereditary breast/ovarian cancer cases to explore their possible roles in breast cancer susceptibility. These results are detailed in the following three chapters.

Chapter 4:

TCHP sequence variants in South African breast and ovarian cancer families

4.1. Introduction

Breast cancer susceptibility genes code for proteins that play multifunctional roles in various processes aimed at maintaining cellular integrity. These include DNA damage repair response, cell cycle checkpoint control and apoptosis (Emmert-Streib, et al. 2014; Roy, et al. 2012; Wolters and Schumacher 2013). Cancer development can largely be explained by the deregulation of such molecular processes that allow cells to acquire capabilities such as irregular cell growth. These functional capabilities were first described by researchers Hanahan and Weinberg who proposed a total of ten hallmark biological events that when acquired allow for the survival, proliferation and dissemination of cancer cells (Hanahan and Weinberg 2011). Tumorigenesis is a multistep processes and is enabled by the most important factor i.e. the development of genomic instability (Kwei, et al. 2010). A combination of rare genetic changes and random genomic rearrangements cause genomic instability and orchestrate hallmark capabilities such as increased cell proliferation and the ability to escape programmed cell death (Negrini, et al. 2010). This significantly affects both cellular and tissue homeostasis and is the fastest path towards tumorigenesis (Spencer, et al. 2004).

Fully functional cell death processes have been shown to reduce the fitness of genetically mutated cells regardless of the genomic alteration present (Enderling and Hahnfeldt 2011). Therefore, maintaining the balance of cell survival and death is equally important for cancer prevention. This process is acted on by multiple complex signalling networks in response to various cellular stresses and safeguards genome integrity (Plati, et al. 2011). Apoptosis acts as a barrier against tumorigenesis and is regulated by multiple signalling mechanisms (Delbridge, et al. 2012). The disruption of this multi-faceted machinery may be the initial event associated with tumour development and progression. Studies have established an inversely proportional association between the rate of cell death and tumour size. Due to this observation, features of the apoptosis pathway are mostly targeted for therapeutic purposes to inhibit tumour expansion (Enderling and Hahnfeldt 2011). Therefore, elements that assist in shifting the disrupted cellular microenvironment to programmed cell death, through oncogene signalling, are worth investigating for their ability to predispose to cancer.

The central components of the programmed cell death pathway are comprised of the caspase protein family members that have been characterised as the core apoptosis machinery (Brentnall, et al. 2013). Recent advances have been made towards the discovery of additional elements that aid in the induction of apoptosis such as the ubiquitously expressed trichoplein keratin filament-binding protein, TCHP (aka mitostatin) (Nishizawa, et al. 2005). Trichoplein is a keratin intermediate filament (IF) protein that co-associates with proteins K8/18 and K16

(Nishizawa, et al. 2005). Intermediate filaments are major contributors of the cytoskeleton in eukaryotic epithelial cells and are of structural importance (Marceau, et al. 2014). TCHP has additional functional roles independent to the supportive role it plays towards stabilising the cellular architecture with keratin IF's. Mitostatin is generally localised with the mitochondria of cells where it alters the morphology of this organelle, promoting apoptosis (Cerqua, et al. 2010; Ibi, et al. 2011; Neill, et al. 2014). Overexpression of this protein has proven to be pro-apoptotic which suggests a possible tumour suppressive role for TCHP. The expression of TCHP is upregulated in normal endothelial tissues and smooth muscles. Reduced levels of expression has been confirmed through immunohistochemistry in various cancer derived cells and tumours from bladder, breast and prostate cancers (Fassan, et al. 2011; Vecchione, et al. 2008). Mutational analysis of this protein has also revealed that mitostatin may be inactivated through various mechanisms that all result in a loss of cellular homeostasis (Kim, et al. 2010). These studies, however suggestive, have not fully investigated whether loss of function mutations in TCHP could be a feature of human cancers. As yet no studies have investigated its capacity as a putative tumour suppressor that may be associated with breast cancer.

The work conducted here reports on the potential involvement of *TCHP* variants in hereditary BC/OVC in South Africa. Patients from 64 families with strong background for breast and ovarian cancer and no *BRCA1/2, PALB2* or *RAD51C* mutations were combined and screened for variants in the TCHP coding region.

4.2. Materials and Methods

4.2.1. Breast/ovarian cancer subjects

As described in Chapter 2, high-risk breast/ovarian cancer families (56) were included in this study. Genomic DNA of each individual was screened for mutations in the coding region of TCHP.

4.2.2. *TCHP* mutation screening

The *TCHP* coding sequence spans from exon2, therefore, exon1 (non-coding exon) was not included during screening. Table 4.1 lists the primers designed in our study and include optimised conditions for sequence analysis. As indicated, 11 fragments were amplified with exons seven and eight analysed as one fragment. Cycle sequencing was performed under the same conditions as PCR using the procedures described in Chapter 2. All premature truncating variants were verified through bi-directional sequencing.

Table 4.1: Primers and amplification conditions for *TCHP*

Exon	Primer sequence	Length	F/R	Ta	PCR Tann(°C)	PCR [$MgCl_2$]	Fragment Size
2	GTTAAAGGGATGAGGCCAAG	20	F	55.4	58	1.5mM	355bp
	AATCCTACGCTCCCTAAATC [a]	20	R	53.35			
3	CTACCTCAGCCTCTTTTACC	20	F	55.4	56	1mM	497bp
	AAGGGGCTAAAGAGACTACA [a]	20	R	53.35			
4	CATGAGGGTTTGAACAGTGG [a]	20	F	55.4	57	1mM	326bp
	CTCTTTTCAGGTGGCTCAAG	20	R	55.4			
5	CCCCATGAAGCGAAGTCAT [a]	19	F	56.47	58	1mM	238bp
	CAGTCTCGCACAGCAGAATA	20	R	55.4			
6	TTATTGAGTCCCTCCCTGAC [a]	20	F	55.4	57	1.5mM	428bp
	TGCTATGGCCTCTTCTTAGG	20	R	55.4			
7&8 *	GTAGTGTCTGCGAAAAATGG [a]	20	F	53.35	57	1.5mM	481bp
	GCTAAGAAAGATGTCAGGCT	20	R	53.35			
9	TAGAAAGGAAAGGGAGGGCA [a]	20	F	55.4	56	1mM	331bp
	TGGGGCATTTCAAGGTATCA	20	R	53.35			
10	TGGATGCTGTCCTTAATTGT [a]	20	F	51.3	54	1mM	364bp
	CCCACTTAAAATGACCAGAGA	21	R	52.5			
11	GAGAGCAAATCCTGTCATCT [a]	20	F	53.35	55	1mM	417bp
	CTGAACCCTGTCCCATGAT	19	R	56.47			
12	CGCAGGCTTACCTTCCTCAT [a]	20	F	56.9	59	1mM	272bp
	GAGTCTGTCTTGCCGGTGTT	20	R	56.9			
13	GCTGGAAATGAAACAGCCA [a]	19	F	54.32	59	1mM	229bp
	GAGCCCTGTAAAACTGCCT	19	R	56.47			

Primers were designed with Primer 3-blast using the NCBI37/Hg19 genomic reference sequence NC_000012.12
PCR primers and optimised conditions for each exon have been indicated
* Note: exon 7 and 8 were amplified in one fragment
[a] Primers used during uni-directional sequencing

DNA sequences from the NCBI reference sequence (RefSeq) database i.e. genomic (NC_000012.12) were used for primer design. mRNA (NM_001143852), cDNA and protein (NP_001137324) sequences were utilized for variant description. The functional effect of all missense variants identified with Sanger sequencing, was evaluated with *in silico* classifier algorithms SIFT (sorting intolerant from tolerant), Polyphen-2 (polymorphism phenotyping) and Align-GVGD (Grantham Variance-Grantham difference). Missense substitutions with scores of SIFT (<0.05), Polyphen2 (>0.85) and A-GVGD (C65) (Adzhubei, et al. 2010; Mathe, et al. 2006; Ng and Henikoff 2003) predicting a likely deleterious effect were considered high priority mutations. A-GVGD manual alignments were generated with protein sequences of ten species, collected from the NCBI protein database. These included; *Homo sapiens* (human), *Mus musculus* (mouse, NP_001137324.1), *Canis lupus familiaris* (dog, XP_534716.3), *Danio rerio* (zebrafish, NP_001035433.1), *Gallus gallus* (chicken, XP_422992.3), *Loxodonta africana* (African elephant, XP_003419404.1), *Xenopus laevis* (African clawed frog, NP_001082520.1), *Rattus norvegicus* (rat, NP_001178595.1), *Pan troglodytes* (chimpanzee, XP_509357.2) and *Bos Taurus* (cattle, NP_001157498.1). The free web-based application, human splice finder vs3.0 (HSFvs3.0) was used to evaluate synonymous and intron mutations.

4.3. Results and Discussion

4.3.1. *TCHP* sequence variants

A total of 13 variants were identified through *TCHP* gene screening (Table 4.2). This included a c.1347G>A mutation previously reported in dbSNP that generates a premature stop codon (p.W449X) and a novel in-frame deletion of one amino acid (c.1382_1384delAGG; p.E461del), both were found in exon12 (Figure 4.1). Eleven additional variants were detected across the TCHP coding sequence. This included; three novel variants (one missense and two intron) and eight previously reported missense (two), synonymous (one) and intronic (five) mutations.

The c.1347G>A nonsense variant was only identified in 1.6% of the 61 BC/OVC families screened within this study. The c.1382_1384delAGG in-frame deletion variant was also detected in 1.6% of BC/OVC families screened. The majority of *TCHP* mutations revealed through Sanger sequencing were intronic. However, the total missense, synonymous and intronic variants identified were found in ~39% (27/69) of all patients.

Table 4.2: Sequence variants identified in the *TCHP* gene

	Site [a]	Nucleotide altered [b]	Amino acid altered [b]	Cases		dbSNP138 [d]	1000Genomes [e]		ESP6500 [f]	
				Number (carrier frequency) [c]	Variant allele frequencies		Global	EUR	ALL	EA
Truncating	Ex12	c.1347G>A	p.Trp449X	1 (0.01)	0.007	rs143201598	-	-	0.00076	0.0011
In-frame deletion	Ex12	c.1382_1384delAGG	p.Glu461del	1 (0.01)	0.007	-	-	-	-	-
Missense	Ex3	c.380A>G	p.Lys127Arg	20 (0.29)	0.144	rs10774978	0.31	0.18	0.3054	0.1582
	Ex6	c.535A>C	p.Thr179Pro	1 (0.01)	0.007	-	-	-	-	-
	Ex11	c.1249G>A	p.Glu417Lys	4 (0.06)	0.03	rs16940680	0.14	0.03	0.1197	0.028
Synonymous	Ex2	c.21G>A	p.Pro7=	13 (0.19)	0.09	rs11539159	0.10	0.1	0.1297	0.0936
Intronic	Ex4	c.456+29T>C	-	10 (0.14)	0.07	rs57860034	0.23	0.07	0.2019	0.0703
	Ex4	c.456+37G>A	-	1 (0.01)	0.007	-	-	-	-	-
	Ex5	c.457-38T>C	-	11 (0.16)	0.08	rs2271318	0.23	0.07	0.2019	0.0703
	Ex6	c.526-26_526-25insCCT	-	11 (0.16)	0.08	rs3830479	0.23	0.07	0.198	0.0695
	Ex7	c.700-34A>G	-	2 (0.03)	0.014	rs147660819	0.007	0.01	0.0047	0.0067
	Ex10	c.1134+74C>T	-	1 (0.01)	0.007	-	-	-	-	-
	Ex13	c.1465-20T>C	-	1 (0.01)	0.007	rs12372049	-	-	0.0035	0.0047

[a] Ex (exon)

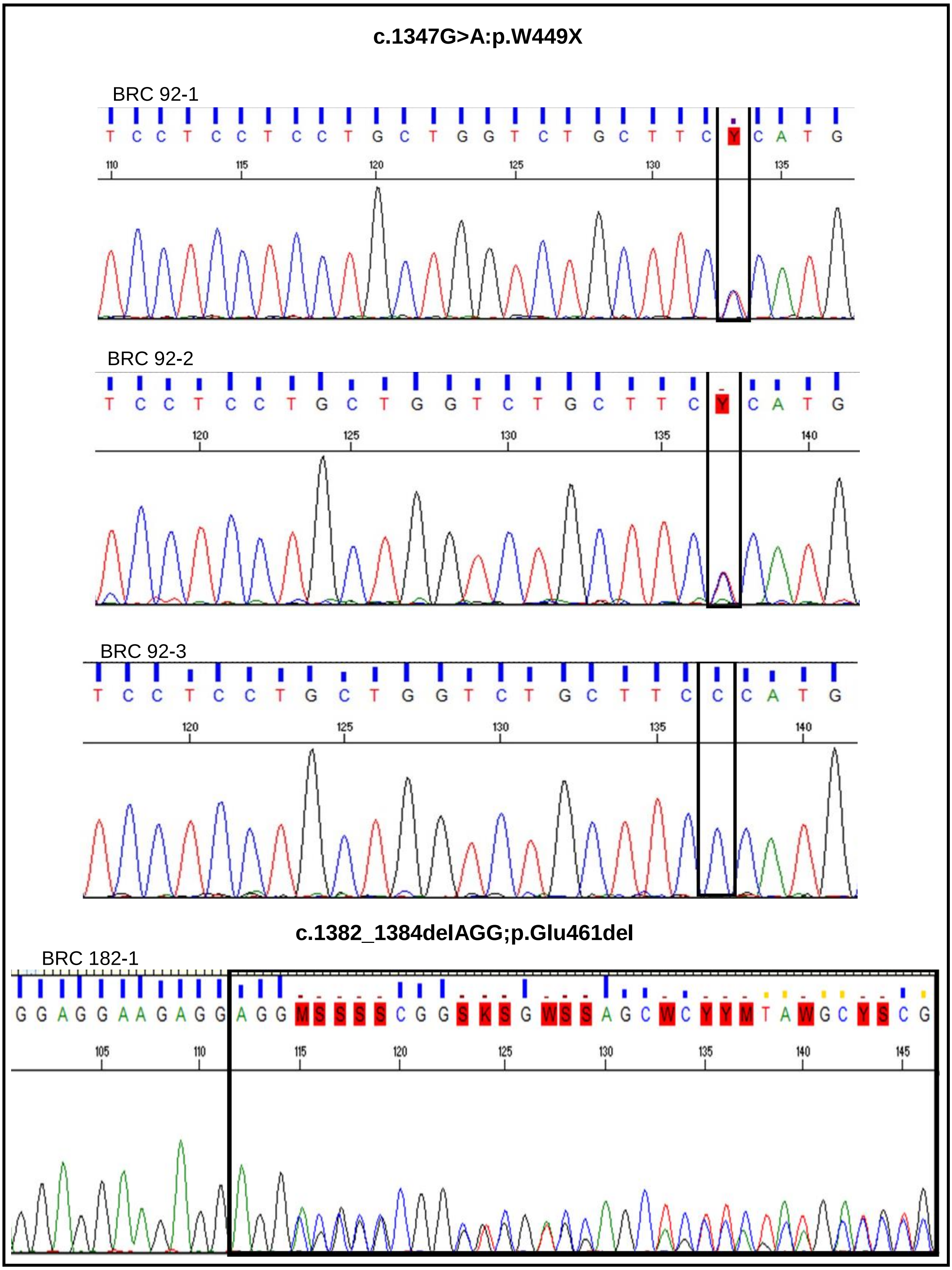
[b] HGVS nomenclature was used

[c] Carrier frequencies were calculated as the number mutation carriers (n) out of the total number tested (N = 69)

[d] All variants were cross-referenced with variant data available through the NCBI dbSNP database, build 38, release date 25.04.2013 (http://www.ncbi.nlm.nih.gov/projects/SNP/)

[e] 1000 Genomes browser, release date 16.10.2014 (http://browser.1000genomes.org). Minor allele frequencies (MAF) have been provided

[f] NHLBI Exome sequencing project (ESP) 6500, release version v.0.0.30, date 03.01.2014 (http://evs.gs.washington.edu/EVS/). Minor allele frequencies have been provided



c.1347G>A:p.W449X

c.1382_1384delAGG;p.Glu461del

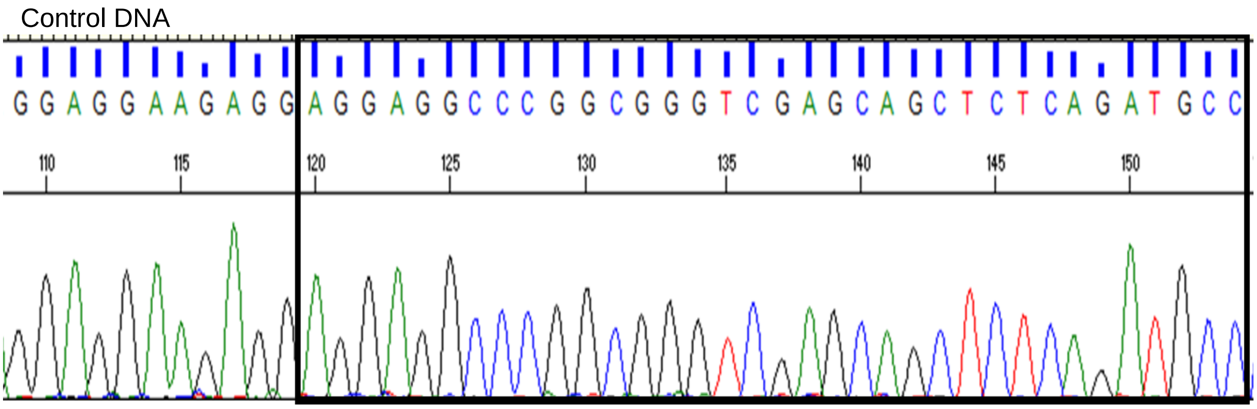


Figure 4.1: Electropherogram illustrating the nonsense and in-frame deletion variant in *TCHP* (NM_001143852) exon12.

Exon variants detected in proband cases by screening for mutations in *TCHP.* Mutation carriers are heterozygous for these variants as depicted by double peaks surrounded within black boxes. Sequences of the reverse (c.1347G>A:p.W449X) and the forward (c.1382_1384delAGG;p.E461del) strand have been provided.

4.3.2. Nonsense and in-frame deletion variants

The p.W449X stop-gain mutation was detected in one family (i.e. BRC 92) and a mutant allele frequency of 0.007 was calculated for the 61 breast/ovarian cancer families. This slightly exceeded the MAF seen in the ESP6500 for individuals of similar ancestry (i.e. European-American). Family 92 had a history of colorectal, liver, prostate cancer including melanoma and were diagnosed with breast cancer at an average age of 60y. The index case BRC 92-2 (III-6) and an affected first degree relative (III-5) carried this nonsense variant (Annexure 3A). These individuals were diagnosed at ages 62 and 50 respectively. The mutation did not always segregate with disease phenotype in this family as illustrated by their cousin (III-4), who did not carry the mutant allele. BRC 92-3 was diagnosed at 55 and may have developed BC by inheriting another disease-causing mutation through a different family line. The first degree relatives of this *TCHP* negative patient have a more extensive background of breast and prostate cancer cases than the relatives of III-5 and III-6. As a rule; variant segregation studies require mutations to be found in up to 50% of available affected relatives to be associated with the disease phenotype (Gracia-Aznarez, et al. 2013). The stop-gain variant was detected in >50% of affected cases within this family. However, additional experimental evidence is needed to prove its role in familial breast cancer. The c.1347G>A mutation is located in codon 449 and will lead to premature protein termination and the loss of ~10% (49 amino acids) of the native molecule. This mutant transcript could possibly be recognised for nonsense mediated decay (NMD) (Schweingruber, et al. 2013).

The novel in-frame deletion variant, p.E461del, was detected in only 1/61 of the breast/ovarian cancer families. Unfortunately segregation analysis was not feasible for this family as no other family members were available. The patient, III-1, has a family history of various cancers including stomach, brain, kidney, lung, breast, glandular cancer and melanoma (Annexure 3B); similar to the background of family 92. However, family 182 was diagnosed with breast cancer at an average age of 37.5y. The three base pair deletion (c.1382_1384delAGG) will result in the loss of one of seven glutamic acids at codon 461. However, very little is known about the structure of TCHP. Therefore, the impact of this in-frame variant is not yet clear.

Our study revealed that only two breast/ovarian cancer families carried this nonsense and amino acid deletion variant in *TCHP*. The late age of disease-onset in both these families suggest that *TCHP* truncating germline variants are similar to variants associated with a low penetrance hereditary breast cancer risk. No definite conclusions can be drawn at this stage.

4.3.3. Revealing the pathogenicity of nonsynonymous and intronic variants

SIFT, PolyPhen-2 and A-GVGD *in silico* analysis of nonsynonymous variants (missense and in-frame deletion) detected in this study revealed that c.1382_1384delAGG, c.380A>G, c.535A>C and c.1249G>A would have no impact on protein function. The majority of missense, intronic and all the synonymous variants, previously reported in dbSNP, were polymorphisms. The p.E417K mutation (c.1249G>A) has been reported in UniProt and was described as a natural variant (Consortium 2015). In addition to this polymorphism, p.K127R (c.380A>G) was identified with a reported minor allele frequency of 0.30 in the 1000 genomes and ESP6500 datasets. The polymorphic variants p.E417K and p.P7P have previously been reported in primary intestinal cancer tissue samples. Two intronic variants (i.e. exon 7:c.700-34A>G & exon13:c.1465-20T>C) have been reported with allele frequencies ≤0.01. In the current study group these variant alleles were discovered at frequencies of 0.014 and 0.007 respectively. However, no evidence of splice site disruption was found (data not shown). Therefore; none of the amino acid substitution or intron variants will have a predicted impact on the function of TCHP and are of unknown significance.

Three intronic variants were observed in ~14% (10/69) of the high-risk BC/OVC patients. This included variants in exon4:c.456+29T>C, exon5:c.457-38T>C and exon6:c.526-22_23insCCT. These variants have been reported in public genome databases such as 1000genomes and the ESP6500 project with allele frequencies averaging at 0.07 in individuals of European decent (Table 4.2). While this is an interesting observation the three variants may not have a significant effect on the TCHP coding sequence.

4.4. Conclusion

Sequence analysis of the *TCHP* coding region identified only one high-risk family with a premature truncation variant, p.W449X (family 92) and one carrying a single amino acid deletion (p.E461del, family 182). The low frequencies of these variants may either be indicative of a novel high-risk susceptibility gene that is rarely mutated or could purely be specific to the two unrelated families. Similar observations have been made with *TP53* which is a well-characterised cancer gene where significantly lower frequencies of pathogenic mutations have been documented in BC cases in comparison to any other tumours (Arcand, et al. 2015; Gasco, et al. 2002).

The decrease/loss of expression and mutation of TCHP has been discovered in different cancer types (including gastric, colorectal, breast, prostate etc.) (Fassan, et al. 2011; Kim, et al. 2010; Vecchione, et al. 2008). The current study is one of only two that have performed direct sequencing of the *TCHP* gene (Kim, et al. 2010) and the first to investigate its possible contribution to familial breast cancer cases. Very little has been published about the role of TCHP in maintaining cellular integrity. To date there are only six published studies which have investigated the functional role of TCHP as a pro-apoptotic protein (Ibi, et al. 2011; Inoko, et al. 2012; Kim, et al. 2010; Neill, et al. 2014; Nishizawa, et al. 2005; Vecchione, et al. 2008). *TCHP* is notably rarely mutated in somatic cancer data. To date, approximately 11 putatively deleterious *TCHP* variants have been identified in somatic cancer tissue ranging from lung, stomach, melanoma, colorectal as well as head and neck squamous cell carcinoma (Cerami, et al. 2012; Cline, et al. 2013; Gao, et al. 2013). Research findings made thus far have suggested that there may be merit in investigating *TCHP* as a human cancer predisposition gene. However, our work did not produce enough evidence to substantiate a link between potential truncation mutations in *TCHP* and breast/ovarian cancer risk. TCHP may still contribute to cancer as published data supports the hypothesis that this protein acts as a tumour suppressor because of its function in apoptosis (Vecchione, et al. 2008).

Chapter 5:

EME2 gene variants in high-risk South African breast and ovarian cancer families

5.1. Introduction

A number of recognised factors contribute to breast cancer development (Jardines, et al. 2011). Family history is an important component for increased BC risk and account for 5-10% of all BC cases (Apostolou and Fostira 2013). Hereditary BC is caused by germline mutations in proteins that play multiple roles in the recognition of DNA lesions and the activation and regulation of DNA repair (Dietlein and Reinhardt 2014). The most regulated is homologous recombination repair (Le Guen, et al. 2014; Loke, et al. 2015). The components of this pathway comprise a third of all genes that have been well-characterised for the role they play in BC/OVC predisposition. Disease-causal variants in homologous recombination repair (HRR) genes are known to contribute to up to 40% of familial cancer cases (Boyd 2014; Petrucelli, et al. 2013). Many studies have suggested that rare variants in HRR genes that have not been characterised for their role in BC predisposition may explain even more of the unknown familial BC risk (Filippini and Vega 2013; Hilbers, et al. 2013; Wen, et al. 2014).

HRR exchanges homologous DNA strands to repair various alternative DNA structures including single strand gaps, double strand breaks, interstrand DNA crosslinks and collapsed replication forks (Holthausen, et al. 2010). Double strand breaks (DSB) are some of the most toxic DNA lesions and must be repaired to preserve genomic stability (Dexheimer 2013). These are caused by either exogenous (e.g. ionising radiation) or endogenous sources (e.g. irregular replication) and can also occur as programmed DSBs. The last-mentioned lesions are site specific and are created by endonuclease enzymes in order to initiate HRR at positions where DNA replication has stalled. Delayed DNA replication can be found at genomic regions containing UV-induced thymine dimers (Mehta and Haber 2014). Stalled forks regress and form D loop holiday junctions which are resolved by the collective action of structure-specific endonucleases e.g. GEN1, BLM and SLX1/SLX4 (Jasin and Rothstein 2013). Therefore, these enzymes may be important components for genome maintenance. One such complex is the MUS81-EME1, which has been well-characterised for the role it plays in facilitating DNA repair (Minocherhomji and Hickson 2014). Its homolog MUS81-EME2 has recently also been under investigation (Pepe and West 2014).

The essential meiotic structure-specific endonuclease subunit 2 (EME2) forms a heterodimer with MUS81 that functions as an XPF family flap/fork endonuclease and facilitates homologous DNA repair (Shin, et al. 2012). The MUS81-EME2 heterodimer complex also functions as a structure-selective endonuclease that cleaves stalled DNA forks to allow fork restoration and progression of replication (Pepe and West 2014). This complex is composed of a catalytic and non-catalytic sub-unit which is characteristic to all XPF/MUS81 endonucleases (e.g. XPF-ERCC1 & FANCM-FAAP24) (Ciccia, et al. 2007; Pepe and West 2014; Shin, et al. 2012). Nucleolytic cleavage of DNA strands is performed by MUS81 as it possesses an active ERCC4 catalytic domain, unlike EME2 (Pepe and West 2013). However, EME2 is proven to direct processing of stalled forks during S-phase specifically (Amangyeld, et al. 2014). Double strand breaks are generated that in-turn facilitates HRR, ultimately creating an active fork (Pepe and West 2013). Unlike the homolog, MUS81-EME1, MUS81-EME2 plays a more versatile role in DNA repair and exhibits a higher nucleolytic activity (Amangyeld, et al. 2014). The MUS81-EME2 complex plays a secondary role in promoting telomere maintenance in ALT (alternative lengthening of telomeres) cells. The loss of EME2 expression has resulted in increased telomere-free chromosome ends in ALT cells (Pepe and West 2014).

The precise biological role of EME2 is still being determined, however, studies have suggested it assists in efficient replication of DNA (Amangyeld, et al. 2014; Pepe and West 2013; Pepe and West 2014). As such, this study set out to assess the potential role of *EME2* as a novel breast cancer susceptibility candidate gene.

5.2. Materials and Methods

5.2.1. *BRCAx* patient selection

Breast/ovarian cancer cases were selected for screening as described in Chapter 2. For this section of the study premature truncating variants (PTV) identified in *EME2* were further assessed by screening DNA samples from 190 BC patients and 15 OVC patients. These patients were negative for *BRCA1/2* mutations and did not have a family history of disease. Specific variants were also screened for in 75 unaffected individuals of Afrikaner ancestry which is representative of the population group under investigation.

5.2.2. *EME2* mutation screening

Genomic DNA of high-risk families and cohorts of women with breast and ovarian cancer were analysed through gene screening. PCR amplification was followed by uni-directional sequencing of the entire EME2 coding region (eight exons/fragments). Optimised conditions for both amplification and sequencing can be found in Table 5.1. Primers were designed for this study with the use of genomic sequences from the NCBI nucleotide sequence database (NC_000016.10) and variants were described according to *EME2* cDNA sequences obtained from the RefSeq database (NM_001257370 and NP_001244299.1). All variants classified as functionally deleterious were confirmed through bi-directional sequencing. Mutations that were detected were also validated by screening the affected family members of index cases for variants of interest (where DNA was available).

Table 5.1: *EME2* primers and amplification conditions

Exon	Primer sequence	Length	F/R	Ta	PCR Tann(°C)	PCR [MgCl$_2$]	Product Size
1	GTCCCAGGCTAAAGTGTTC	19	F	56.47	57	1mM	449bp
	TCTCCATTCGGCACAAAC [a]	18	R	55.4			
2	GGTTTGTGCCGAATGGAGAC [a]	20	F	57.45	60	1mM	371bp
	CTCCATCCTTGCTTTACCCA	20	R	55.4			
3	CTTGGCAGGAAAGGGAACAC [a]	20	F	57.45	60	1mM	421bp
	CTCCATCTCACCCTAGAAAC	20	R	55.4			
4	TTCAGGCTTGCTGTTCTGC [a]	19	F	56.47	60	1mM	273bp
	CACAGCAACCCCAGAAGTGT	20	R	57.45			
5	AGCTGATCCCACTTCTCCAG [a]	20	F	57.45	60	1mM	275bp
	TACCACGCTGGGAACCAAAC	20	R	57.45			
6	GTACATGGGGCAGCTATCAG	20	F	57.45	60	1mM	346bp
	TGCTGTGCAGAAGGAGAAGG [a]	20	R	57.45			
7	AAGGCTTCTCTCTGTCCCCA	20	F	57.45	60	1mM	288bp
	GAGCAGGCAAAGGCATGAGA [a]	20	R	57.45			
8	GTGGCTGATGCAGTTGTC [a]	18	F	57.67	61	1mM	359bp
	CAGGGTGTCTGGTCTGT	17	R	59.01			

Primers were designed with Primer 3-blast using the NCBI37/Hg19 genomic reference sequence, NC_000016.10
[a] Primers used during uni-directional sequencing

The potential consequence of all nonsynonymous, synonymous and intron mutations were evaluated with the *in silico* mutation analysis tools SIFT, Polyphen-2, Align-GVGD and human splice finder, as described in Chapter 2. Multiple sequence alignments for use in A-GVGD were made with amino acid sequences from seven species *Homo sapiens* (human), *Mus musculus* (mouse, NP_001156574.1), *Rattus norvegicus* (rat, XP_006246109.1), *Xenopus (Silurana) tropicalis* (western clawed frog, NP_001072860.1), *Bos taurus* (cattle, XP_005224673.1), *Macaca mulatta* (Rhesus macaque, XP_001118625.2), *Gallus gallus* (red junglefowl, XP_004945459.1).

Protein structure modelling was accomplished by submitting the wild-type and mutant amino acid sequence of EME2 to the SWISS-MODEL workspace (Biasini, et al. 2014). Altered/deleted amino acids have been emphasized in order to indicate the location of each variant in both models.

5.2.3. Chi square analysis

Statistical analysis was performed in order to compare the distribution of variants of interest in patients versus 75 unaffected individuals. Only variants identified in >1 index case were investigated. Calculations investigating statistical significance were used to assess the association of putative pathogenic/disease-causal variants with breast cancer risk. Based on the heterozygous model χ^2- test were performed with familial *EME2* variant carrier frequencies in comparison with unaffected individuals in order to assess the significance of a hereditary component. Carrier frequencies of breast and ovarian cancer cohorts, without a family history, was used to determine the ultimate significance of observed rare *EME2* variants and BC risk. Applying chi square methods tested the deviation from Hardy-Weinberg equilibrium and p-values lower than 0.05 indicated a significant association.

In combination with odds ratios and relative risk estimates were calculated with the use of two-by-two tables. Crude values were calculated according to Altman, 1991, using the MedCalc vs15.11 web-based free statistical calculator (Altman 1991). Values corresponding to 95% confidence intervals were considered significant.

5.3. Results and Discussion

5.3.1. *EME2* germline variants

The entire EME2 coding region was first screened in a further 56 high-risk breast/ovarian cancer families. The nonsense mutation was detected in 5.3% (3/56) of these families. Two other potentially deleterious variants i.e. a frameshift (p.G55VfsX19) and a novel in-frame (p.E98_L100del) were identified as well. These rare variants also account for 5.3% of the high risk BC/OVC families. All the *EME2* variants are listed in Table 5.2 and Figure 5.1 illustrates their gene locations. Exons 1, 2 and 7 were labelled as potential hotspot regions and were targeted to validate the role of *EME2* as a potential novel candidate BC predisposition gene. Germline DNA was obtained for two cohorts including 190 patients with BC and 15 with OVC without a family history for the disease. Three separate variants including the p.G55VfsX19, an additional novel mutation in exon7 i.e. p.S269del (c.805_807delTCC) and the p.Q322X variant were found in 11.6% (11/190) of the BC cases. Of the ovarian cancer patients, 13% (2/15) carried the p.Q322X variant.

A total of 13 variants of uncertain significance (VUS) were observed among the breast/ovarian cancer patients included in this study. These consisted of missense (five), synonymous (six) and intron (two) variants, six of which were novel (Table 5.2 & Figure 5.1).

Table 5.2: ***EME2*** **gene variants identified in South African familial breast/ovarian cancer cases**

	Exon	Nucleotide altered [a]	Amino acid altered [a]	Familial BC Cases	Cancer cases - no family history		dbSNP [d]
				Number (carrier frequency)[b]	Number BC (carrier frequency) [c]	Number OVC (carrier frequency) [c]	
Truncating	Ex1	c.162delG	p.Gly55ValfsX19	2 (0.03)	9 (0.05)	2 (0.13)	rs377074694
	Ex7	c.964C>T	p.Gln322X	4 (0.06)	11 (0.06)	-	rs61753375
In-frame deletion	Ex2	c.293_301del	p.Glu98_Leu100del	1 (0.01)	-	-	-
	Ex7	c.805_807delTCC	p.Ser269del	-	2 (0.01)	-	-
Missense	Ex1	c.58C>T	p.Arg20Trp	-	1 (0.005)	-	rs376006516
	Ex6	c.772C>T	p.Pro258Ser	1 (0.01)	-	-	-
	Ex7	c.872C>T	p.Ala291Val	1 (0.01)	-	-	rs114640322
	Ex8	c.1049G>A	p.Arg350His	5 (0.07)	1 (0.005)	-	rs61753376
	Ex8	c.1055G>A	p.Arg352His	1 (0.01)	-	-	rs147967590
Synonymous	Ex1	c.216C>G	p.Val72=	30 (0.43)	7 (0.04)	1 (0.06)	rs761065
	Ex2	c.270C>T	p.Ala90=	2 (0.02)	-	-	rs199761703
	Ex5	c.630C>T	p.Ala210=	1 (0.01)	-	-	rs139829498
	Ex6	c.699C>T	p.Asp233=	2 (0.02)	-	-	rs139103048
	Ex7	c.810C>T	p.Phe270=	-	2 (0.01)	-	-
	Ex7	c.951C>A	p.Ser317=	-	1 (0.005)	-	-
Intronic	Ex3	c.385-74C>T		3 (0.04)	-	-	rs141044525
	Ex8	c.970-28G>A		2 (0.02)	-	-	rs190542947

[a] The HGVS nomenclature was used

[b] Frequencies were calculated as the number of *EME2* mutation carriers (n) out of the total number of individuals tested (N=69)

[c] Carrier frequencies were calculated as the number mutation carriers (n) out of the total number tested N= 190 (BC), 15 (OVC)

[d] All variants were cross-referenced with variant data available through the NCBI dbSNP database, build 38, release date 25.04.2013 (http://www.ncbi.nlm.nih.gov/projects/SNP/)

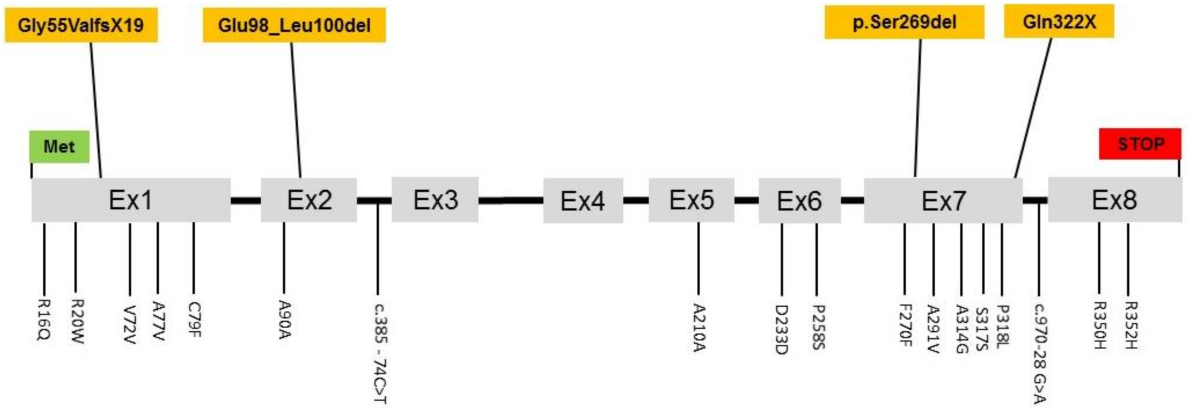


Figure 5.1: Schematic illustration of sequence variants identified from *EME2* whole gene screening.

Truncating and in-frame deletion variants are highlighted in yellow.

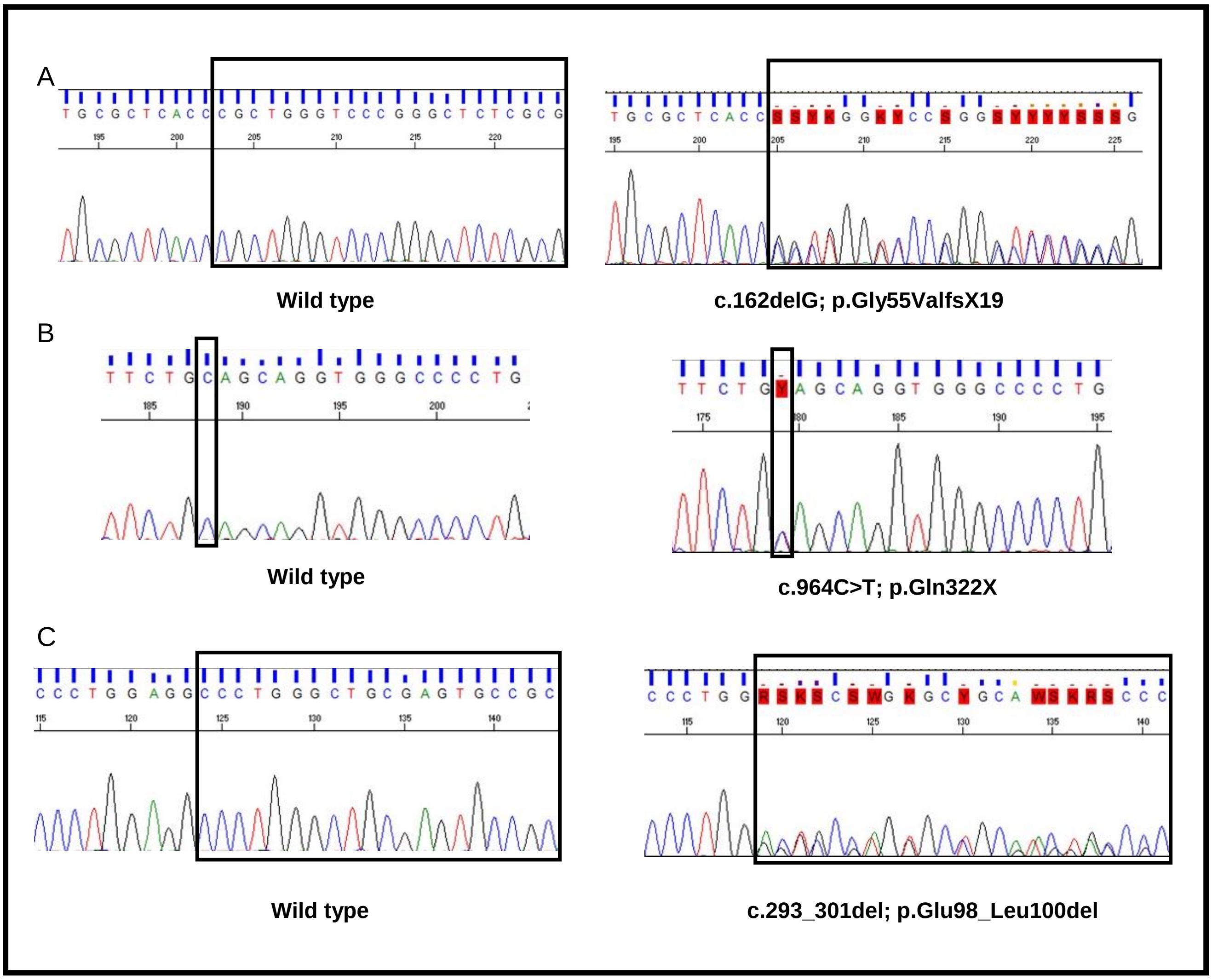
5.3.2. Deleterious variants identified in *EME2*

5.3.2.1. Frameshift variant, p.G55VfsX19

The exon1 frameshift variant c.162delG (Figure 5.2A) was identified in 3% of all the familial BC/OVC index cases (WES and Sanger sequencing, Table 5.2). The *EME2* p.G55VfsX19 variant was identified in 2/3 available cases in family 4 and in one of the two cases screened in family 157 (Annexure 3B), thus carried by approximately half of the individuals in these two breast cancer families. Among the patients without a BC/OVC family history the *EME2* truncation variant was found in ~5% (9/190) of the breast- and ~13% (2/15) of the ovarian cancer cases (Table 5.2). Public databases 1000genomes and ESP6500 have previously reported this variant allele at frequencies of 1% and 2% in individuals of European decent. The variant has not been identified amongst European individuals tested in the ExAC dataset (Table 5.3). *EME2* p.G55VfsX19 generates a premature termination codon (PTC) (Annexure 6A) and as such may trigger nonsense mediated decay.

Table 5.3: Allele frequencies of *EME2* truncating and in-frame deletion variants

Nucleotide altered	**Minor allele frequencies**					
	Familial BC Cases	**Cancer cases- no family history**	**1000 Genomes**	**ESP6500**	**ExAC (Fin)**	**ExAC (NFin)**
c.162delG	0.014	0.025	0.01	0.022	-	-
c.964C>T	0.028	0.025	0.017	0.014	0.004	0.013
c.293_301del	0.007	-	-	-	-	-
c.805_807delTCC	-	0.004	-	-	-	-



Wild type

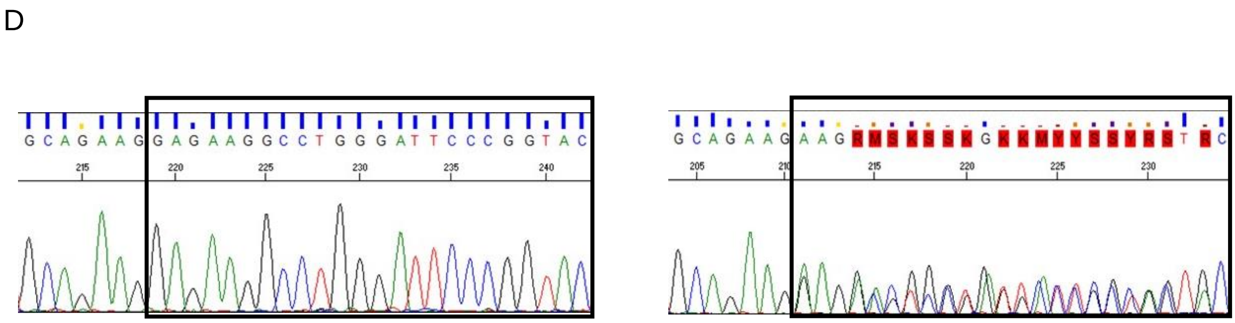
c.162delG; p.Gly55ValfsX19

Wild type

c.964C>T; p.Gln322X

Wild type

c.293_301del; p.Glu98_Leu100del



Wild type

c.805_807delTCC; p.Ser269del

Figure 5.2: Electropherogram of *EME2* putative pathogenic variants.

The frameshift c.162delG (A), nonsense c.964C>T (B) variant and in-frame deletions c.293_301del (C) & c.805_807delTCC (D). Sequences of reverse strands have been provided for all mutations with the exception of the exon 2:c.293_301 deletion variant. Heterozygous germline variants are depicted by double peaks which are indicated in black boxes.

5.3.2.2. Nonsense variant, p.Q322X

A stop-gain variant in exon7 of *EME2* i.e. p.Q322X (c.964C>T, Figure 5.2B) was discovered in ~6% of high-risk breast/ovarian cancer patients from four separate families (Table 5.2). One of the families (BRC 9), had a history of five BC affected individuals. The index case from family 9 was previously diagnosed with ovarian cancer at age 56 and further screening revealed that four of her first and second degree relatives carried this mutation (Annexure 3A). The c.964C>T nonsense variant was absent in her daughter, unaffected at the time of DNA sample collection. Relatives of *EME2* mutation carriers of family 96 and 121 were screened for the variant of interest. None of the relatives, screened for the variant of interest, tested positive for this mutation. The two non-carrier daughters of family 96 may have inherited a mutant allele from their paternal family line explaining their early age breast cancer diagnosis. The same may be said for the individual from family 121 carrying the wild-type allele, however no information was available to describe the cancer history on the paternal side of these index cases.

On performing mutation analysis in two cohorts of BC and OVC cases without a family history for their disease, the *EME2* c.964C>T variant was identified in 5.7% of the BC patients (11/190, Table 5.2) only. This variant allele has previously been reported in European and

European American population groups at frequencies of 1.7% and 1.4% respectively. Within the ExAC database this mutation has been identified at frequencies of 1% and 0.4% in European non-Finnish and Finnish individuals respectively (Table 5.3). *EME2* p.Q322X may result in the loss of 15% of the c-terminus of this protein. If expressed *EME2* p.Q322X will lack three out of seven important hydrophobic residues in the formation of its helix-hairpin-helix $(HhH)_2$ domain. (Annexure 6A). The HhH domains mediate interactions between EME2 and DNA strands in combination with MUS81 (Pepe and West 2013; Shao and Grishin 2000).

5.3.2.3. Novel in-frame variants

The two novel mutations that were discovered included a 9-bp in-frame deletion in exon2 of *EME2* (c.293_301del, Figure 5.2C). This potentially deleterious variant was identified in one of the high-risk families (i.e. BRC 190) (Table 5.2). The p.E98_L100del mutation was not detected in any of the relatives of this index case. The second novel variant discovered in *EME2* exon7 was a 3-bp deletion (c.805_807delTCC, Figure 5.2D) found in 1% of the breast cancer cases without a family history. The allele frequencies of both novel potentially pathogenic variants were very low and together they were detected in 1.1% of all the unexplained high-risk breast/ovarian cancer patients. This may be indicative of rare high-risk alleles, however, currently there is not enough evidence to support such a claim.

The impact of protein in-frame truncating variants were assessed through protein modelling (Annexure 6B). The two in-frame deletion variants in exon 2 and 7 were mapped on a 3D secondary EME2 model. The *EME2* exon2:c.293_301 deletion may affect codons 98-100 which form part of an α-helix loop. The exon7:c.805_807delTCC deletion will result in the loss of codon 269 which is located in a coiled-coil protein region. However, neither the novel in-frame variants are predicted to significantly impair EME2 structure and/or function.

As such, a total of 35 selected breast and ovarian cancer patients, screened for mutations in either the entire coding region or potential mutation hotspot regions, carried putatively deleterious *EME2* variants. Two mutations may lead to premature protein truncation and include a nonsense (Q322X) and frameshift (G55VfsX19) mutation and two are novel in-frame variants (E98_L100del & S269del). Mutation positive cases consist of 7/61 breast/ovarian cancer families and 24/205 breast/ovarian cancer patients without a family history for the disease.

5.3.2.4. Statistical analysis results

Statistical analysis was performed of the frameshift and nonsense variant that were identified in more than one index case. *EME2* c.162delG was found in 6.7% of the Afrikaners unaffected with disease indicating that the mutation was not associated with increased breast/ovarian cancer risk (p-value=0.05, Table 5.4). This was verified by relative risk calculations of all BC/OVC patients carrying the *EME2* c.162delG variant (RR: 0.584, 95% CI:0.17-1.90, p-value: 0.37). The high carrier frequencies observed amongst the unaffected individuals screened characterises *EME2* c.162delG as a polymorphism. *EME2* c.964C>T was identified in 2.7% of the unaffected cases. Therefore it was not associated with breast/ovarian cancer (p-value=0.5, Table 5.4). This was verified by the observed RR (RR: 1.41, 95% CI:0.78-2.56, p-value:0.25). In essence, our results could not verify that p.Q322X has any effect on breast cancer susceptibility either. This suggests that it may be a common polymorphism in the Afrikaner population group.

Table 5.4: Analysing the association of *EME2* truncating variants with BC/OVC risk

	BC/OVC families			**Cohort BRW and OVC cases**		
	χ² - Test statistic	**Interpret (p=0.05; df=1)**	**p-value**	**χ² - Test statistic**	**Interpret (p=0.05; df=1)**	**p-value**
p.Gly55Val fsX19	1.12	No association	0.1<*p*<0.5	0.373	No association	0.1<*p*<0.5
p.Gln322X	0.889	No association	0.1<*p*<0.5	0.92	No association	0.1<*p*<0.5

Formula of chi-square test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$

The evidence produced thus far does not prove that *EME2* p.G55VfsX19 or p.Q322X confers a risk for breast cancer, however, it does not exclude their potential role in the progression of this disease. The two truncating novel variants i.e. p.E98_L100del (c.293_301del) and p.S269del (c.805_807delTCC) may be worth investigating further. Collectively both in-frame deletions were detected in 1.2% of all BC/OVC cases (allele frequency 0.005). The discovery of four truncating variants are amongst other coding variants. Missense and non-coding variants that were identified were then analysed.

5.3.3. Analysis of amino acid substitutions and intronic variants

The total missense variants consisted of one novel and four that have previously been identified. All known non-truncation variants were queried against public databases 1000genomes and ESP6500 and 90% had allele frequencies <0.01. Only one (i.e. R350H) had previously been reported with minor allele frequencies >0.01. *In silico* predictive tools, SIFT and PolyPhen-2 produced predictive scores for two nonsynonymous amino acid substitutions (p.R20W and p.P258S) that reflect a potentially damaging effect on protein structure (Table 5.5). The sequence alignment based tool, A-GVGD, predicted that both mutants were neutral and therefore unlikely pathogenic, excluding them from further consideration (Mathe, et al. 2006).

Table 5.5: *In silico* predictions indicating the potential effect of the amino acid substitutions on protein function

Amino acid altered	***In silico* prediction tools**		
	SIFT	**PolyPhen-2**	**Align-GVGD**
p.Arg20Trp	0.03 - Deleterious	0.94 - Probably damaging	C0 - Less likely
p.Pro258Ser	0.01 - Deleterious	1.00 - Probably damaging	C0 - Less likely
p.Ala291Val	0.09 - Tolerant	0.124 - Benign	C0 - Less likely
p.Arg350His	0.06 - Tolerant	0.034 - Benign	C0 - Less likely
p.Arg352His	0.15 - Tolerant	0.198 - Benign	C0 - Less likely

The synonymous variants identified consisted of five rare and one common mutation. One polymorphic allele (c.216C>G) in exon1 was found to be more prominent in patients of European descent (Table 5.2). This finding is also seen amongst European population groups in public genome variation databases. Intronic variants were only identified among breast/ovarian cancer families. *In silico* algorithms accessed through human splice finder were not able to predict significant alteration of splicing events. The relatives of familial BC/OVC individuals carrying missense, synonymous or non-coding variants were not screened as the mutations were considered to be of little significance. Future functional analysis may assists in elucidating their potential impact on protein function.

5.4. Conclusion

Screening of *EME2* led to the discovery of four mutations predicted as potential protein truncation variants. These are *EME2* p.G55VfsX19, p.E98_L100del, p.S269del and p.Q322X identified at allele frequencies ranging from 0.007 – 0.02 in cancer patients. Of the two known mutations *EME2:*c.162delG was predicted to have the most damaging structural/functional effect on the protein and was detected in up to 50% of the available family members making it a variant of interest for breast cancer predisposition. *EME2:*c.964C>T displayed characteristics of a low to moderate disease allele as it displayed incomplete segregation (~40%) amongst the available family members of high-risk BC/OVC cases carrying this mutation.

Neither the truncating (frameshift and nonsense) variants were significantly associated with susceptibility to breast/ovarian cancer. Both appeared to be polymorphic gene variants amongst subjects unaffected by disease. These individuals were representative of the population of this country. It is interesting that the c.162delG allele frequency amongst the unaffected individuals (0.03) was three times higher than what has been reported for individuals of similar descent (Europeans) in the 1000genomes database. In 2001 Garte and colleagues determined that a small degree of genetic heterogeneity exists within ethnic populations of the same descent. A gene variant could be classified as a polymorphic SNP in one group and be monomorphic in another group from a different national origin. Such a demographic variable could affect the true population frequency of rare alleles under investigation for an association to disease susceptibility (Garte, et al. 2001).

Even though the evidence produced thus far does not conclusively prove that p.G55VfsX19 or p.Q322X confers a risk for breast/ovarian cancer, it does not exclude their potential role in the progression of this disease. However, the two novel in-frame deletion variants could potentially be pathogenic. Although subjects unaffected by disease were not screened for the exon2:c.293_301del variant the exon7:c.805_807delTCC mutation was not identified in any of these healthy individuals. Therefore these putatively deleterious variants could still be analysed further in future investigations.

To our knowledge this is the first study investigating *EME2* as a potential candidate gene in high-risk breast and ovarian cancer patients. More evidence is needed to determine whether germline truncation variants in *EME2* may account for some of the unexplained breast/ovarian cancer cases found in South Africa and worldwide. However, future studies with greater number affected patients may help determine whether these variants are true loss-of-function alleles that may in-fact predispose to disease. Such gene level meta-analyses are useful for validating the role of rare gene variants in BC susceptibility and to curate newly identified gene variants in order to determine their pathogenicity (Feng, et al. 2015; Loke, et al. 2015).

Although the results obtained from our study do not support an appreciable association between germline variants in *EME2* and BC/OVC risk, we cannot categorically rule out that loss-of-function variants may not disrupt its ability to form a heterodimer with MUS81. Recent evidence has suggested that EME2 could play a role in HRR while associated with the MUS81 endonuclease (Amangyeld, et al. 2014; Pepe and West 2014). This heterodimer cleaves nicked nucleotide strands and channels aberrant DNA duplexes through recombinant repair processes with increased efficiency and versatility than its less active counterpart, MUS81-EME1 (Pepe and West 2013). Computational analysis verified that all four protein truncation variants may have a negative impact on EME2 protein structure with the exon1:c.162delG and exon7: c.964C>T mutation being possible targets for mRNA degradation. The two in-frame variants are located in α-helix or coiled-coil regions of the protein structure. Deletion mutations tend to destabilise most secondary protein structures and as a result, protein function (Batra 2009; Henzler Wildman, et al. 2002). The MUS81-EME2 complex also contributes towards maintaining telomere length, which may also be disrupted by the truncation variants observed in this study (Pepe and West 2013). However, the functional role of this heterodimer may still be fulfilled by hemizygous expression of EME2.

Double strand break tolerance is not uniquely dependant on MUS81-EME2 complexes but includes the involvement of additional factors with a redundant functional role including ERCC4-ERCC1 & FANCM-FAAP24 (Pepe and West 2014). Variants that disrupt the expression of EME2 may still diminish DNA repair efficiency somewhat and increase genomic instability which is one of the most important hallmarks of breast/ovarian cancer progression (Hanahan and Weinberg 2011). However future functional studies are needed to support this theory.

Chapter 6:

HELQ gene variants in high-risk non-*BRCA1/2* South African families

6.1. Introduction

The missing heritability seen in high-risk *BRCAx* families may best be described by rare variant alleles in recently characterised genes (Hilbers, et al. 2013). Such novel high-risk alleles may be more uncommon than known variants in *BRCA1* and *BRCA2* that account for 60% of familial breast cancer cases (Hilbers, et al. 2013; Karami and Mehdipour 2013). Gene regulatory network analyses revealed that the most functionally significant variants which contribute towards the development of BC are located in genes that are involved in the maintenance of genome stability (Emmert-Streib, et al. 2014).

Homologous recombination (HRR) and non-homologous end joining (NHEJ) are two pathways which repair double strand DNA (dsDNA) breaks and interstrand crosslinked (ICL) DNA (Dexheimer 2013; Le Guen, et al. 2014). The Fanconi Anemia (FA) pathway is a widely integrated process comprised of up to 15 genes which not only provide cellular recovery of ICL DNA lesions but collaborates with proteins that are mainly involved in HRR, translesion synthesis and nucleotide excision repair thereby connecting these mechanisms (Kottemann and Smogorzewska 2013). FA and HRR are two pivotal, error-free modes of DNA repair and are comprised of multiple integrating factors (Moldovan, et al. 2012). Many known BC genes code for proteins that play critical roles in regulating these pathways in response to dsDNA damage (Wang, et al. 2015a; Yoshida and Miki 2004). DNA helicases are examples of such tumour suppressors as they also play pivotal roles in DNA repair and have been linked to cancer predisposition (Brosh and Robert 2013). This family of proteins is classically known for unwinding dsDNA during replication but also functions towards genome maintenance, ultimately promoting overall cellular homeostasis (Khan, et al. 2015). Recent evidence has confirmed that helicase family proteins, such as the RecQ proteins, WRN and BLM are involved in breast cancer. Germline variants in these proteins predispose to hereditary diseases such as Blooms -, Werner – and Rothmund–Thomson syndrome (Croteau, et al. 2014). Individuals affected by these diseases are part of families where high incidences of breast/ovarian cancers are observed (Broustas and Lieberman 2014). Considerable efforts have been made to identify and describe the function of DNA helicases (Brosh and Robert 2013; Croteau, et al. 2014). Helicase Q (a.k.a. Helicase 308) is an ATP-dependent enzyme that unwinds DNA in a 3' to 5' direction (Marini and Wood 2002; Tafel, et al. 2011). Evidence has been produced which suggest that HELQ forms a potential physical association with replication protein A (RPA). These proteins were shown to co-express and interact via the HELQ conserved C terminal region to bind to DNA structures for strand displacement during DNA repair (Woodman, et al. 2011).

HELQ binds to RAD51 and contributes towards ICL repair (Clauson, et al. 2013). Recent studies have revealed that HELQ associates directly with the RAD51 paralogue complex BCDX2 (RAD51B, C, D and XRCC2), promoting efficient homologous recombination repair (Adelman and Boulton 2010; Adelman, et al. 2013). The RAD51 paralogue is comprised of a group of related proteins that mediate the repair of dsDNA lesions through homologous recombination (Chun, et al. 2013). Even though the exact mechanism of HELQ remains unresolved, findings from Takata and colleagues have confirmed that it plays a BCDX2-directed role at interstrand crosslinked sites (Takata, et al. 2013). Studies suggest that HELQ may be a putative tumour suppressor protein due to its dual role in DNA replication and repair (Adelman, et al. 2013).

Alterations in amino acid regions that correspond to the HELQ helicase-like core domain results in cellular hypersensitivity to DNA crosslinking agents. Defects in this protein have been associated with HRR defective, mitomycin C (MMC) hypersensitive endocervical cancer cells (Takata, et al. 2013). Mice that carry functionally disruptive mutations in the HELQ coding region possess haematopoietic stem and progenitor cells with increased ICL DNA lesions and are predisposed to tumour development. Mice heterozygous for loss-of function HELQ mutations present with similar, yet less severe symptoms (Adelman, et al. 2013; Luebben, et al. 2013; Williams and Michael 2010).

The interactions between *HELQ* and *RAD51* paralogues has classified it as a recognised ovarian cancer gene (Adelman, et al. 2013). Heterozygous loss of its function is comparable to that of RAD51C and D (Takata, et al. 2013). The prominent role of HELQ in DNA damage repair has been investigated, but further research is still being done to explain its potential as a breast cancer predisposing gene (Hamdi, et al. 2012; Pelttari, et al. 2015; Tafel, et al. 2011). Due to the interconnected role that helicase-Q plays with proteins that are coded by well-known breast cancer genes, studies have also investigated the involvement of *HELQ* gene mutations in breast cancer predisposition (2012; Pelttari, et al. 2015; Wang, et al. 2015b). We hypothesize that germline variants in the *HELQ* gene may contribute to the non-*BRCA1/BRCA2* breast/ovarian cancer families in our local study population.

6.2. Materials and Methods

6.2.1. *BRCAx* families

Fifty-six additional families with high-risk breast/ovarian cancer were selected in this study. The 61 cases (described in Chapter 2) were screened for germline variants in the entire HELQ coding region.

6.2.2. *HELQ* mutation screening

HELQ coding regions were amplified using previously extracted blood DNA samples. Primers were designed for this study to amplify the 18 exons (including exon-intron boundaries) of this gene in 21 fragments (Table 6.1). Target exon regions were amplified, purified and analysed through uni-directional cycle sequencing as explained in Chapter 2. Potential deleterious variants of interest were verified through bi-directional sequencing in proband cases and their family members (where available). Sequence traces were visualised and compared to reference sequences obtained from the NCBI RefSeq database i.e. NC_000004.12 (genomic) and NM_133636 (mRNA) (2015). Variants were described with the *HELQ* cDNA and protein sequence (NP_598375.2) according to HGVS guidelines (Dunnen and Antonarakis 2000).

Table 6.1: *HELQ* Primers and amplification conditions

Exon	Primer sequence	Length	F/R	Ta	PCR Tann(ºC)	PCR $[MgCl_2]$	Product Size
1	CCCAAGGTGGATGTAGAAGCG [a]	21	1F	58.3	60	1.5	455bp
	CTGGGAAGGATGCCAAAAGT	20	1R	55.4			
2A	ACTGGAATTGACCATAAGCG	20	2AF	53.35	56	1.5	497bp
	ATTTTGCTGGGGTTGCTCTAT [a]	21	2AR	52.47			
2B	TGTACCTTCCTCACAGGCTAT [a]	21	2BF	54.42	58	1.5	394bp
	GTCTCTCACTTTGCTGGGTA	20	2BR	55.4			
2C	CATGACTGGAAATGCGAAGG [a]	20	2CF	55.4	59	1.5	359bp
	GTAGCAGGACAGGTAGGAATGG	22	2CR	57.27			
3	CCTTATGAGAGGATTGCTCCAC [a]	22	3F	55.4	53	1.5	385bp
	TACTTCTCAACAATACAAAGT	21	3R	46.61			
4A	TTACTGCAGCCTCAATTCTCTG [a]	22	4AF	53.53	56	1.5	409bp
	AAGGAGTTCACCAAGCTATGTC	22	4AR	53.53			
4B	GTTGAAGAATATGCTGGAAGCAA	23	4BF	50.94	53	1.5	315bp
	ATCAAGCATAGAATGTCTTCACCT [a]	24	4BR	50.27			
5	AGCTCATTAACCTAAATACCAGTGA	25	5F	49.6	54	2.0	373bp
	ACAAAGGCAAAAAGCTTAAGTATC [a]	24	5R	48.57			
6	TCAGAGCACAAAAGACTTGCTA [a]	22	6F	51.67	56	1.5	315bp
	CATGGTGAGTAAAATATGCCCC	22	6R	53.53			
7	AGGGAAAAAGAGACCAAAACAGA	23	7F	50.94	54	1.5	423bp
	TGTCCAACAAATGTCAAGCATC [a]	22	7R	51.67			
8	ACCTTTCAAATCACACTGGTAA	22	8F	49.81	54	2.0	379bp
	TTTCTTTCTTCAAGGTCAAGGC [a]	22	8R	51.67			
9	CCACAAAAAGATTGAGAGACTGC [a]	23	9F	52.73	56	1.5	463bp
	GAGCCTGGTCAACACTTACTTA	22	9R	53.53			
10	ATCATAGATTCTTGGTAAACAC [a]	22	10F	47.97	51	1.5	365bp
	AAAGATGATTTCATAAAAGGGC	22	10R	47.94			
11	TTTTGGTCTCCTTCTGGCTCCAC [a]	23	11F	56.29	58	2.0	398bp

	CATTAATCACTGAAGGGGATGAGA	24	11R	51.98			
12	AGCTTTGAGGTATGGAAATGAGT [a]	23	12F	50.94	54	2.5	369bp
	TTTCCAATGTGTTTCCTATGAT	22	12R	47.94			
13	CCCCTCTGGTAATTGTGAAAGA	22	13F	53.53	58	1.5	343bp
	TGGGAAGACTGCTTGAGCCTA [a]	21	13R	56.38			
14	ACTGGATTTGGTTCACATGACA	22	14F	51.67	55	2.0	333bp
	GGGTGCTCAGGTATGTTTTAAG [a]	22	14R	53.53			
15	ACTTCAGTTTTCCTTTTGTCCA [a]	22	15F	49.81	54	1.5	351bp
	ATGCTGACATACTGACTAACTGG	23	15R	52.73			
16	CCCCTATCACACAAACCAAACT	22	16F	53.53	57	1.5	296bp
	TAAGGCTCTTGCTATCACTTCC [a]	22	16R	53.53			
17	TATCCCAGTCTCTTACATGCAG [a]	22	17F	53.53	55	2.0	298bp
	ACAAGGAAATAACACTACGCAAG	23	17R	50.94			
18	TCTGATGTGTAAGGATTCAGG [a]	21	18F	52.47	54	2.0	231bp
	GAAAATTCTCATCAGATAGC	20	18R	49.25			

Primers were designed with Primer 3-blast using the NCBI37/Hg19 genomic reference sequence, NC_000004.12
PCR primers and optimised conditions for each exon have been indicated
a: Primers used during uni-directional sequencing

The potential functional impact of amino acid substitutions and intronic variants were assessed using SIFT, PolyPhen-2, A-GVGD and human splice finder vs3.0 *in silico* tools as described in Chapter 2. Multiple sequence alignments were made using amino acid sequences from eight species including; *Homo sapiens* (human), *Mus musculus* (mouse, NP_001074576.1), *Rattus norvegicus* (Rat, NP_001014156.2), *Bos taurus* (cattle, XP_010804613.1), *Macaca mulatta* (Rhesus macaque, XP_001104832.1), *Pan troglodytes* (chimpanzee, XP_003310356.1), *Gallus gallus* (chicken, XP_004941184.1), *Danio rerio* (zebrafish, NP_001269352.1).

6.3. Results and Discussion

6.3.1. *HELQ* gene variants

HELQ gene screening revealed a total of 28 germline sequence variants in 50 of the high-risk families (WES and Sanger sequencing). These include one frameshift, five missense, three synonymous and 19 intronic variants as well as one variant in the 5' untranslated region (UTR) (Table 6.2). Variants in *HELQ* were distributed throughout the coding region. The highest frequency of cases (>80%) carried germline mutations in non-coding regions and known polymorphic variants of unknown significance. A total of six rare coding variants were detected in 5% (4/69) of this study group and were mainly found in the N-terminal region of *HELQ*. This included four previously reported and two novel variants in 3.2% (2/61) and 3.2% (2/61) of the BC/OVC families respectively. We identified three variants of potential significance i.e. a frameshift (c.1358delG;p.R453KfsX9) in exon4, a missense (c.1924T>C;p.Y642H) in exon9 and an insertion mutation (c.1191+1_1191+2insGT) at the 5' (aka donor) splice region of exon3. The frameshift mutation was predicted to disrupt the expression of this gene and may be pathogenic.

Table 6.2: Sequence variants identified in the *HELQ* gene

	Exon [a]	Nucleotide altered [b]	Amino acid altered [b]	Number (carrier frequency) [c]	dbSNP [d]	Minor allele frequencies				
						This study	1000 Genomes [e]	ESP6500 [f]	ExAC (Fin) [g]	ExAC (NFin) [g]
Truncating	Ex4	c.1358delG	p.Arg453LysfsX9	1 (0.01)	-	0.007	-	-	-	-
Missense	Ex1	c.47A>G	p.Lys16Arg	1 (0.01)	rs150540222	0.007	0.00	0.002	-	0.002
	Ex2	c.704T>C	p.Leu235Pro	1 (0.01)	rs17006837	0.007	0.00	0.003	-	0.002
	Ex2	c.916G>A	p.Val306Ile	33 (0.48)	rs1494961	0.23	0.51	0.50	0.52	0.49
	Ex8	c.1753C>T	p.Pro585Ser	1 (0.01)	rs6817280	0.007	0	0.002	-	0.0008
	Ex9	c.1924T>C	p.Tyr642His	1 (0.01)	-	0.007	-	-	-	-
Synonymous	Ex3	c.1036T>C	p.Leu346=	34 (0.49)	rs13141136	0.24	0.38	0.41	0.46	0.40
	Ex6	c.1482T>C	p.Ile494=	36 (0.52)	rs7665103	0.26	0.38	0.41	0.47	0.41
	Ex12	c.2325T>C	p.His775=	1 (0.01)	rs59255439	0.007	-	0.002	-	0.001
Intronic	Ex1	c.-29_-27delACG		2 (0.03)	rs139503945	0.014	0.00	0.002	-	0.002
	Ex2	c.1013-12delT		4 (0.06)	rs532161833	0.028	0.01	-	-	-
	Ex3	c.1191+1_1191+2inGT		1 (0.01)	-	0.007	-	-	-	-
	Ex4	c.1192-61_1192-38del24		27 (0.39)	rs11272773	0.19	0.49	0.39	-	-
	Ex5	c.1393-47C>T		28 (0.40)	rs11099600	0.20	0.49	0.5	0.52	0.49
	Ex5	c.1393-53T>C		1 (0.01)	rs141664955	0.007	0.01	0.01	-	-
	Ex5	c.1393-65G>A		5 (0.07)	rs116634000	0.03	0.04	0.04	-	-
	Ex6	c.1466-41A>G		44 (0.63)	rs6535473	0.31	0.49	0.50	0.52	0.50
	Ex6	c.1563+66G>A		10 (0.14)	rs6845316	0.07	0.11	-	-	-
	Ex7	c.1564-32C>T		6 (0.08)	rs114619599	0.04	0.02	0.03	0.02	0.03
	Ex7	c.1662+6A>G		1 (0.01)	rs56253838	0.007	-	0.002	-	0.0009
	Ex7	c.1662+15A>G		4 (0.05)	rs4693088	0.03	0.11		0.06	0.09

	Exon [a]	Nucleotide altered [b]	Amino acid altered [b]	Number (carrier frequency) [c]	dbSNP [d]	Minor allele frequencies				
						This study	1000 Genomes [e]	ESP6500 [f]	ExAC (Fin)[g]	ExAC (NFin)[g]
Intronic	Ex7	c.1662+15A>C		24 (0.34)	rs4693088	0.17	0.38	0.45	0.46	0.41
	Ex7	c.1662+15A>T		7 (0.10)	rs4693088	0.05	0.11	-	-	-
	Ex10	c.2190+44G>A		2 (0.03)	rs373644961	0.01	-	-	-	0.00001
	Ex11	c.2295+67A>G		39 (0.56)	rs12645412	0.28	0.38	-	-	-
	Ex12	c.2296-17A>G		1 (0.01)	rs59175583	0.007	-	0.002	-	0.001
	Ex16	c.3063+34T>G		4 (0.06)	rs115889524	0.03	0.02	0.023	0.008	0.02
	Ex18	c.3199-21C>T		5 (0.07)	rs62303752	0.03	0.04	0.03	0.01	0.04

[a] Ex (exon)

[b] HGVS nomenclature was used

[c] Carrier frequencies were calculated as the number mutation carriers (n) out of the total number tested (N = 69)

[d] All variants were cross-referenced with variant data available through the NCBI dbSNP database, build 38, release date 25.04.2013 (http://www.ncbi.nlm.nih.gov/projects/SNP/)

[e] 1000 Genomes browser, release date 16.10.2014 (http://browser.1000genomes.org).

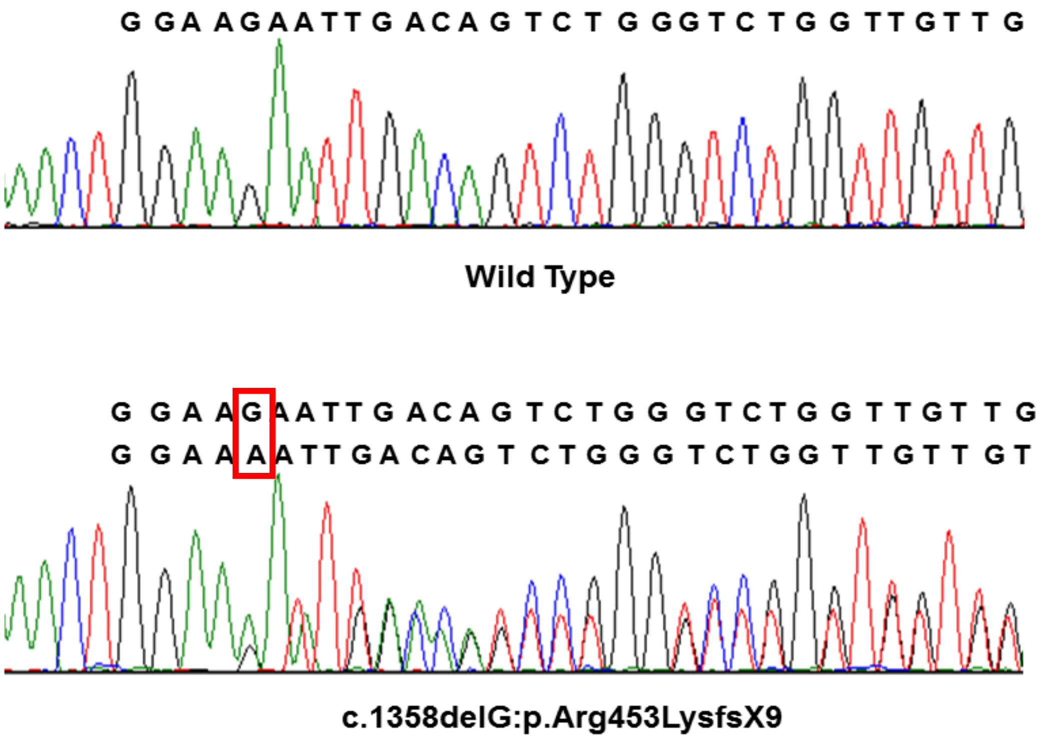
[f] NHLBI Exome sequencing project (ESP) 6500, release version v.0.0.30, date 03.01.2014 (http://evs.gs.washington.edu/EVS/)

[g] European Finnish (Fin) and non-Finnish (NFin) minor allele frequencies from the Exome Aggregation Consortium (ExAC), release version v.0.3.1 (http://exac.broadinstitute.org)

6.3.2. Truncating variant

The *HELQ* frameshift variant (c.1358delG, Figure 6.1) was discovered in only one individual of the high-risk BC/OVC families i.e. family 95. This variant allele was detected in a woman diagnosed with breast cancer at the age of 30 years (Annexure 2A). Unfortunately no further segregation analysis could not be done as her mother was deceased. DNA from a maternal aunt and cousin tested negative for the frameshift. The frameshift mutation (p.R453KfsX9) could be maternally or paternally inherited or may be a de novo event. As no further information regarding the paternal family history is available, the origin of this frameshift mutation is unknown.

This mutation may have a deleterious effect on the protein as it introduces a stop codon only nine residues downstream from an amino acid substitution at position 453. This will activate nonsense mediated decay that will have an impact on the gene expression levels. *HELQ* p.R453KfsX9 affects the expression of five isoforms including the major isoform 1 (Uhlén and Fagerberg 2015; Uhlén, et al. 2015).



Wild Type

c.1358delG:p.Arg453LysfsX9

Figure 6.1: Electropherogram of the frameshift *HELQ* variant c.1358delG.

Forward (sense) strand sequences of fragment 4B have been provided. Wild type and mutant sequences are depicted with double peaks indicating the heterozygous germline variant. The deletion at c.1358 (red box) results in a frameshift and causes a truncation of the polypeptide chain.

6.3.3. The potential significance of amino acid substitutions

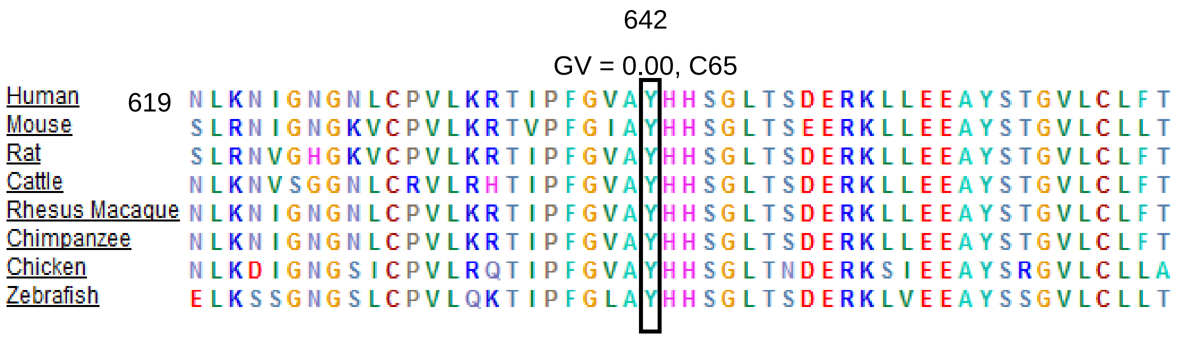
In silico analysis of the five missense variants was performed with SIFT, PolyPhen-2 and A-GVGD (Table 6.3). Two of the missense variants identified were localised in exon 2 (fragment 2A). One novel missense variant (ex9:c.1924T>C;p.Y642H) was detected in 1.4% (1/69) of the familial BC/OVC cases (0.007 allele frequency, Table 6.2). The novel p.Y642H variant detected in a patient (BRC 117-1) that was diagnosed with bilateral breast cancer at the age of 48 years. The patient's sister was also diagnosed with bilateral breast cancer at the age of 42 years. Unfortunately no DNA was available for the other affected members of this high-risk breast cancer family and no further segregation analysis could be done.

Of the five missense variants, only the p.Y642H novel mutation may be potentially significant, as supported by all three variant prediction tools. The non-conservative substitution of a polar tyrosine to a basic histidine results in the change of a tyrosyl side chain (Tyr) to an imidazole residue (His). SIFT and PolyPhen-2 analysis predicted that the missense alteration may have a potentially damaging effect on the protein. A-GVGD was used to calculate conservation scores from the multiple alignments submitted (i.e. GV) and quantified differences in the biochemical properties of wild-type and mutant amino acids (i.e. GD) based on these alignments (Mathe, et al. 2006). The T>C substitution at codon 642 fell outside the range of natural variation (GV: 0.00, GD: 83.33) observed at this position in the multiple sequence alignments. Both GV and GD scores indicated that the substitution was observed in an invariant genome location. This suggests that the p.Y642H variant in this position may be classified as a likely deleterious amino acid substitution. The multiple sequence alignments used to determine GV values are shown in Figure 6.2, which illustrates that codon 642 was conserved across all eight species. The scores from all three *in silico* tools indicated that the *HELQ* p.Y642H mutation may alter protein function.

One of the five missense variants discovered in the present study (ex2:c.916G>A) has previously been investigated as an allele with a potentially modifying effect on the risk conferred to *BRCA2* mutation carriers (Hamdi, et al. 2012). Despite this, *in silico* programs predicted that the variant would have no impact on protein structure or function (Table 6.3). The p.V306I variant was identified in 34% (21/61) of our BC families and was detected with an allele frequency of 0.23. This was close to prior reports where the common SNP was discovered at allele frequencies of 0.5 in individuals of European descent. A recent GWAS study by Hamdi et al, investigated the potential association of genetic variants on the 4q21 locus with BC risk and suggested that c.916G>A may affect transcription factor binding and be linked to the expression of certain target genes in breast cancer tissue (Hamdi, et al. 2016).

Table 6.3: *In silico* predictions of the effect of *HELQ* missense substitutions

Amino acid altered	*In silico* prediction tools		
	SIFT	**PolyPhen-2**	**Align-GVGD**
p.Lys16Arg	0.02 - Deleterious	0.141 - Benign	C25 - less likely
p.Leu235Pro	0.11 - Tolerated	0.002 - Benign	C0 - less likely
p.Val306Ile	1 - Tolerated	0 - Benign	C0 - less likely
p.Pro585Ser	0.03 - Deleterious	0.324 - Benign	Class C25 - less likely
p.Tyr642His	0.0 - Deleterious	1.00 - Probably damaging	Class C65 - most likely



```
                                              642
                                        GV = 0.00, C65
Human          619 N L K N I G N G N L C P V L K R T I P F G V A Y H H S G L T S D E R K L L E E A Y S T G V L C L F T
Mouse              S L R N I G N G K V C P V L K R T V P F G I A Y H H S G L T S E E R K L L E E A Y S T G V L C L L T
Rat                S L R N V G H G K V C P V L K R T I P F G V A Y H H S G L T S D E R K L L E E A Y S T G V L C L F T
Cattle             N L K N V S G G N L C R V L R H T I P F G V A Y H H S G L T S D E R K L L E E A Y S T G V L C L F T
Rhesus Macaque     N L K N I G N G N L C P V L K R T I P F G V A Y H H S G L T S D E R K L L E E A Y S T G V L C L F T
Chimpanzee         N L K N I G N G N L C P V L K R T I P F G V A Y H H S G L T S D E R K L L E E A Y S T G V L C L F T
Chicken            N L K D I G N G S I C P V L R Q T I P F G V A Y H H S G L T N D E R K S I E E A Y S R G V L C L L A
Zebrafish          E L K S S G N G S L C P V L Q K T I P F G L A Y H H S G L T S D E R K L V E E A Y S S G V L C L L T
```

Figure 6.2: Multiple sequence alignments of a conserved HELQ polypeptide region.

This window displays the HELQ amino acid sequence containing codon 642. Multiple sequence alignments illustrate that this residue is conserved across all eight species (GV=0). A-GVGD predicted that a histidine substitution in this residue would have a likely deleterious effect on this protein.

Three of the missense variants i.e. p.K16R, p.L235P, p.P585S have reported minor allele frequencies <0.01. It is the common perception that such rare amino acid substitutions could potentially exert some biological effect (MacArthur, et al. 2014). However, these changes were not predicted to be deleterious by all three *in silico* tools i.e. SIFT, PolyPhen-2 or A-GVGD. These three rare variants were initially classified as having no functional consequence as no evidence supporting their involvement in disease susceptibility was found. We then set out to explore the potential impact of intronic variants on splicing.

6.3.4. The impact of intron variants

The web-based *in silico* tool, human splice finder (HSFvs3.0), was used to determine the effect of intron *HELQ* sequence variants. We applied thresholds for the disruption of existing splice sites and auxiliary sequences of -5% (HSF) and -15% (MaxEntScan) variation between wild-type and mutant sequences. Predictions suggesting the creation of *de novo* splice acceptor or donor sites were documented if minimum increased levels of variation of at least +60% (HSF) and +200% (MaxEntScan) were observed. Applying these thresholds have previously proven to aid in the accurate prediction of splice site variants (Whiley, et al. 2011). The intronic variants predicted to affect splicing events are listed in Table 6.4. HELQ coding variants identified in the present study were assessed as well, however, none were predicted to significantly alter any splicing recognition sequences.

Table 6.4: Intronic *HELQ* sequence variants predicted to impact splicing events

	Ex	Nucleotide altered	Variant interpretation	HSF predictions		MaxEntScan	
				Mut score$^{\alpha}$	Variation (%)$^{\beta}$	Mut score$^{\alpha}$	Variation (%)$^{\beta}$
Canonical Splice region	Ex3	c.1191+1_1191 +2insGT	Activation cryptic donor site	72.33	+492.38	6.64	0
Intronic variants	Ex2	c.1013-12delT	Broken WT acceptor site	20.22	-78.16	-11.19	-203.32
	Ex4	c.1192-61_1192-38del24	Broken WT branch point	8.5	-87.69	-29.58	-5113.5
	Ex10	c.2190+44G>A	Activation cryptic acceptor site	82.94	+53.62	7.91	+19875

CV = consensus value
$^{\alpha}$ Mutant HSF or MaxEntScan score
$^{\beta}$ Variation between wild type and mutant score
- Variants have been grouped according to their location "Canonical Splice region" include variants located up to 2bp in the intron. "Intronic variants" include variants >2bp up/down stream intron-exon boundaries

Only four intronic variants were predicted as noteworthy splice signal variants and include the disruption of one wild-type acceptor site, the activation of two cryptic splice sites and the aberration of a wild-type branch point consensus sequence. The *HELQ* exon2 variant (c.1013-12delT) was identified in 6% (4/69) of index cases (Table 6.2) from three BC/OVC families. This mutation was found to significantly influence CV values (-78%, HSF) at the wild-type acceptor splice region. The *HELQ* c.1013-12delT variant, predicted to interrupt this WT acceptor site, will affect the reference "GTT" sequence at this intron region. This introduces a *de novo* acceptor site (GT) with increased output score variation of +262.06% for HSF and +186.27% for MaxEntScan (not shown here), compensating for the disruption of the wild-type site. A novel insertion *HELQ* c.1191+1_1191+2insGT mutation was predicted to generate a cryptic donor site. This variant occurred in a GT-rich position at the intron-exon boundary of exon3, as such, predictive scores are equal to cryptic donor splice sites contained in this genomic region. Therefore, despite *in silico* predictions, this insertion at the intron 3 donor site will have no impact on splicing. The one variant expected to potentially disrupt a wild-type branch point consensus sequence (*HELQ:* exon4c.1192-61_1192-38del24) was detected in 39% (27/69) of our study group from high-risk BC/OVC families (Table 6.2). This variant appears to be common amongst individuals of European descent (50%) which decreases the potential biological significance of the mutation.

The intron variant, *HELQ* c.2190+44G>A was identified in 3% (2/69) of the high-risk BC/OVC cases and is a very rare variant allele among individuals of European descent (similar to our study group). HSF predicted that this variant creates a *de novo* acceptor site with increased scores of 53% for HSF and 19875 for MaxEntScan. The HSF algorithm defines all strong sites as regions presenting with CV scores higher than 80 (Desmet, et al. 2009). As such, the G>A substitution may generate a strong cryptic acceptor site in the intron region of exon10 with a CV value (82.94) that is slightly greater than the corresponding wild-type acceptor site of exon11 (82.4). Correct assembly of the complex spliceosome machinery requires the recognition of consensus sequences that are vital for intron removal (Chen and Moore 2014). Detailed inspection of the *de novo* cryptic 3' splice site (3'ss) revealed that the nearest branch point heptamer sequence forms part of the wild type 5' splice site (5'ss). As such, this may be the less favourable acceptor site that will not be correctly recognised by the splicing apparatus.

Lastly, we evaluated a variant located in the 5' untranslated region c.-29_-27delACG, identified in 3% (2/61) of the BC/OVC families. The deletion was not located in a genomic region which showed great evidence of regulatory factor binding in the Regulome database. As such there was no evidence that it will have an impact on putative transcription element binding sites.

6.4. Conclusion

Recently, several research studies have indicated that *HELQ* is a candidate ovarian cancer gene that, when mutated, could confer an increased risk for disease (Luebben, et al. 2013; Nhung 2014; Takata, et al. 2013; Wang, et al. 2015b). This was motivated by the release of previous evidence which has verified that HELQ is a 3'–5' superfamily 2 helicase that plays a role in crosslinked DNA repair (Nhung 2014; Tafel, et al. 2011). As in the case of many DNA repair pathway genes, *HELQ* may potentially act as a gene of interest in several different cancers (Chun, et al. 2013). We considered *HELQ* as a candidate for screening and set out to determine whether mutations in this gene plays a role within South African breast and/or ovarian cancer families.

Recent studies have investigated the potential role of *HELQ* in breast cancer predisposition. As yet, research has only been conducted amongst Finnish and Canadian patients (Hamdi, et al. 2012; Pelttari, et al. 2015). However, these studies have only discovered putatively pathogenic missense and polymorphic mutations amongst high-risk BC/OVC families (Hamdi, et al. 2012; Pelttari, et al. 2015). The current study discovered 28 variants in high-risk BC/OVC families that included one frameshift (ex4: c.1358delG;p.R453KfsX9) and a putatively pathogenic missense mutation (ex9:c.1924T>C;p.Y642H) in two of the BC/OVC families. These two potentially pathogenic mutations may account for a carrier frequency of 2.8%. The frameshift mutation (p.R453KfsX9) is one of a total of three novel variants that were identified and it may result in nonsense mediate mRNA degradation. The biological impact of such a truncation variant may be similar to what was observed by Adelman and colleagues who examined HELQ deficient mice (Adelman, et al. 2013; Luebben, et al. 2013; Nhung 2014). These mice exhibited fertility defects, pronounced ovarian germ cell atrophy, ICL sensitivity and were predisposed to ovarian tumours or pituitary adenomas. This revealed that HELQ may have a tumour suppressive role due to the contributing part it plays in replication-coupled DNA repair (Adelman, et al. 2013; Luebben, et al. 2013).

The missense substitution (p.Y642H) was classified potentially pathogenic/deleterious by all three integrated *in silico* tools (SIFT, PolyPhen-2, Align-GVGD). *HELQ* c.1924T>C was the second novel mutation detected in this study. This variant may require further investigation, however, the lack of segregation analysis information may not support this. No DNA had been collected for retrospective study in this family and as such we could not perform additional analyses. In addition to this amino acid substitution, a putative cryptic splice site-activating mutation was found in intron10. *HELQ* c.2190+44G>A was detected in 3% (2/61) of the breast/ovarian cancer families at an allele frequency of 1%. Nevertheless, This potential cryptic acceptor site may not disrupt cleavage at the wild-type 3’ss due to its proximity to the 5’ss or wild-type donor site. None of the intronic variants that were identified significantly altered exonic splicing enhancer or silencer sequences. However, no guidelines are available to assess the significance of variant predictions in ESE/ESS sites. The poor accuracy of these predictions limits their clinical significance (Houdayer, et al. 2012). The pathogenicity of any of the missense, synonymous and potential splice-site variants of interest discovered in this study are ultimately unknown as we can only verify their true impact through functional studies.

This is the first study that has investigated the role of *HELQ* mutations in South African breast/ovarian cancer families. It is one of only three studies worldwide that has investigated the involvement of *HELQ* in hereditary breast/ovarian cancer. Thus far no overtly deleterious germline variants have previously been discovered in familial breast/ovarian cancer cases (Hamdi, et al. 2012; Pelttari, et al. 2015). While we could not fully validate the potentially pathogenic variants we detected within this study population, it does not categorically rule out *HELQ* as a potential gene of interest. Mutations that prevent the expression of a functional protein are characteristically rare (Lee, et al. 2014b). Such *HELQ* mutations could contribute to a small portion of the familial breast/ovarian cancer cases with unknown genetic risk factors that account for their disease. Analysing a larger set of patients may reveal the presence of very rare mutations which will be a worthwhile strategy to pursue in future.

Chapter 7:

Germline sequence variants in DNA repair genes of South African breast/ovarian cancer families with and without *BRCA* mutations

7.1. Introduction

Next generation sequencing has proven to be a good method for discovering novel genes that may predispose to breast/ovarian cancer (Hilbers, et al. 2013). The challenge still lies in mining the large amount of data in order to find mutations with functional consequences and prioritize them for validation (Li, et al. 2012). The search for variants of significance has proven to be more successful when primarily investigating candidate genes that are implicated in the process of DNA damage recognition and repair (Lu, et al. 2015). Defects in the complex DNA damage response processes account for the greatest amount of hereditary breast/ovarian cancer cases (Shuen and Foulkes 2011).

Previous studies that have evaluated pathogenic mutation signatures in DNA damage signalling and repair pathway genes have contributed towards understanding the genetic architecture that exists in many cancer types (Lu, et al. 2015). This has many implications for prevention and treatment (Alexandrov and Stratton 2014). However, the research done thus far has mainly focused on characterising mutational signatures in cancer-derived somatic data. Studies have shown that somatic mutations occur as a consequence of intrinsic mutational patterns that contribute to destabilising the DNA damage signalling and repair machinery (Alexandrov, et al. 2013a; Alexandrov, et al. 2013b; Alexandrov and Stratton 2014).

WES analysis conducted as part of the present study has investigated two DNA repair genes (*EME2* and *HELQ*) as potential candidate BC/OVC genes. This, however, has not explained the missing heritability amongst all the *BRCAx* families incorporated in the present study. It may be interesting to investigate what other, potentially less penetrant, germ-line variants our patient group may carry in genes associated with the highly integrated DNA damage signalling and repair pathway. The exome variant data of non-*BRCA1/2* index cases and carriers of truncating *BRCA1/2* mutations may be compared to examine any differences in the genetic variation amongst these patient groups. The entire sequence variation should be investigated to determine what germline variants can be found in the DNA repair genes of persons who are negative for *BRCA1/2* mutations versus patients that carry *BRCA* pathogenic mutations. This study set out to evaluate genetic variation in 516 DNA damage signalling and repair pathway genes carried by BC patients from high-risk families with and without *BRCA1/2* disease-causing mutations. Whole exome sequencing (WES) of two high-risk South-African breast cancer groups had been performed. One group was negative for mutations in *BRCA1/2* and the other was comprised of individuals carrying disease-causing *BRCA1/2* mutations.

DNA repair mechanisms are the most important means of defence against carcinogenesis (Alberg, et al. 2013). Analysing the contribution of rare and common germline variation in the genes of DNA repair processes may be a better approach of filtering and analysing data. This may facilitate the discovery of gene alterations that contribute towards unknown hereditary breast cancer risk.

7.2. Methods and Materials

7.2.1. Study patients

The analyses described in this chapter were performed with 12 high-risk BC patients from South African families with a significant history for BC/OVC (≥3 affected cases). This included the eight individuals from five families without pathogenic *BRCA* mutations. The remaining individuals, from four families, were positive for pathogenic *BRCA* mutations. One was the *BRCA1* mutation carrier and three additional individuals were included with *BRCA2* mutations.

7.2.2. Whole exome sequencing

Paired-end exome sequencing of *BRCA2* mutation-positive samples were sequenced by Omega Bio-services (Norcross, GA, USA). The Agilent Sureselect Human All exon version 5 (Santa Clara, CA, USA) sequence capture method was used to enrich for all known exome regions within the human genome at Omega. Sequencing was performed on the HiSeq 2000 genome analyzer IIx (Illumina, San Diego, CA, USA).

7.2.3. Sequence variants in DNA repair pathway genes

Exome variant data was obtained from high quality germline sequences of all 12 cases by using the variant analysis pipeline described in Chapter 2. A list of 516 genes was compiled of components from DNA damage repair pathways. This list consisted of DNA damage signalling and repair genes obtained through literature searches and from a recent study by Smith *et al*, 2016 (Annexure 5). VariantDB was used to analyse all variants identified in the DNA damage repair genes that had passed the quality assessment criteria (Vandeweyer, et al. 2014).

Variants were filtered to include all mutations with rare allele frequencies i.e. MAF ≤0.01 (1000genomes) and ≤0.03 (ESP6500) as well as low frequency alleles with MAF ≤0.05. It is plausible that the analyses of both rare and low frequency alleles may help explain some genetic variation that contributes to disease risk. Mutation classes such as protein truncation variants (PTV) were primarily focused on. This included frameshift, nonsense and splice-site (+/- 1 to 2bp) alterations. Non-synonymous variants were comprised of in-frame mutations and missense mutations with an inferred pathogenic effect according to the above-mentioned criteria. The computational predictions were made by the four web-based *in silico* tools that were incorporated during variant annotation.

7.3. Results and Discussion

7.3.1. Illumina exome sequencing coverage statistics and variant discovery

The whole exome sequence data quality of both Illumina sequencing runs at BGI and Omega bio -services were comparable (Table 7.1). Overall, an average of 37 million PE reads, 90bp in size were generated for all the samples with PHRED scaled base call quality scores ≥Q30. Final QC analysis results indicated that up to 99% base call accuracy was achieved in all forward and reverse paired-end sequence reads. The average depth for on-target sequence reads was 40X with 80% of the reads aligning to enriched target intervals. At the minimum sequence read depth of 10X more than 93% of the targeted regions were covered.

Sequences were filtered according to their quality and depth, resulting in the discovery of an average of 27 000 variants (25 857 SNV and 1423 insertions/deletions, Table 7.1) in the germline of each individual carrying *BRCA1/2* pathogenic variants. Of these, 19 098 were coding variants. This was similar to the results described in Chapter 3 where an average of 26 000 mutations (25 066 SNV and 1260 insertions/deletions) were discovered in high-risk *BRCAx* cases and these patients carried up to 18 690 coding variants.

Marginal differences in the number of variants identified in the high-risk BC patients may be expected as an updated version of the Agilent Sureselect Human All exon target enrichment kit (i.e. V5) was used for *BRCA2* R2645NfsX2 positive patients. The newer version of this kit is designed to allow for better coverage of the exome. Even though the *BRCA1* p.R1704X positive and *BRCAx* patients were analysed with version 4 of this capture method we did not observe a significant difference in the number of coding variants detected. The Agilent Sureselect platform has proven to be the most suitable means of enrichment for whole exome sequencing (WES) in comparison to other platforms (Meienberg, et al. 2015).

Table 7.1: Whole exome sequencing coverage and variant call results

	Average length of reads [a]	Total mapped reads	% Mapped reads [b]	% Sequence duplication levels [c]	% Reads on target [d]	% Target bases 10 fold Coverage [d]	% Target bases 20 fold Coverage [d]	% Target bases 40 fold Coverage [d]	Mean coverage for target bases [d]	Ti/Tv ratio[e]	Total variants (SNV: Indel)[e]
***BRCAx* patients**											
BRC9-3	87	36,928,598	99.4	20.2	78.0	92.4	78.2	42.6	39.7	2.814	25 732:1280
BRC73-1	88	36,661,032	99.5	24.1	81.7	92.6	79.3	45.4	41.5	2.901	25 582:1291
BRC92-2	88	36,735,443	99.4	24.2	80.6	92.3	78.8	44.9	41.1	2.865	22 267:1269
BRC92-3	88	38,875,739	99.3	25.8	77.5	92.5	79.3	45.8	41.7	2.840	25 243:1280
BRC94-1	87	37,397,412	99.4	23.4	80.9	92.6	79.4	45.4	41.4	2.898	25 178:1262
BRC94-2	88	36,340,432	99.4	18.6	79.1	92.3	78.1	42.7	39.7	2.836	25 213:1241
BRC95-2	87	37,477,048	99.4	16.8	76.0	92.3	77.6	40.9	38.6	2.836	25 741:1240
BRC95-3	87	36,039,619	99.5	23.2	83.2	92.8	79.3	45.0	41.3	2.837	25 578:1263
***BRCA1* pathogenic mutation-positive (√p.R1704X)**											
BRC71-1	87	39,513,267	99.2	20.8	73.2	92.5	78.6	43.6	40.3	2.814	25 106:1270
***BRCA2* pathogenic mutation-positive (√p.R2645NfsX2)**											
BRW112 [f]	100	34,199,877	98.7	29.9	76.5	94.5	81.4	48.0	42.0	2.850	25 918:1454
BRW181 [f]	100	34,996,360	99.0	23.3	75.9	94.9	82.4	44.0	42.0	2.874	25 925:1491
BRW248 [f]	100	35,103,844	98.8	20.4	70.8	94.5	80.0	42.0	39.0	2.851	26 481:1477

[a] Average length of reads after processing with FastX against bases with PHRED Scaled Quality scores <30

[b] Total mapped reads (assessment with the CLC Genomics workbench software v5) / total reads √100

[c] Sequence duplication levels : (FastQC analysis F + FastQC analysis R) / 2

[d] CLC Genomics workbench v5 mapping report

[e] According to VariantDB annotation

[f] BRW112, BRW181 and BRW248 was performed by Omega Bio-services. Coverage results were determined from the SureCall QC report √*BRCA1* (NM_007297), *BRCA2* (NM_000059)

As illustrated in Table 7.1, we identified on average 94% single base changes in the exome data of all the patients.

7.3.2. Quantifying sequence variants in 516 DNA damage signalling and repair pathway genes

The number of sequence variants carried by the study group was quantified in a list of 516 well-characterised and putative DNA repair response pathway genes. The list of genes was compiled from a recent study performed by Smith *et al.* and included genes that have been implicated in DNA repair, based on literature searches and the AmiGO browser (Annexure 5) (Smith, et al. 2016). Genes were divided into 12 pathway categories consisting of eight major mechanisms of DNA damage signalling and repair and seven additional groups that are functionally related to the removal of DNA lesions (Sehgal and Singh 2014). We then determined the number of rare truncating and putatively pathogenic missense variants in these genes for all index cases and the pathways they form part of.

Of the total sequence variants identified in DNA repair genes the average ratio of SNV to indels was 62 and 55 for *BRCAx* and *BRCA1/2* mutation-positive individuals respectively (Table 7.2). More SNVs were identified in the DNA repair genes of all index cases in comparison to insertion and deletion variants. Small insertion and deletion variants may have a more damaging effect on protein function, and could contribute to disease (MacArthur, et al. 2014). As such, this observation is expected. Deleterious variants will be rare (in less than 1% of the population) as a result of purifying selection (Lee, et al. 2014b).

Regardless of the presence or absence of a *BRCA* pathogenic variant each of the high-risk BC patients carried an approximate similar number of insertions/deletions in components involved in DNA damage recognition and repair (Table 7.2). *BRCA1/2* pathogenic mutation-positive individuals (except BRW248) carried an average number of only one to two truncating (frameshift & nonsense) variants which included their disease-causing *BRCA* mutation. *BRCAx* individuals possessed a range of one to four truncation mutations (Table 7.2). Of the total truncating variants, one-two were rare and low frequency variant alleles (i.e. ≤0.01 and ≤0.05 respectively) and approximately one-two were common variants in *BRCAx* and *BRCA1/2* truncation-positive individuals. Whilst the true functional effect of common variants remains uncertain, these common truncating mutations may also contribute to the inter-individual differences in DNA repair capacity (Nagel, et al. 2014; Panoutsopoulou, et al. 2013).

A range of 190 - 206 nonsynonymous (missense and in-frame indels) and 242 - 263 synonymous germline mutations were detected in *BRCAx* patients. The *BRCA1*/2 mutation-positive individuals carried 186 - 200 nonsynonymous and 240 - 270 synonymous mutations in the listed DNA repair genes (Table 7.2). The majority of germline nonsynonymous (missense and in-frame indels) variants identified in individuals with and without *BRCA1/2* mutations had common minor allele frequencies (>0.05) and approximately 6% were rare (<0.01) and low frequency (>0.01 & ≤0.05) variants. Studies have suggested that missense variants with a common and low allele frequency may influence the expression of pathway genes with a small effect, potentially contributing to the overall disease risk (Gibson 2012). However, the four *in silico* prediction tools incorporated did not consider any of the mutations as pathogenic. The nonsynonymous SNVs were classified as “variants of unknown significance”.

Table 7.2: The number of sequence variants detected in genes related to DNA damage signalling and repair pathways

	Total (SNV: Indel)	Missense	Synonymous	Stop-gain	Frameshift insertion/deletion	In-frame insertion/deletion	Intronic
***BRCA* truncation mutation negative**							
BRC9-3	449:10	203	245	1	3	3	62
BRC73-1	425:11	188	237	0	2	5	54
BRC92-2	431:5	188	242	0	0	4	64
BRC92-3	428:6	180	248	0	1	2	65
BRC94-1	427:4	198	229	0	0	1	48
BRC94-2	444:6	207	235	2	1	2	58
BRC95-2	449:8	186	263	0	1	4	59
BRC95-3	442:6	197	243	2	1	3	54
***BRCA1* mutation-positive (p.R1704X)**							
BRC71-1	439:8	190	247	2	1	5	58
***BRCA2* mutation-positive (p.R2645NfsX2)**							
BRW112	453:9	182	270	1	1	4	69
BRW181	459:9	198	260	1	1	6	71
BRW248	470:7	206	264	0	1	4	76

The number of variants in 516 DNA damage signalling and repair genes obtained from literature searches and a recent study by Smith *et al*, 2016 (Smith, et al. 2016). Rare and common exome sequence variants identified within germline DNA of each index case have been included. Mutational events in the eight *BRCA*-negative patients were compared to the *BRCA1*- and three *BRCA2*- truncation mutation carriers
√ The number of truncation variants for these individuals include their disease-causing *BRCA* mutations

7.3.3. Gene variants and their associated DNA repair processes

The majority of sequence variants detected in index cases were common to low frequency alleles but we found that the two patient groups carried many different types of mutations across different gene components of DNA repair. To evaluate the inter-individual differences between subjects, both common and rare gene variants were mapped according to their DNA repair process. Grouping gene variants in this manner also allowed us to distinguish the most likely genes of interest according to their molecular function from a background of genetic variation. Following visual inspection only 10 truncating mutations were identified in 10 DNA damage signalling and repair gene variants across all sequenced individuals (Table 7.3). A range of one - two truncating mutations was identified in each of the *BRCAx* patients (except BRC92-2 and BRC94-1). These genes are predominantly associated with DNA damage response processes such as; signalling DNA damage (DDS), interstrand cross-linked DNA repair through the fanconi anemia pathway (FA), homologous recombination (HRR), trans lesion synthesis (TLS), base excision repair (BER) and the modulation of nucleotide pools (NT Pools). This included both error free and prone mechanisms of repair (Dietlein and Reinhardt 2014). Two of the gene variants i.e. *RRM2B* c.211dupC and *UBE2NL* c.266T>G were identified in *BRCA1/2* positive cases as well (except for BRW248, Table 7.3 & 7.4). The allele frequencies of five deleterious mutations were greater than 0.05. While such variation may be noteworthy it cannot be investigated as candidate variants for breast/ovarian cancer susceptibility.

The major portion of germline variants in the current study group were amino acid substitutions. Even though the allele frequency of >90% of these substitution variants were common polymorphisms the two patient groups shared only 10% of the total missense and in-frame mutations identified. None of the nonsynonymous polymorphisms were considered putatively pathogenic. As such only gene variants with very low allele frequencies (≤0.005) were grouped according to their associated pathway (Annexure 5B). The study group carried a total of 105 nonsynonymous (missense and in-frame) mutations in 82 genes. Known and novel nonsynonymous variants identified in each individual were observed in 11/15 major DNA damage signalling and repair processes. This included mechanisms such as BER, NER, TLS, MMR, DDS, HRR, NHEJ, FA, EPN, PARP and DNAP. According to computational prediction these rare variants were not damaging to protein function and the genetic variation may be non-specific to disease. However, it should be noted that their true effect on protein function is depended on many additional factors and may not solely be based on *in silico* predictions (Chang and Wang 2012).

Only one of the *BRCAx* families (BRC94) carried two rare sequence variants i.e. *ATM* c.5417T>C:p.I1806T and *BLM* c.2480T>C:p.M827T in DNA damage signalling and homologous recombination repair genes (Table 7.4). Rare germline missense variants in *ATM* and *BLM* have been associated with low risk BC predisposition with an additive effect when in combination with other alleles (Lu, et al. 2015; Suhasini and Brosh Jr 2013). In order to make any meaningful conclusions from these results further investigations are needed. It may be interesting to perform segregation analysis of these alleles to determine the prevalence of these variants within other *BRCAx* families. Screening population matched healthy individuals from families that do not have a history of BC/OVC could also determine whether the sequence variation observed in this study was of any significance to their disease risk.

Table 7.3: Nonsense and frameshift variants in DNA damage repair associated genes of 12 index cases

Function[a]	**Gene**	**Variant[b]**	**Consequence[b]**	***BRCAx* cases n=8[c]**	***BRCA1/2* cases n=4[c]**	**1000G EUR[d]**	**ESP EA[e]**	**Functional annotation**
BER	*TDG*	c.287dupA	p.E96EfsX7	3		0.43	0.42	frameshift
HRR	*EME2*	c.964C>T	p.Q322X	1		0.016	0.015	stop-gain
	GEN1	c.2515_2519del	p.K839EfsX1	1		0.09	0.1	frameshift deletion
DDS	*HELB*	c.1566delG	p.E522DfsX4	1		0.023	0.026	frameshift deletion
TLS	*POLN*	c.2509delC	p.Q837SfsX7	1		0.002	0.0034	frameshift deletion
FA	*HELQ*	c.1358delG	p.R453KfsX9	1		-	0	frameshift deletion
NT Pools	*RRM2B*	c.211dupC	p.R71RfsX27	3	1 (*BRCA1*)	0.085	-	frameshift deletion
Other	*ASTE1*	c.1894dupA	p.R632KfsX10	1		-	-	frameshift insertion
	UBE2NL	c.266T>G	p.L89X	1	3 (*BRCA1/2*)	0.49	0.67	stop-gain
	CEP170	c.2081C>A	p.S694X	2		-	-	stop-gain

[a] BER= base excision repair, HRR= homologous recombination repair, DDS= DNA damage signalling, TLS= translesion synthesis, FA= fanconi anemia, NT pools= modulation of nucleotide pools, Other=putative DNA repair genes not classified as part of a specific mechanism. Mechanisms are listed as in the DNA repair genetic association studies database (DR-GAS) (Sehgal and Singh 2014)

[b] HGVS nomenclature was used

[c] Number of mutation-positive index cases

[d] All variants were cross referenced with variant data from European individuals made available through the 1000 Genomes browser, release date 16.10.2014 (http://browser.1000genomes.org). Minor allele frequencies have been provided

[e] Minor allele frequencies from European-American individuals was obtained from the NHLBI Exome sequencing project (ESP) 6500, release version v.0.0.30, date 03.01.2014 (http://evs.gs.washington.edu/EVS/)

Table 7.4: Rare (MAF <0.01) nonsynonymous variants and truncating mutations in DNA damage signalling and repair genes

Function	Gene	Variant (HGVS nomenclature)	*BRCAx*								*BRCA1*$^+$	*BRCA2*$^+$		
			9-3	73-1	92-2	92-3	94-1	94-2	95-2	95-3	71-1	112	181	248
SNV ≤0.01														
BER	*NEIL2*	c.22C>G:p.P8A	√											
	DNA2	c.2996G>A:p.R999H		√										
	NEIL1	c.421G>C:p.E141Q												√
	LIG3	c.2861G>A:p.R954H				√								
DDS	*ATR*	c.2776T>C:p.F926L												√
	UIMC1	c.1138T>C:p.S380P								√				
	ATM	c.1229T>C:p.V410A				√								
		c.5417T>C:p.I1806T					√	√						
	FOXM1	c.425G>A:p.G142E							√					
	TOP2A	c.3041C>T:p.T1014M						√						
DNAP	*POLG*	c.1550G>T:p.G517V					√							
		c.1174C>G:p.L392V									√			
EDP	*EXO1*	c.809A>T:p.D270V											√	
	SPO11	c.433A>G:p.R145G			√									
FA	*FANCC*	c.1081C>T:p.R361W							√					
		c.632C>G:p.P211R												√
	BRCA2	c.7828G>A:p.V2610M		√										
	FANCA	c.2859C>G:p.D953E											√	
HRR	*RAD50*	c.1153C>T:p.R385C					√							
	MCM9	c.3286A>G:p.M1096V										√		
		c.1915C>G:p.L639V			√									
		c.302C>T:p.S101L		√										
	NBN	c.643C>T:p.R215W								√				
	RAD54B	c.1343A>G:p.N448S											√	

Function	Gene	Variant (HGVS nomenclature)	*BRCAx*								*BRCA1*$^+$	*BRCA2*$^+$		
			9-3	73-1	92-2	92-3	94-1	94-2	95-2	95-3	71-1	112	181	248
SNV ≤0.01														
HRR	*MUS81*	c.896C>T:p.T299M		√										
	BLM	c.2480T>C:p.M827T					√	√						
	EME2	c.772C>T:p.P258S				√								
		c.804_806del:p.S268_269del											√	
	BRCA1	c.4394G>T:p.S1465I												√
	EME1	c.1636C>T:p.R546C							√					
MMR	*MSH6*	c.2827C>T:p.P943S											√	
		c.3335G>A:p.R1112H						√						
		c.3461C>T:p.T1154M							√					
	MLH1	c.376T>A:p.Y126N						√						
	MSH3	c.187C>G:p.P63A		√										
		c.2041C>T:p.P681S								√				
NER	*CHD1L*	c.1486G>A:p.G496R						√						
	ERCC3	c.2111C>T:p.S704L									√			
	ERCC4	c.1563C>G:p.S521R												√
	ERCC2	c.1187G>A:p.S396N		√										
NHEJ	*WRN*	c.1066A>G:p.K356E				√								
PARP	*PARP1*	c.1148C>A:p.S383Y	√											
		c.450G>T:p.Q150H							√					
TLS	*POLQ*	c.7393G>A:p.E2465K												√
		c.4635C>A:p.H1545Q			√									
		c.673C>T:p.H225Y			√									
	POLN	c.270G>T:p.Q90H			√									
	POLK	c.85G>A:p.E29K	√											

Function	**Gene**	**Variant (HGVS nomenclature)**	***BRCAx***								***BRCA1***$^{+}$	***BRCA2***$^{+}$		
			9-3	**73-1**	**92-2**	**92-3**	**94-1**	**94-2**	**95-2**	**95-3**	**71-1**	**112**	**181**	**248**
Truncating mutations														
BER	*TDG*	c.287dupA:p.E96EfsX7	√	√						√	√			
HRR	*EME2*	c.964C>T:p.Q322X	√											
	GEN1	c.2515_2519del:p.K839EfsX1		√										
DDS	*HELB*	c.1566delG:p.E522DfsX4						√						
TLS	*POLN*	c.2509delC:p.Q837SfsX7	√											
FA	*HELQ*	c.1358delG:p.R453KfsX9							√					
NT Pools	*RRM2B*	c.211dupC:p.R71RfsX27	√			√					√			

Putatively pathogenic nonsynonymous variants consist of missense and in-frame mutations
Genes not associated with a specific mechanism have been removed from this table
The disease-causing variants of *BRCA1/2* PTV positive patients have not been included in these listed variants

7.3.4. Variants of potential relevance to BC susceptibility

The four rare truncating mutations that were discovered in four of the *BRCAx* cases (three families) and were comprised of two novel and three known mutations that have reported minor allele frequencies ≤0.01 (Table 7.3). These truncation variants were in *EME2* (p.Q322X, rs61753375), *HELQ* (p.R453KfsX9), *POLN* (p.Q837SfsX7, rs3833632) and *CEP170* (p.S694X, rs199575184). This supports the findings that were made during our discovery of family-specific variants in Chapter 3. In addition to these mutations variant analysis highlighted ASTE1 as it is implicated in DNA repair. However, frameshift mutations in this receptor protein are associated with a high density of tumour-infiltrating lymphocytes (TIL) in colon cancers (Goyal, et al. 2016). This promotes a cytotoxic anti-tumour response and is related to controlling colon cancer progression (Maby, et al. 2015).

As previously noted, HELQ, EME2 and POLN play unique roles in pathways of DNA damage signalling and repair such as interstrand crosslinked DNA repair, processing stalled replication forks and the extension of intermediate D-loop structures during TLS and HRR (Pepe and West 2014; Takata, et al. 2013; Takata, et al. 2015). Rare variant alleles may have an impact on protein function which could affect the efficient repair of DNA lesions at different interfaces or contribute to the genomic instability within each index case. The observations made during the current analyses have reiterated the potential involvement of *EME2* and *HELQ* gene variants in hereditary BC (Lee, et al. 2014b; MacArthur, et al. 2014). *POLN* may be the following candidate gene worth investigating as rare putatively disruptive mutations could also contribute to the currently unexplained familial breast cancer risk.

In this study, both truncating and nonsynonymous variants were analysed in DNA damage signalling and repair genes. The observations that were made are similar to previous reports made by studies that investigated the link between a significant decrease in DNA repair capacity and breast cancer development (Matta, et al. 2012). Mutations in the DNA repair processes are frequently observed in hereditary breast cancer and is the driver of tumorigenesis (Kobayashi, et al. 2013; Le Guen, et al. 2014; Liu, et al. 2014; Patrono, et al. 2014; Santarpia, et al. 2013). Approximately half of the mutations identified in our cases were common, with allele frequencies that are representative of polymorphic variants. The functional significance of DNA repair polymorphisms is still largely unknown. Previous studies have suggested that such polymorphisms may contribute towards decreased DNA repair capacity, which may be plausible (Patrono, et al. 2014). However, this could be investigated through further functional studies. Genome instability is induced by the presence of unrepaired DNA lesions (Barlow, et al. 2013) and any loss or decrease in the function of products geared towards the detection, stabilisation and removal of nucleotide damage, strand breaks, inter- or intra-stand crosslinks may ultimately contribute to the development of cancer (Amemiya, et al. 2015; Johnson, et al. 2007; Le Guen, et al. 2014; Nicolay, et al. 2012). Mutations, shown through functional analysis, to disrupt the expression of proteins involved in DNA repair processes may be candidates for BC susceptibility. Collectively “common” changes potentially lead to suboptimal repair mechanisms rather than being the sole cause of cancer development (Brandt-Rauf, et al. 2013; Matta, et al. 2012). Future studies are needed to completely validate the role of rare truncation variants in *EME2, HELQ* and *POLN*, unique to two of our high-risk *BRCAx* families, in hereditary breast cancer. Despite these rare variants; *BRCA* negative index cases shared similar levels of sequence variation with the *BRCA1* and *BRCA2* mutation-positive individuals in genomic regions that code for DNA repair proteins.

7.4. Conclusion

Whole exome sequence analysis provided detailed information on the germline variants carried by high-risk familial breast cancer cases. The results presented here may represent some of the genetic variation carried by index cases from breast/ovarian cancer families. We observed similar sequence variation between the group negative for truncating *BRCA* mutations*,* the *BRCA1* p.R1704X carrier and those that were positive for a *BRCA2* premature truncation variant. Marginal differences were observed in DNA damage signalling and repair pathway gene variants carried by the high-risk breast cancer patients we investigated. More than 90% of the mutations that were detected were "common". According to the evolutionary theory rare deleterious alterations (i.e. nonsense & frameshift) will likely be rare (allele frequency 0.01) and that truncating variants, which have a severe impact on protein function will be even more rare (<0.01) (Lee, et al. 2014b). We found that *BRCA1/2* mutation-positive patients possessed more indels in genomic regions that contain DNA repair gene sequences and this may suggest that these individuals carry a higher burden of mutations in the components of this pathway. However, we identified more sequence variation in the coding regions (exons) of DNA repair genes in *BRCAx* patients.

Germline variants were then categorised according to 15 key DNA damage response processes. Deleterious mutations and rare missense variants were identified in genes associated with 12 DNA damage recognition and repair mechanisms including well characterised mechanisms such as HRR, DDS, NHEJ, MMR, TLS etc. Homologous recombination repair is the main cellular mechanism that repairs double strand breaks. HRR is dependent on the exchange of a homologous DNA/sister chromosome segment and is an error-free pathway (Dietlein and Reinhardt 2014). Therefore, germline truncation mutations in genes associated with this and other mechanisms that contribute to cellular tolerance and repair of DNA lesions i.e. NHEJ, DDS, TLS & MMR, may affect the genetic stability of the cell (Goodman and Woodgate 2013; Grabarz, et al. 2012; Nicolay, et al. 2012; Poulogiannis, et al. 2010). The observed mutations in these pathway genes may explain the manner in which germline mutations (in *BRCAx* patients) contribute towards breast cancer development. Defects in these processes are of great significance as pathogenic alterations in their genes confer risk for disease, are targeted in therapeutic strategies (e.g. DNA-PKcs inhibition) and are drivers of tumorigenesis (Dietlein and Reinhardt 2014; Le Guen, et al. 2014).

Unlike protein truncating genetic variants, it is less clear what the effect of missense mutations in relevant DNA repair pathways are (Brandt-Rauf, et al. 2013; Gibson 2012). Our analysis of missense variants were restricted to those changes predicted as deleterious by all four *in silico* tools. As we did not make use of any additional technologies to supplement the exome sequence results (i.e. segregation analysis/Sanger sequence validation) strict inclusion criteria were applied for interpreting amino acid substitution variants (Foley, et al. 2015). None of the nonsynonymous variants (missense and in-frame) were found to be putatively pathogenic according to these criteria.

Observations made in the current study are similar to previous pathway-based analyses in somatic cancer studies. These prior investigations have revealed that driver mutations are mainly found in DNA repair genes as they are vital to maintaining genomic stability during DNA replication (Alexandrov, et al. 2013a; Alexandrov, et al. 2013b; Alexandrov and Stratton 2014). The analyses performed in this study are based on germline variant data and reports on the genetic variation carried by high-risk hereditary breast cancer cases. Overall, we observed that hereditary breast cancer cases have genetically heterogeneous DNA repair gene sequences. The extent of their genetic variation could account for the high number of affected cases in their families and may contribute to their elevated disease risk. The analysis of 516 DNA damage signalling and repair pathway genes has shown that truncating mutations in *EME2, HELQ* and *POLN* genes may be involved in breast cancer susceptibility. Studies that have investigated the patterns of variants in breast cancers have discovered variants of interest in *HELQ,* whereas further studies are still needed for *EME2* and *POLN* (Lhota, et al. 2016). As yet, we cannot confirm that the germline variants identified in our study account for the unknown familial BC cases. More evidence may be needed to support the significance of such rare truncating variants in DNA damage repair pathway genes and their contribution to carcinogenesis. At this stage we cannot exclude the role of these uncharacterised, potentially pathogenic, truncation variants either as they may contribute towards subtle defects in protein expression, increased cancer risk or different responses to therapeutic treatment strategies. However, future validation, co-segregation studies and functional analysis of mutations is needed to support the observations made.

Chapter 8:

Conclusion

8.1. Whole exome sequencing of high-risk BC/OVC families

Comprehensive testing with the use of next generation sequencing has led to the discovery of many recently characterised BC predisposing genes. However, the search for novel components that account for the large portion of unexplained familial BC cases is still on-going (Snape, et al. 2012). The missing heritability of breast cancer has brought on the widespread search for a potential "*BRCA3*" gene that when mutated, confers higher risk for BC to carriers than the general population lifetime risk (Boyd 2014). This as-yet unknown gene could likely be a tumour suppressor gene and may carry mutations that are even more rare than variants in *BRCA1/2* (Boyd 2014; Easton 1999). Almost all of the BC susceptibility genes to date are related to fundamental processes such as cell proliferation and maintaining DNA fidelity which are collectively referred to as the FA/BRCA pathways (Vuorela, et al. 2011).

The primary aim of this study was to discover rare pathogenic germline variants in SA patients from families with a strong history for breast/ovarian cancer that do not carry mutations in *BRCA1/2* or any other known BC gene (i.e. *BRCAx*). An adapted GATK variant filtration and prioritisation pipeline was used to analyse whole exome sequence data and identify germline variants from patients from six high-risk *BRCAx* BC/OVC families. Mutant alleles in BC genes that act as high-risk disease markers have been identified by focusing on familial groups. Such patients may be enriched with damaging germline mutations that confer a significant increase in their lifetime risk for BC, higher than the general population (Wen, et al. 2014). This study made use of a family-specific approach to discover novel and known potentially pathogenic mutations with minor allele frequencies <1%. Mutations that were computationally predicted to affect protein function and those that may play an important role in breast cancer were prioritised. Four gene variants were considered the most potential candidates for validation. After sequence verification three high-priority genes (*TCHP, EME2, HELQ*) were screened further to identify their potential significance in breast cancer predisposition. These promising potential candidate genes coded for a novel putative tumour suppressor (*TCHP*) that triggers pro-apoptotic pathways (i.e. evoking mitophagy), an endonuclease (*EME2*) that resolves stalled replication forks and a helicase protein (*HELQ*) which promotes crosslinked DNA repair (Neill, et al. 2014; Pepe and West 2014; Takata, et al. 2013).

TCHP has been evaluated as a key regulator of tumour cell mitophagy and angiostacis (Neill, et al. 2014). Mutations in proteins that play a role in promoting apoptosis may also be linked to cancer initiation (Delbridge, et al. 2012). Two potentially deleterious, rare *TCHP* variants (p.W449X and p.E461del) were discovered in only 3.3% of all the families that formed part of the present study. Due to the incomplete segregation of both variants in exon12, *TCHP* was not investigated further. The results obtained in this study could not prove that *TCHP* gene variants play a role in the disease susceptibility of the chosen patient cohort. Neither one of the *TCHP* variants (p.W449X and p.E461del) was detected in a significant number of related BC/OVC patients. For either of these mutations to be considered, the variants of interest should have segregated amongst a significant number of closely related patients (Feng, et al. 2015). The cellular role of this putative tumour suppressor protein still suggests that loss-of-function variants in its coding sequence may contribute to many other types of cancer (Kim, et al. 2010; Vecchione, et al. 2008).

EME2 was also investigated as a potential candidate breast cancer gene. Through whole gene screening two truncating and two in-frame deletion mutations were identified. Together these variants may account for 11.4% of the *BRCAx* families and 11.7% of the *BRCA1/2*-negative BC/OVC cohort. However, a significant association between the nonsense (Ex7:p.Q322X) and frameshift (Ex1:p.G55VfsX19) mutations and breast cancer was not found. Both variants appeared to be polymorphisms in the small group of unaffected individuals. Two of the four potentially damaging mutations, were in-frame deletions and may still be of some functional significance. This includes an exon2:p.E98_L100del mutation that was detected in 1.4% and exon7:p.Ser269del found in ~1% of the breast/ovarian cancer patients. These variants are located in alpha helix and coiled-coiled regions of EME2 and may be damaging to protein function. While the low carrier frequency of these potentially pathogenic variants may indicate otherwise, *EME2* may still warrant further investigation.

HELQ was also investigated as a potential novel candidate gene and the deleterious variant allele (i.e. ex4:c.1358delG) was found in 1.6% of the *BRCAx* families (1.4% cases). *HELQ* p.R453KfsX9 is a novel variant that may be deleterious to protein function which explains the extremely low/rare carrier frequencies observed (Lee, et al. 2014b). The Cancer Research UK team have recently recognised that pathogenic variants in helicase Q (aka HEL-308) is associated with an increased chance of developing a rare form of ovarian cancer (Adelman, et al. 2013). These observations motivated the study of germline variants in this gene and the potential role they may play in BC susceptibility. Thus far only two other studies have investigated the role of *HELQ* mutations in familial breast/ovarian cancer families but have discovered only putatively deleterious missense variants (Hamdi, et al. 2012; Pelttari, et al. 2015). This current study also detected a novel putatively pathogenic missense variant (p.Y642H) in one of the BC families. Similar to previous studies our study cannot rule out the potential that rare risk variants in *HELQ* may account for a portion of unknown breast/ovarian cancer cases.

The exome data of our study population was then evaluated further. Exome variant data from patients carrying pathogenic mutations in *BRCA2* were incorporated as well. The objective of this section of the current project was to make use of a candidate gene approach and discover variants in DNA signalling and repair genes as potential susceptibility alleles. The study population then included patients from hereditary breast cancer families that were positive for pathogenic mutations in *BRCA1/2* genes. The level of sequence variation carried by *BRCAx* individuals was compared to mutations found in *BRCA1* and *BRCA2* mutation-positive patients.

Sequence variants in a list of 516 well-characterised and putative DNA repair response pathway genes were analysed. Overall, 10 potential loss of function and 105 rare missense variants were identified, spanning ~90 of the 516 clinically relevant DNA repair pathway component genes. The majority of germline mutations of index cases were found in pathway components of both error-free and error-prone DNA damage recognition and repair pathways. This study did not have enough evidence to suggest that any of the amino acid substitution mutations were genetic factors which may play a collaborative role towards driving high-penetrant familial breast cancer development. Three rare truncating variants were highlighted as variants of interest in four *BRCAx* cases (two families). This included *EME2* and *HELQ* which reaffirmed the reason for screening these candidates as part of the present study. Our analyses also highlighted *POLN* as a potential gene of interest. This gene may be screened in future to determine the prevalence of *POLN* mutations in familial breast cancer cases. The genes code for proteins that play vital roles in their associated DNA repair processes and contain rare mutations that are potentially damaging to protein function.

The majority of familial breast/ovarian cancer is accounted for by rare, high penetrant variant alleles in proteins associated with DNA repair and genome maintenance pathways (Boyd 2014). Various studies have indicated that these processes contain the most plausible candidates for familial BC risk which was suggested by the current study as well. Natural selection keeps the frequency of variants in these genes low, supporting the hypothesis that rare gene variants will have a moderate to high impact on disease susceptibility (Vuorela, et al. 2011). Such as-yet unidentified low frequency, high-risk mutant alleles may only account for a small fraction of *BRCAx* families (Mavaddat, et al. 2010a). Therefore, it may be difficult to detect these variants and challenging to validate their pathogenicity (Vuorela, et al. 2011).

The discovery of variants in new susceptibility genes may account for the remaining 40% of unknown familial breast cancer cases. The existence of a *BRCA3* gene may not be the answer to the missing heritability seen worldwide. This may be accounted for by variants that are population- or family-specific due to founder effects (Bodmer and Tomlinson 2010).

8.2. Challenges in elucidating the missing heritability of familial breast cancer

The discovery of breast cancer predisposition genes through next generation techniques such as whole exome sequencing has proven to be challenging. To date 12 studies have attempted to identify high- and moderate penetrant variants in novel BC susceptibility genes with limited success (Cybulski, et al. 2015; Gracia-Aznarez, et al. 2013; Hilbers, et al. 2013; Kiiski, et al. 2014; Lynch, et al. 2013; Noh, et al. 2015; Park, et al. 2011; Park, et al. 2012; Park, et al. 2014; Snape, et al. 2012; Thompson, et al. 2012; Wen, et al. 2014). These studies considered one of two main strategies in their study design which included WES of multiple BC-affected family members (i.e. family-based approach) or a cohort of unrelated patients diagnosed with BC at an early age (≤45years) (Chandler, et al. 2016). The current study incorporated the first approach and identified variants in genes that are potential candidates in hereditary breast cancer.

To improve the chances of capturing variants in candidate genes, more than one method for variant discovery should be used (DePristo, et al. 2011; Pabinger, et al. 2013). Our study could not identify copy number- (CNV) and complex structural variants (SV). Computational tools that call such variants have performed best when using whole genome sequence (WGS) data in comparison to WES. The high genome coverage achieved with whole genome sequencing is the most optimal means of detecting complex structural variation (Spurrell 2013). This includes variants that extend beyond the size of an input read such as large deletions/insertions, inversions, translocations, tandem duplications (Moncunill, et al. 2014). Sensitive and specific discovery of CNV from WES can be very difficult as it is dependent on uniform read depths and accurate prediction of breakpoints from the targeted sequence (Krumm, et al. 2012; Tan, et al. 2014). Several CNV call tools have recently been developed and the improvement of such methods may help better the analysis of WES data (Tan, et al. 2014).

Identifying genes of interest from a large set of variant data has become a challenging task that has been termed “the prioritization problem” (Tranchevent, et al. 2011). It is at this stage of the analyses where various issues arise that may prevent or hamper the discovery of possible candidate genes (Bodmer and Tomlinson 2010). Several strategies were used in order to empower our analysis overall. Index cases were carefully selected to enrich for high-risk *BRCAx* families and improve the power of variant analysis (Feng, et al. 2015). In addition to stringent quality filtration settings, selecting rare protein truncation variants (frameshift, in-frame, stopgain, and canonical splice donor/acceptor) was the primary focus. However, previous studies have shown that identifying PTV's is not enough to prove causality (Snape, et al. 2012). Many *BRCA1/2* missense mutations have been described as pathogenic and are linked to breast cancer predisposition (Machackova, et al. 2001). This study did not identify missense mutations that were predicted to affect protein function by the four *in silico* tools incorporated during variant prioritisation. Making use of less stringent inclusion criteria may be a worthwhile future strategy to pursue.

Overall, whole exome sequencing has resulted in the discovery of a modest amount of novel BC susceptibility genes in studies conducted to date (Chandler, et al. 2016). Similar to some of these research efforts, the WES variants identified in our study were not significantly associated with breast/ovarian cancer susceptibility. However, the family-based approach has still proven to be the more successful study design to follow with approximately 10% of prior studies discovering novel BC-susceptibility genes (Park, et al. 2012; Park, et al. 2014). Further analysis in larger sample sets is needed to further assess the promising candidate genes identified in the present study (Feng, et al. 2011).

8.3. Concluding statements and future directions

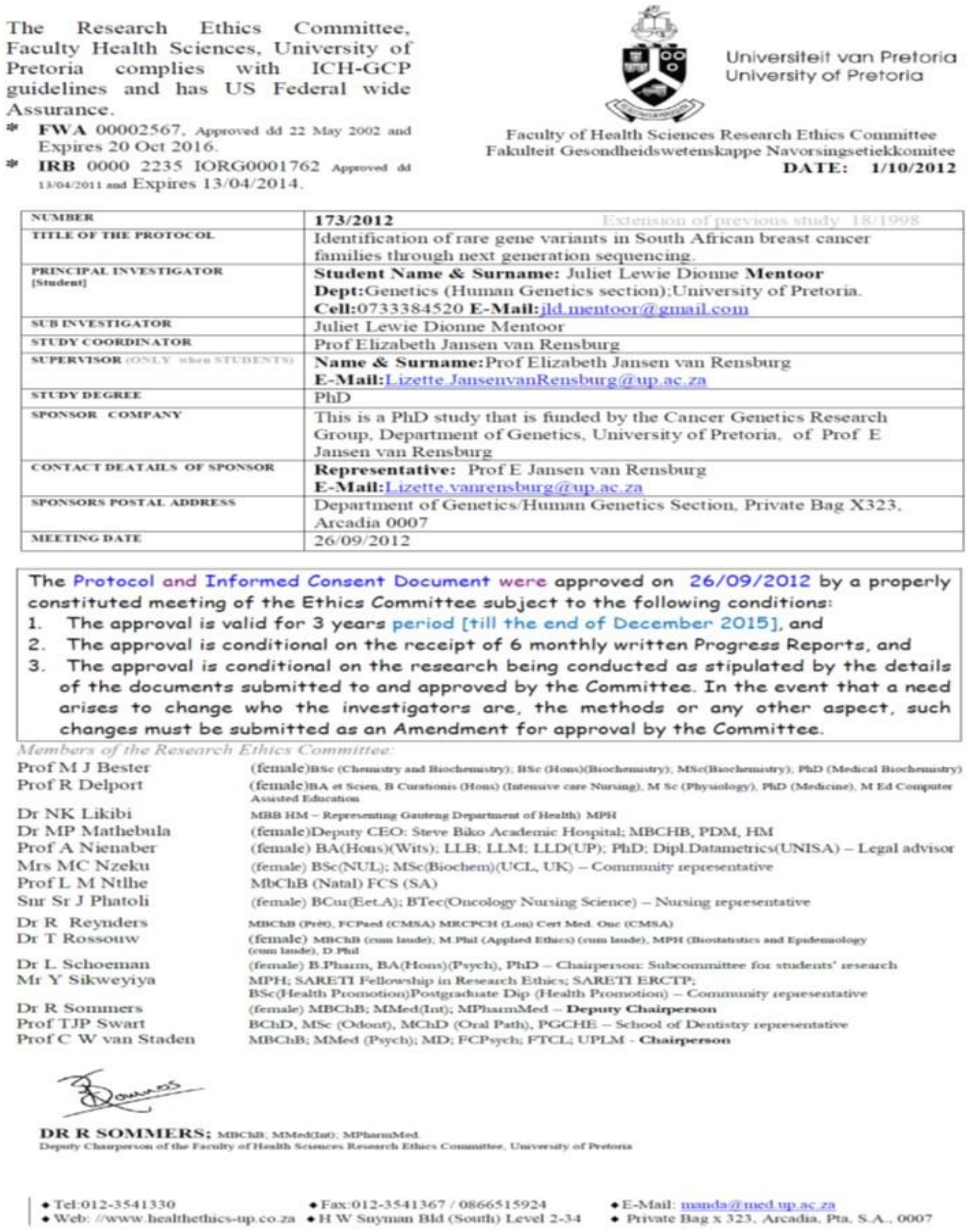
Regardless of some of the challenges this study faced we successfully obtained germline variant data from the exomes of high-risk familial BC/OVC cases. Patients were selected from families that are primarily of European descent and form part of the Afrikaner population group in South Africa. This sub-population likely have less genetic heterogeneity and have families that display founder effects similar to the Ashkenazi Jewish population (Easton 1999; van der Merwe, et al. 2012). Once detected, variants must be prioritised based on allele frequency and the predicted effect on protein function. As such, this study may have contributed towards improving bioinformatics pipelines that are implemented in the search for rare disease causing variants. However, it may be more useful to apply multiple variant interpretation and prioritisation strategies. This could increase the chance of identifying rare variants that are significantly more common in cases versus healthy controls and is associated with hereditary breast cancer (Bodmer and Tomlinson 2010). In future, gene candidates may also be stratified according to their expression in breast and ovarian tissues (MacArthur, et al. 2014). The quality of experimental protein expression data will impact the success of this prioritisation strategy (Nitsch, et al. 2010).

Since the introduction of exome sequencing more candidate genes have been discovered that may contribute to some unexplained hereditary breast cancer cases when mutated (Gracia-Aznarez, et al. 2013; Quintáns, et al. 2014). None of the variants that were investigated in the present study showed a conclusive association to hereditary breast/ovarian cancer cases in South Africa. *EME2* and *HELQ* gene variants could potentially be added to gene capture panels to increase the likelihood of discovering additional pathogenic variants. The analysis of such rare variants can prove to be especially challenging and requires the inclusion of large patient cohorts (Lee, et al. 2014b). Re-sequencing such genes in thousands of cases will help validate the clinical significance of the mutations that were discovered. This may be a massive undertaking that has motivated the formation of collaborations between research facilities that are attempting to achieve a similar aim (Complexo, et al. 2013). Future collaborations may help validate the association of *HELQ* and *EME2* gene variants with BC/OVC susceptibility.

The results generated from our research have been presented at two national and three international conferences. A manuscript has been submitted to the journal: *Familial Cancer* (Annexure 7).

Annexure 1:

Ethics clearance certificate for the study.



The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 20 Oct 2016.
- IRB 0000 2235 IORG0001762 Approved dd 13/04/2011 and Expires 13/04/2014.

Universiteit van Pretoria
University of Pretoria

Faculty of Health Sciences Research Ethics Committee
Fakulteit Gesondheidswetenskappe Navorsingsetiekkomitee

DATE: 1/10/2012

NUMBER	**173/2012** Extension of previous study 18/1998
TITLE OF THE PROTOCOL	Identification of rare gene variants in South African breast cancer families through next generation sequencing.
PRINCIPAL INVESTIGATOR [Student]	**Student Name & Surname:** Juliet Lewie Dionne **Mentoor** **Dept:**Genetics (Human Genetics section);University of Pretoria. **Cell:**0733384520 **E-Mail:**jld.mentoor@gmail.com
SUB INVESTIGATOR	Juliet Lewie Dionne Mentoor
STUDY COORDINATOR	Prof Elizabeth Jansen van Rensburg
SUPERVISOR (ONLY when STUDENTS)	**Name & Surname:**Prof Elizabeth Jansen van Rensburg **E-Mail:**Lizette.JansenvanRensburg@up.ac.za
STUDY DEGREE	PhD
SPONSOR COMPANY	This is a PhD study that is funded by the Cancer Genetics Research Group, Department of Genetics, University of Pretoria, of Prof E Jansen van Rensburg
CONTACT DEATAILS OF SPONSOR	**Representative:** Prof E Jansen van Rensburg **E-Mail:**Lizette.vanrensburg@up.ac.za
SPONSORS POSTAL ADDRESS	Department of Genetics/Human Genetics Section, Private Bag X323, Arcadia 0007
MEETING DATE	26/09/2012

The Protocol and Informed Consent Document were approved on 26/09/2012 by a properly constituted meeting of the Ethics Committee subject to the following conditions:

1. The approval is valid for 3 years period [till the end of December 2015], and
2. The approval is conditional on the receipt of 6 monthly written Progress Reports, and
3. The approval is conditional on the research being conducted as stipulated by the details of the documents submitted to and approved by the Committee. In the event that a need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

Members of the Research Ethics Committee:

Prof M J Bester	(female)BSc (Chemistry and Biochemistry); BSc (Hons)(Biochemistry); MSc(Biochemistry); PhD (Medical Biochemistry)
Prof R Delport	(female)BA et Scien, B Curationis (Hons) (Intensive care Nursing), M Sc (Physiology), PhD (Medicine), M Ed Computer Assisted Education
Dr NK Likibi	MBB HM – Representing Gauteng Department of Health) MPH
Dr MP Mathebula	(female)Deputy CEO: Steve Biko Academic Hospital; MBCHB, PDM, HM
Prof A Nienaber	(female) BA(Hons)(Wits); LLB; LLM; LLD(UP); PhD; Dipl.Datametrics(UNISA) – Legal advisor
Mrs MC Nzeku	(female) BSc(NUL); MSc(Biochem)(UCL, UK) – Community representative
Prof L M Ntlhe	MbChB (Natal) FCS (SA)
Snr Sr J Phatoli	(female) BCur(Eet.A); BTec(Oncology Nursing Science) – Nursing representative
Dr R Reynders	MBChB (Prêt), FCPaed (CMSA) MRCPCH (Lon) Cert Med. Onc (CMSA)
Dr T Rossouw	(female) MBChB (cum laude); M.Phil (Applied Ethics) (cum laude), MPH (Biostatistics and Epidemiology (cum laude), D.Phil
Dr L Schoeman	(female) B.Pharm, BA(Hons)(Psych), PhD – Chairperson: Subcommittee for students' research
Mr Y Sikweyiya	MPH; SARETI Fellowship in Research Ethics; SARETI ERCTP; BSc(Health Promotion)Postgraduate Dip (Health Promotion) – Community representative
Dr R Sommers	(female) MBChB; MMed(Int); MPharmMed – **Deputy Chairperson**
Prof TJP Swart	BChD, MSc (Odont), MChD (Oral Path), PGCHE – School of Dentistry representative
Prof C W van Staden	MBChB; MMed (Psych); MD; FCPsych; FTCL; UPLM - **Chairperson**

DR R SOMMERS; MBChB; MMed(Int); MPharmMed.
Deputy Chairperson of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria

- Tel:012-3541330
- Web: //www.healthethics-up.co.za
- Fax:012-3541367 / 0866515924
- H W Snyman Bld (South) Level 2-34
- E-Mail: manda@med.up.ac.za
- Private Bag x 323, Arcadia, Pta, S.A., 0007

Annexure 2:

(A) - Pedigrees of cases selected for whole exome sequencing

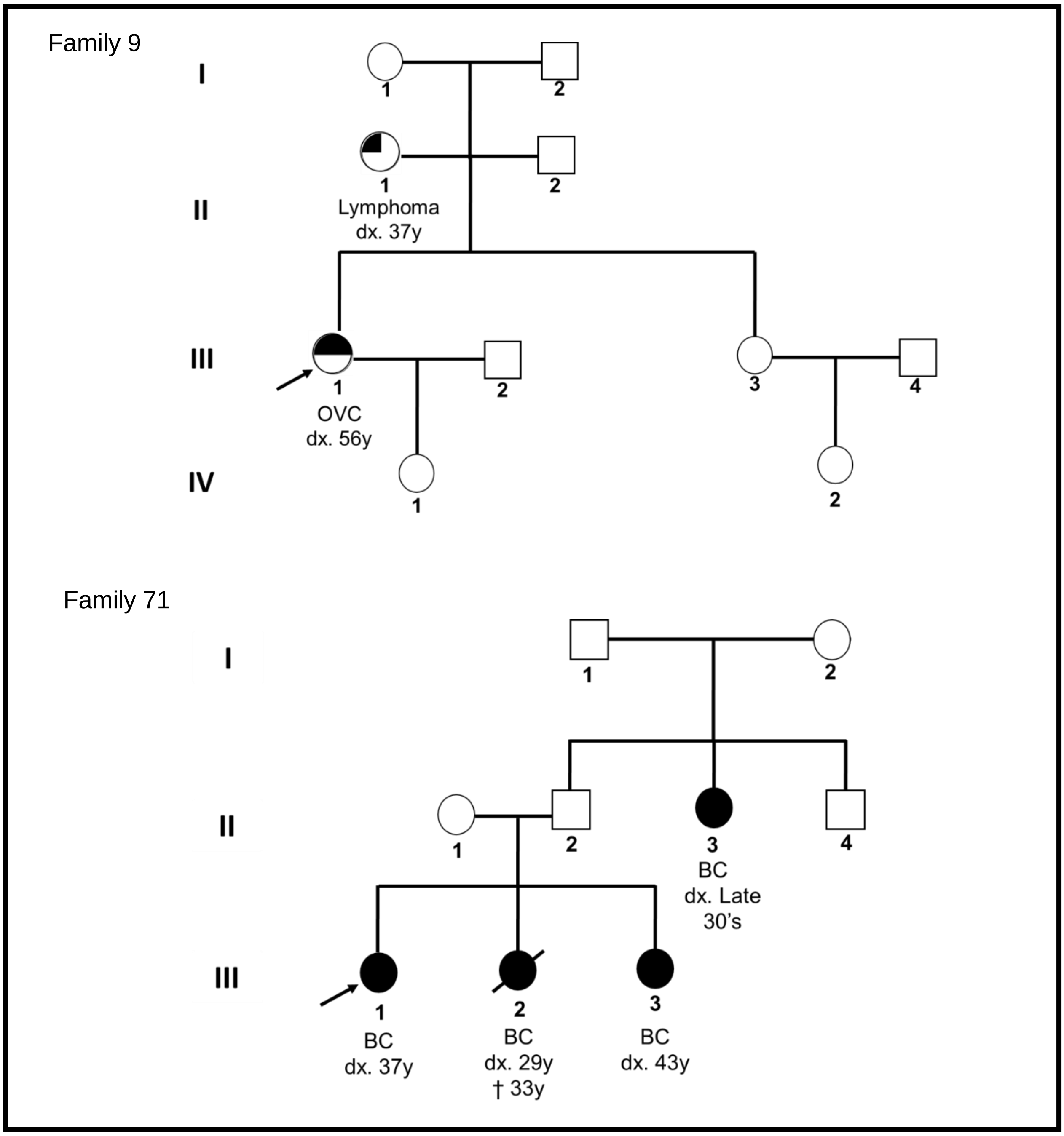
(B) - Pedigrees of families included for gene screening.

Pedigrees have been abbreviated, excluding some distant family members unaffected by disease. A key is shown below that will apply to all pedigrees. Males and females are represented by squares and circles respectively. Shapes containing numbers provide the number of individuals. Deaths due to cancer are illustrated with a line through the shape. Index cases have been indicated with an arrow. Abbreviations/symbols include dx. = age at diagnosis, † = age at death, wt = wild type allele, y = years, BC = breast cancer, OVC = ovarian, DCIS = Ductal carcinoma in situ, PC = prostate cancer, CC = colon cancer GC = gastric cancer and Ca = other observed cancers.

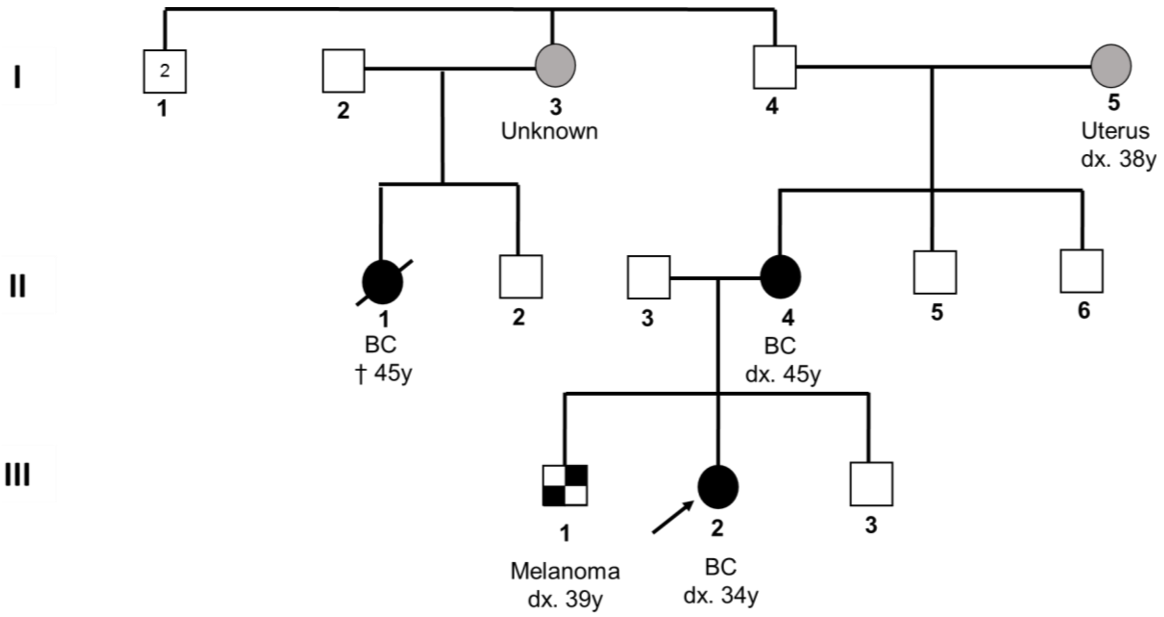
KEY

Breast Cancer	●
Ovarian Cancer	◓
Prostate Cancer	⬓
Lymphoma	◴
Non-cancer related death	⍁ ⊘
Colon Cancer	◱ ◵
Other Cancers	▩ ◍
Mole Cancer	◵
Intestinal Cancer	◐
Bladder Cancer	◑
Gastric Cancer	▚
Melanoma	▚
Liver Cancer	◷
Unknown gender	◇

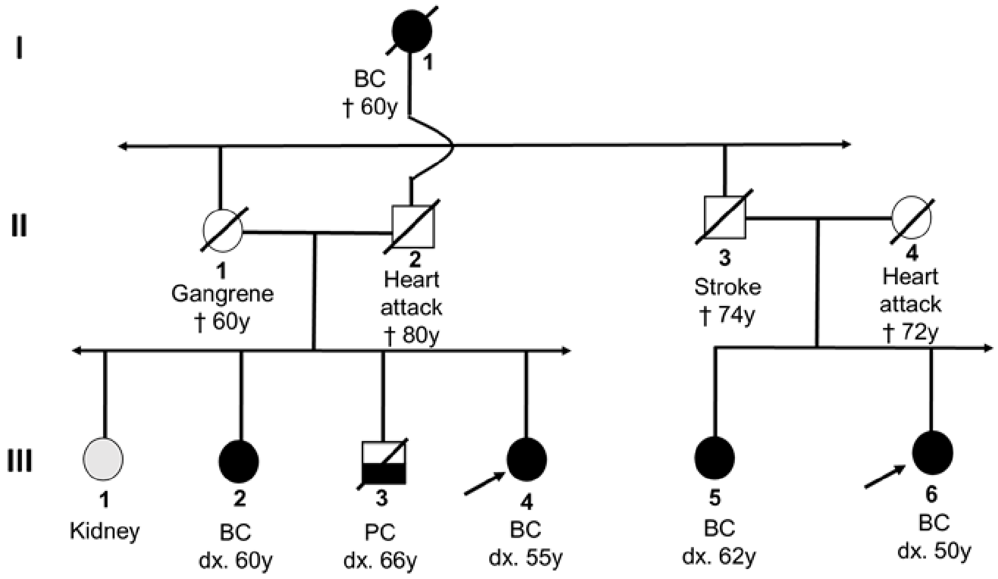
(A) Pedigrees of subjects that had undergone whole exome sequencing.

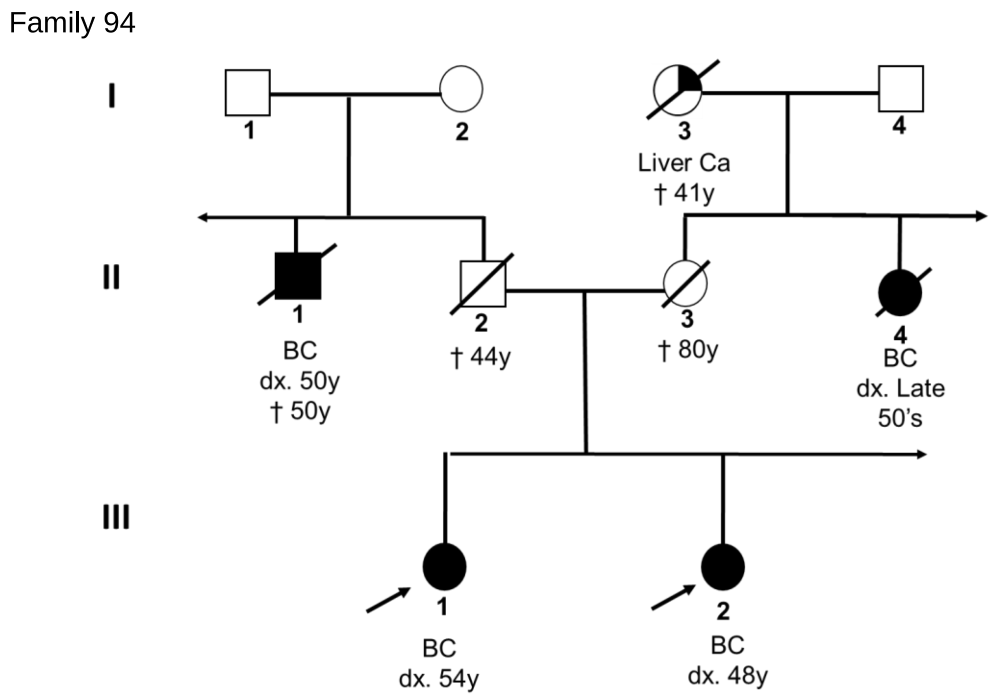


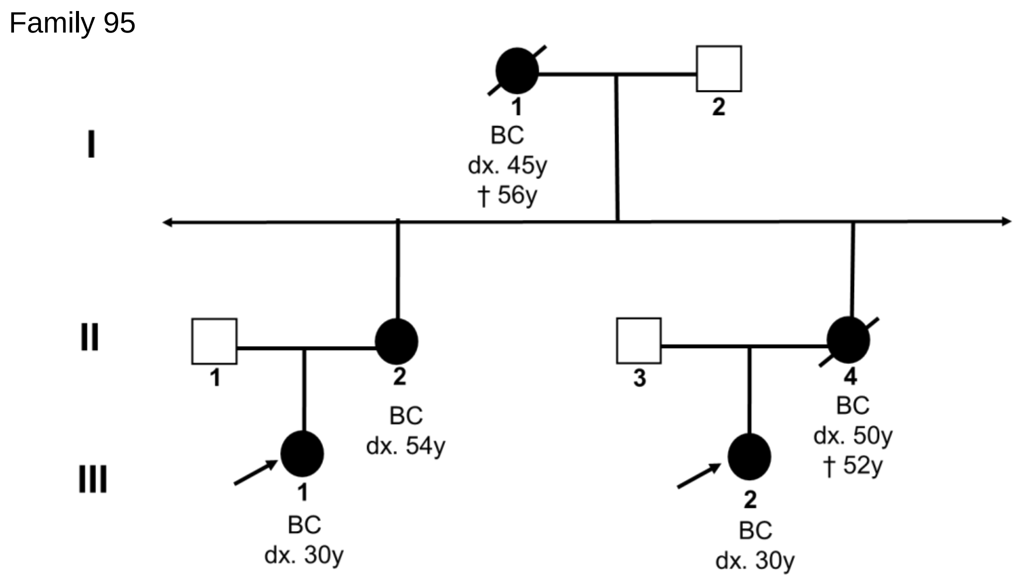
Family 73



Family 92

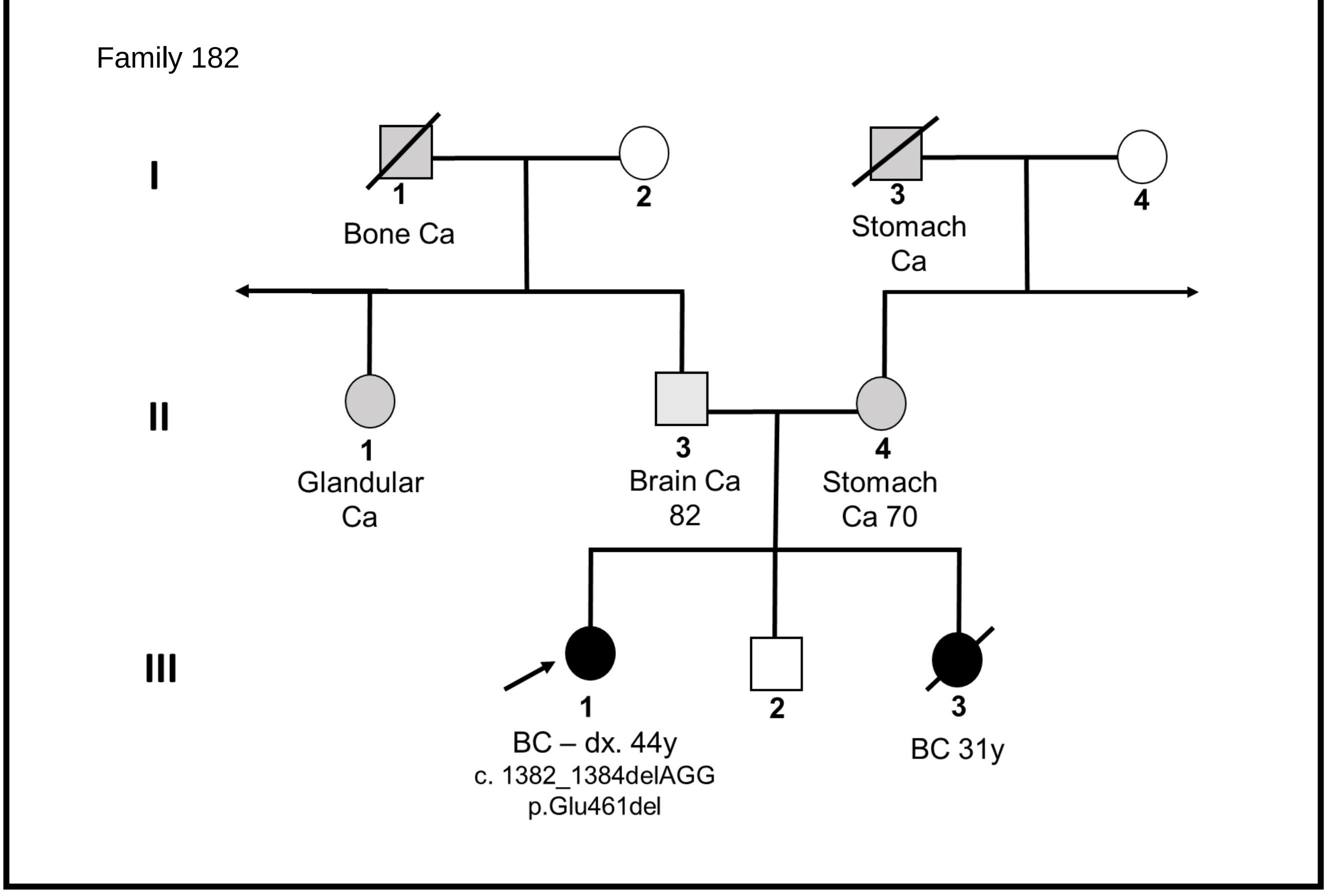



Family 94
I
II
III
1
2
3
Liver Ca
† 41y
4
1
BC
dx. 50y
† 50y
2
† 44y
3
† 80y
4
BC
dx. Late
50's
1
BC
dx. 54y
2
BC
dx. 48y


Family 95
I
1
BC
dx. 45y
† 56y
2
II
1
2
BC
dx. 54y
3
4
BC
dx. 50y
† 52y
III
1
BC
dx. 30y
2
BC
dx. 30y

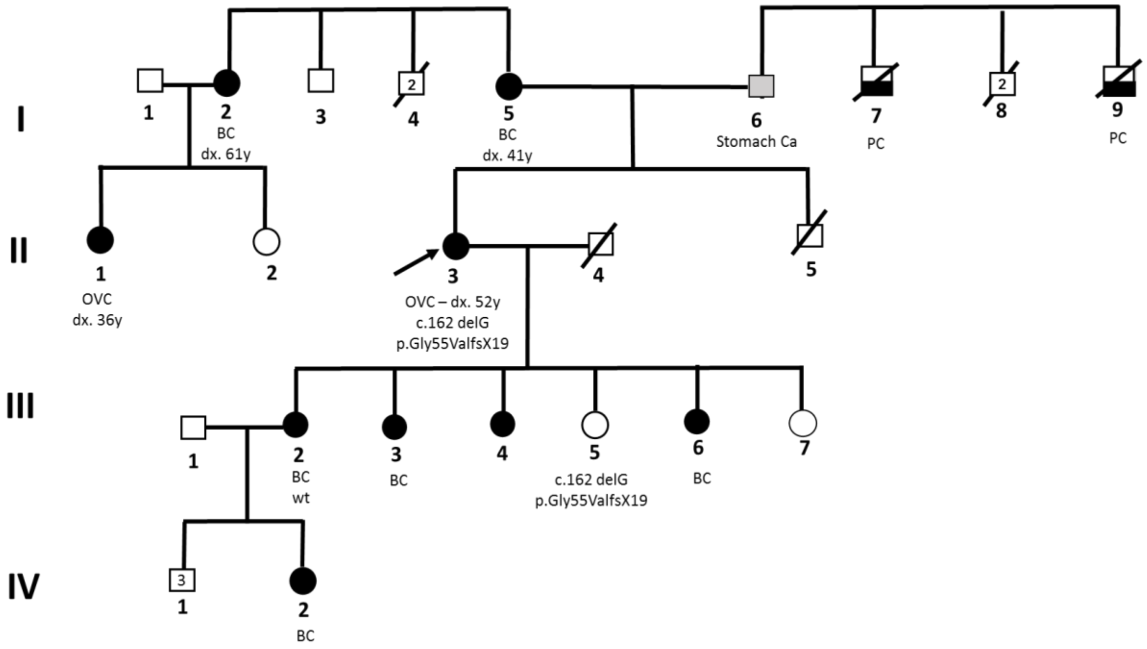
(B) Pedigrees of *TCHP* and *EME2* mutation carriers

TCHP c.1382_1384delAGG

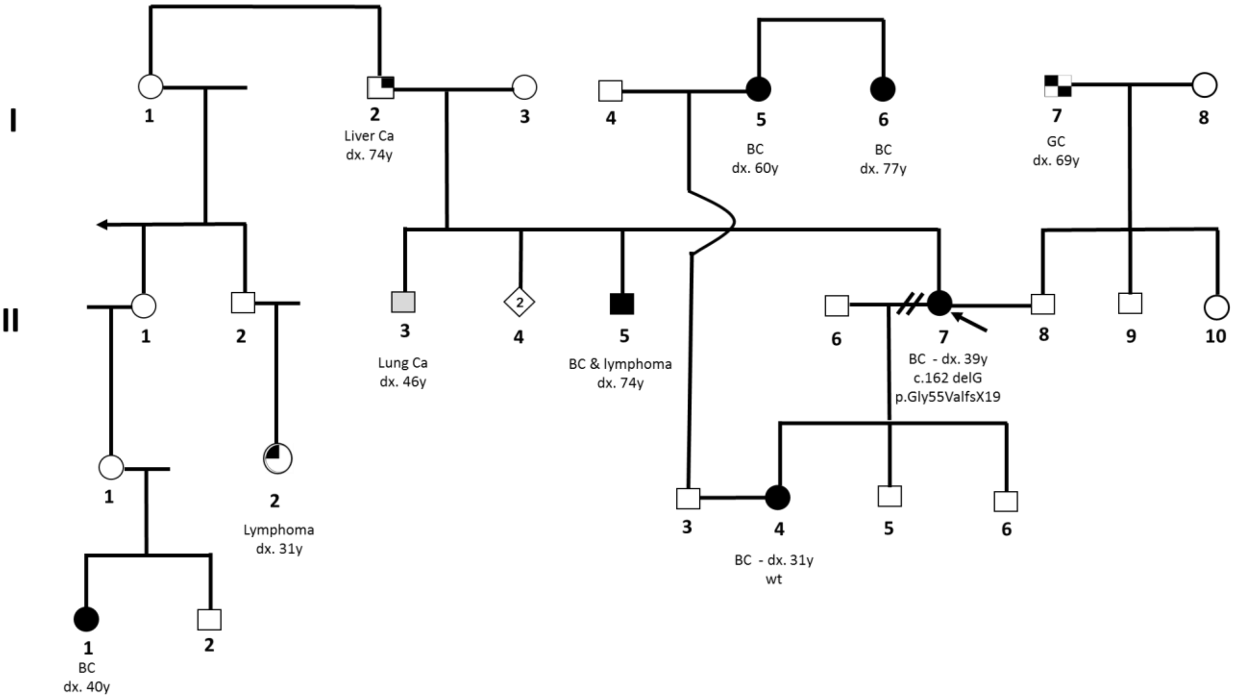


***EME2* c.162delG**

Family 4

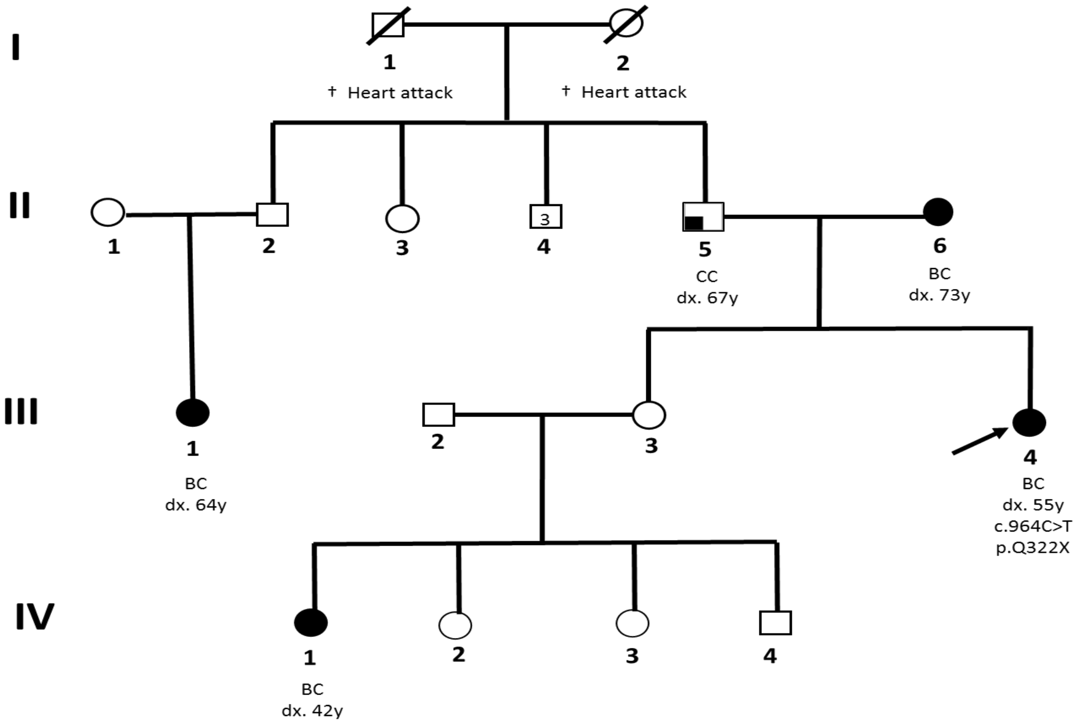


Family 157

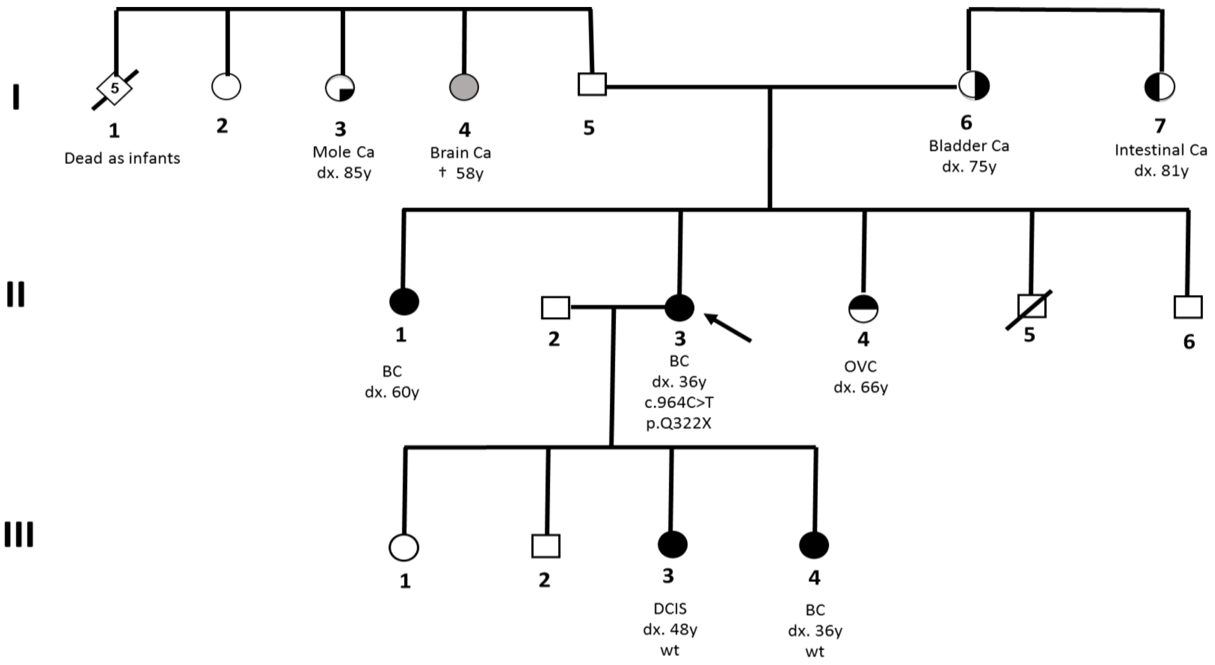


***EME2* c.964C>T**

Family 93

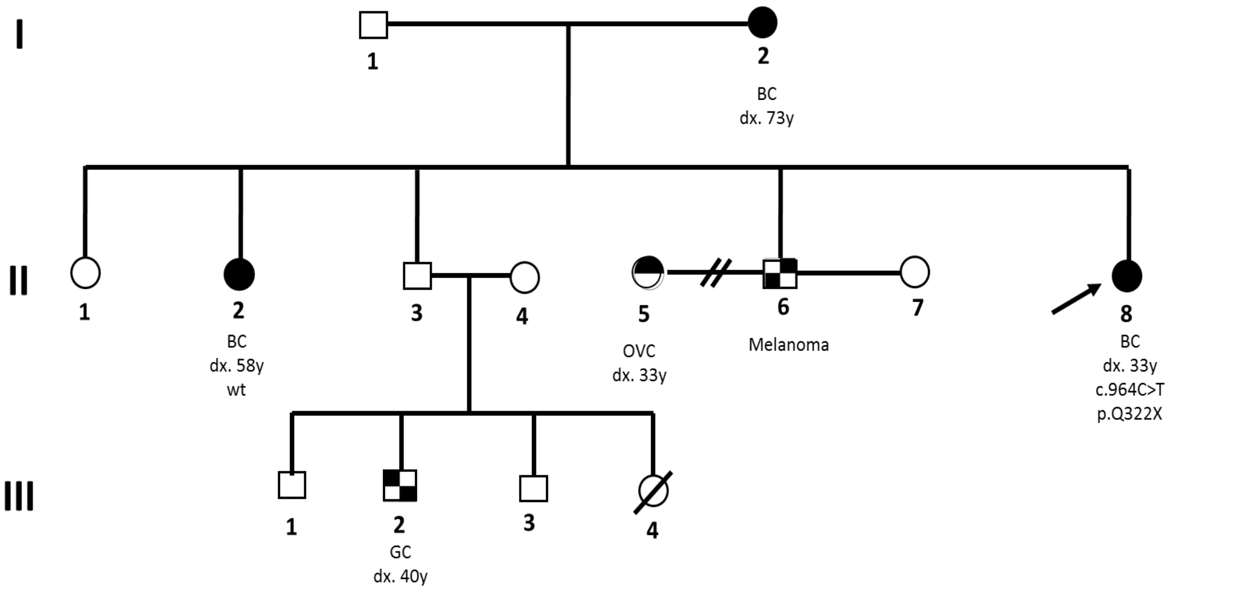


Family 96



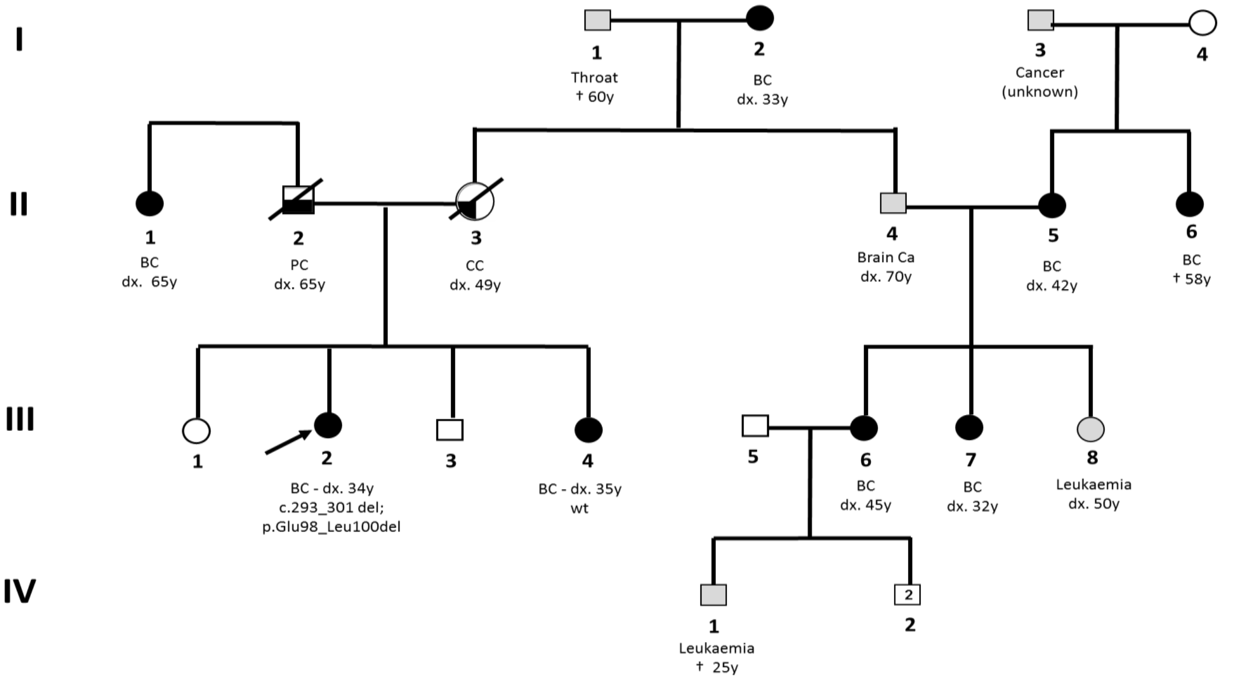
***EME2* c.964C>T**

Family 121



***EME2* c.293_301del**

Family 190



Annexure 3:

Scripts used in whole exome sequence data analysis, variant calling and annotation.

Data analysis was performed with the use of various tools that form part of the Genome Analysis Toolkit pipeline. Some third-party tools have been included and all steps taken towards variant discovery are described below. All steps have been listed by using BRC 71-1 as an example, however, the same steps were repeated for all samples unless otherwise stated.

All Linux commands have been italicized

❖ Raw Data analysis

Whole exome sequence data was received from the Beijing Genome Institute as .zip files through an internet file transfer program (FTP) address temporarily created for the transferal of the large dataset. Adapter region sequences were pre-removed from sequence reads and "clean data" for each index case were saved. Downloaded sequence data have been kept on the local Lustre filesystem set up at the Bioinformatics and Computational Biology Unit, University of Pretoria. Duplicates of this data was transferred onto a local laboratory computer for use.

- Sequence data files were unzipped using gunzip
  - ➢ *gunzip filename*
  - ➢ Each file contained paired-end reads described as BRC-71-1_1.fastq (forward reads) and BRC-71-1_2.fastq (reverse reads) for the two lanes used during the sequencing run
- Data from the two sequence lanes for forward reads were concatenated (i.e. Forward [lane1 + lane 2])
  - ➢ *cat filenameLane1 filenameLane2 > ../BRC-71-1_1.fastq*
  - ➢ This concatenated data from two forward read lanes into one file. The combined file was placed into the BRC-71-1 directory
- This was repeated for reverse strands from two sequence lanes
  - ➢ *cat filenameLane1 filenameLane2 > ../BRC-71-1_2.fastq*
- Quality control FastQC analysis
  - ➢ *fastqc BRC-71-1_1.fastq*
  - ➢ *fastqc BRC-71-1_2.fastq*
  - ➢ Each result generated was viewed to assess the necessity for base trimming

- Trimming reads
  - ➢ *fastx_trimmer -f 1 -l 85 -Q33 -i BRC-71-1_2.fastq -o BRC-71-1_2t.fastq*
  - ➢ This removed the last 5 bases from reverse reads, generally of lower quality than all other bases

❖ Variant calling with the Genome Analysis Toolkit (GATK)
(Van der Auwera, et al. 2013)

- The Genome Analysis Toolkit release package was downloaded for this study

- The latest version of GATK made use of various sources including the reference genome, human variant databases and a set of known insertion/deletion sites. The dataset was downloaded from the GSA FTP server, found in the directory named “bundle”

GATKvs2.8 resource bundle

Name	**Description**	**Purpose**
HG19 (build37)	gatk_resource_bundle/2.8/hg19/ucsc.hg19.fasta	Reference Genome
Mills 1000	gatk_resource_bundle/2.8/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf	Known insertion deletion regions for re-alignment
HAPMap 3.3.	gatk_resource_bundle/2.8/hg19/hapmap_3.3.hg19.sites.vcf	Variant recalibration
2.5 OMNI	gatk_resource_bundle/2.8/hg19/1000G_omni2.5.hg19.sites.vcf	Variant recalibration
dbSNP	gatk_resource_bundle/2.8/hg19/dbsnp_135.hg19.vcf	Variant recalibration

(Ball, et al. 2012; Li, et al. 2009)

- Paired-end left-hand reads were aligned
  - ➢ *bwa aln -f BRC71-1_1.sai -1 -t 8 GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta BRC71-1_1.fastq*
- This was repeated for right-hand (aka reverse) reads
  - ➢ *bwa aln -f BRC71-1_2.sai -2 -t 8 GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta BRC71-1_2t.fastq*
- To convert SAI to SAM (Paired-ends) and add sample headers
  - ➢ *bwa sampe -f BRC71-1.sam -r '@RG\tID:FLOWCELL1.LANE1\tPL:ILLUMINA\tLB:LIB-EXOME-1\tSM:BRC71-1'*

```
GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta BRC71-1_1.sai BRC71-1_2.sai BRC71-1_1.fastq BRC71-1_2t.fastq
```

- To convert SAM to BAM

```
➢ samtools import GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta BRC71-1.sam BRC71-1.bam
```

- To sort the BAM file

```
➢ samtools sort BRC71-1.bam BRC71-1.sorted
```

- To index the BAM file

```
➢ samtools index BRC71-1.sorted.bam BRC71-1.sorted.bam.bai
```

- Marking duplicate reads on bam file which has been sorted using samtools

```
➢ java -jar picard-tools-1.70/MarkDuplicates.jar INPUT=BRC71-1.sorted.bam OUTPUT=BRC71-1.sorted.dedup.bam METRICS_FILE=BRC71-1.sorted.bam.dupmetrics ASSUME_SORTED=true VALIDATION_STRINGENCY=LENIENT
```

- To index the sorted BAM file

```
➢ samtools index BRC-71-1.sorted.dedup.bam BRC-71-1.sorted.dedup.bam.bai
```

- Creating realignment targets by incorporating known indels

```
➢ java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -nt 8 -T RealignerTargetCreator -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -I BRC71-1.sorted.dedup.bam -o BRC71-1.intervals -known GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf
```

- Realign around indels incorporating known indels

```
➢ java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -I BRC71-1.sorted.dedup.bam -targetIntervals BRC71-1.intervals --out BRC71-1.sorted.realigned.bam -known GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf
```

- Fix mate-paired reads

```
➢ java -jar picard-tools-1.70/FixMateInformation.jar INPUT=BRC71-1.sorted.realigned.bam OUTPUT=BRC71-1_bam.sorted.realigned.fixed.bam
```

```
GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta VALIDATION_STRINGENCY=LENIENT -CREATE_INDEX=true
```

- Recalibrate base quality: CountCovariates

  -
    ```
    java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountCovariates -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -knownSites GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/dbsnp_135.hg19.vcf -I BRC71-1.sorted.realigned.bam -cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate -cov DinucCovariate -recalFile BRC71-1.sorted.realigned.bam.recal_data.csv
    ```
- Recalibrate base quality: TableRecalibration

  -
    ```
    java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T TableRecalibration -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -I BRC71-1.sorted.realigned.bam -o BRC71-1.sorted.realigned.recal.bam -recalFile BRC71-1.sorted.realigned.bam.recal_data.csv
    ```
- Recalibrate base quality: AnalyzeCovariates

  -
    ```
    java -jar GenomeAnalysisTK/AnalyzeCovariates.jar -recalFile BRC71-1.sorted.realigned.bam.recal_data.csv  -outputDir sorted.realigned.bam.recal_data_dir -ignoreQ 5
    ```
- CLC Genomics workbench was incorporated to assess the coverage distribution for target enriched regions obtained through high-throughput sequencing

  - The function “create statistics for target regions” was selected under the “Resequencing” subfolder in the “toolbox” bar

  - Target regions enriched during DNA library preparation were imported as a target_interval_hg19.bed file

  - Default sets of pre-defined coverage thresholds were used as the report type

  - Minimum coverage was set at 10X

  - A detailed report was generated containing the read mapping efficiency statistics for targeted exome intervals

- Perform SNP and Indel variant calling

  -
    ```
    java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -nt 8 -T UnifiedGenotyper -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta
    ```

```
-I BRC71-1.sorted.realigned.recal.bam --dbsnp GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/dbsnp_135.hg19.vcf -glm BOTH -dcov 1000 -A DepthOfCoverage -A AlleleBalance -L target_intervals.bed -o variants.raw.vcf -stand_call_conf 50.0 -stand_emit_conf 10.0
```

- Select only SNPs

➢
```
java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T SelectVariants -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta --variant variants.raw.vcf -o snp.raw.vcf -selectType SNP
```

- Select only indels

➢
```
java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T SelectVariants -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta --variant variants.raw.vcf -o indel.raw.vcf -selectType INDEL
```

- Build SNP error model with VQSR

➢
```
java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -nt 8 -T VariantRecalibrator -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -input snp.raw.vcf --maxGaussians 4 -resource:hapmap,VCF,known=false,training=true,truth=true,prior=15.0 GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/hapmap_3.3.hg19.sites.vcf -resource:omni,VCF,known=false,training=true,truth=false,prior=12.0 GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/1000G_omni2.5.hg19.sites.vcf -resource:dbsnp,VCF,known=true,training=false,truth=false,prior=6.0 GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/dbsnp_135.hg19.vcf -an QD -an HaplotypeScore -an MQRankSum -an ReadPosRankSum -an HRun -an FS -an MQ -mode SNP -recalFile snp.recal -tranchesFile snp.tranches -rscriptFile snp.plots.R
```

- Recalibrate SNPs

➢
```
java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T ApplyRecalibration -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -input snp.raw.vcf --ts_filter_level 99.0 -tranchesFile snp.tranches -recalFile snp.recal -o snp.recal.filtered.vcf
```

- Hard filter SNP variants according to listed parameters
  - ➢ *java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T VariantFiltration -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -o snp.HDpfiltered.vcf --variant snp.recal.filtered.vcf --filterExpression "QD < 2.0" --filterExpression "MQ < 40.0" --filterExpression "FS > 60.0" --filterExpression "HaplotypeScore > 13.0" --filterExpression "QUAL >= 30" --filterExpression "DP >= 10" --filterName QDFilter --filterName MQFilter --filterName FSFilter --filterName HaplotypeFilter --filterName QUALFilter --filterName DepthFilter*

- Hard filter indel variants according to listed parameters
  - ➢ *java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T VariantFiltration R data/GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta -o indel.HDpfiltered.vcf --variant indel.raw.vcf --filterExpression "QD < 2.0" --filterExpression "FS > 200.0" --filterExpression "QUAL >= 30" --filterExpression "DP >= 10" --filterName QDFilter --filterName FSFilter --filterName QUALFilter --filterName DepthFilter*
- Combine SNP and Indel variant data
  - ➢ *java -jar GenomeAnalysisTK/GenomeAnalysisTK.jar -T CombineVariants -R GRCGalaxyDatabase/H.sapiens/gatk_resource_bundle/hg19/ucsc.hg19.fasta --variant snp.HDfiltered.vcf --variant indel.HDfiltered.vcf -o finalvariants.vcf -genotypeMergeOptions UNSORTED*
- Convert SNPs and indels to annovar format
  - ➢ */usr/local/annovar/convert2annovar.pl --format vcf4 --includeinfo finalvariants.vcf > variants.annovar*
- Run annovar to produce csv output
  - ➢ */usr/local/annovar/summarize_annovar.pl --outfile annovar.out --buildver hg19 --ver1000g 1000g2010nov --verdbsnp 137 indel.annovar GRCGalaxyDatabase/H.sapiens/annovar/humandb/*
- Variant annotation files were transferred from the laboratory cluster to a standard desktop computer for investigation in tabular format as this is compatible with Microsoft excel
  - ➢ WinSCP was used to securely transfer files to from the personal home/<user> directory within the cluster
- Variant (.vcf) data was also transferred for further analysis with the VariantDB analysis tool (Vandeweyer, et al. 2014)

Biological interpretation and filtration of whole exome variants

❖ VariantDB analysis tool

Analysis of exome sequence data was performed with each sample case. The step by step filtration protocol has been indicated below using BRC 9-3 as an example. Identical analyses were performed for all 9 index cases.

- Sample upload

- A unique sample identifying barcode was created with the BRC9-3 patient code as the display name
- Instances where two individuals, from one family, were sequenced were described
- Variant call data generated from the GATK pipeline was directly uploaded
- Each sample was associated with two files in variant call format (.vcf) as both single nucleotide mutations and insertion/deletions were uploaded simultaneously

- Variant analysis

- The uploaded variant files were annotated by creating separate analyses for each affected index case

- Variant filtration

- The total raw variants identified by the Unified genotyper was filtered based on the following criteria
- A quality based filtration step was set to keep all bases with a base call score no less than 20 and sequence read depth ≥10X
- Variants were kept if observed with allele frequencies ≤1,0% when compared to the most updated releases from the 1000 Genomes project and public Complete Genomics genomes dataset or ≤3,0% when compared to the NHLBI Exome sequencing project (ESP) 6500
- Variants were further filtered to keep all truncating variants (i.e. frameshift insertion/deletion or stop codon gain) or mutations predicted to possibly disrupt splice sites up to 2.0 bases into intron
- Single amino acid substitutions were filtered to select those computationally predicted as putatively pathogenic according to four *in silico* tools (incl. LJB SIFT, LJB PolyPhen2, LJB CADD and LJB GERP)

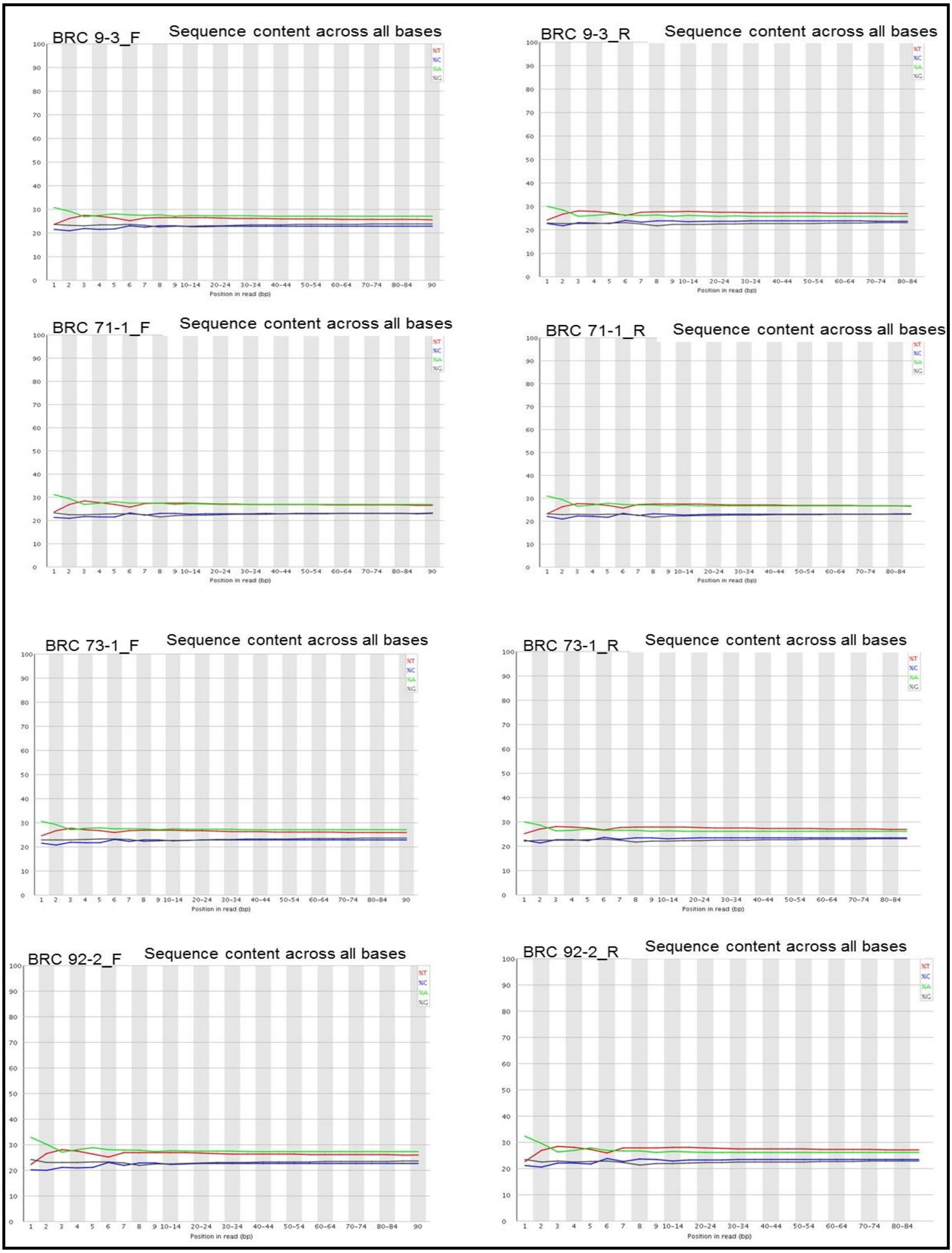
- Case sample variants clearly associated as loss of function mutations, clinically asserted as pathogenic or those that were heterozygous at a specific allele were further selected

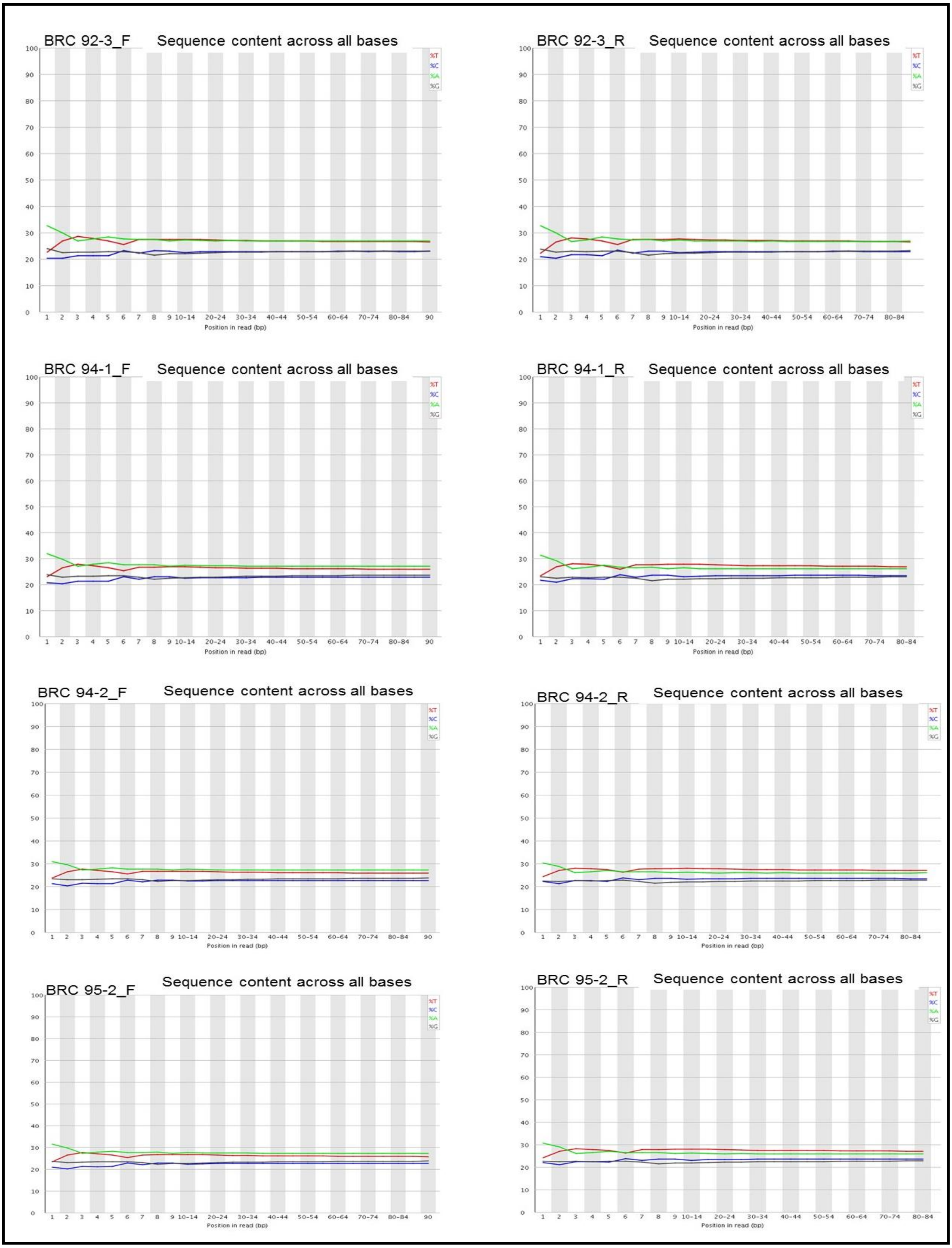
- Variant prioritisation

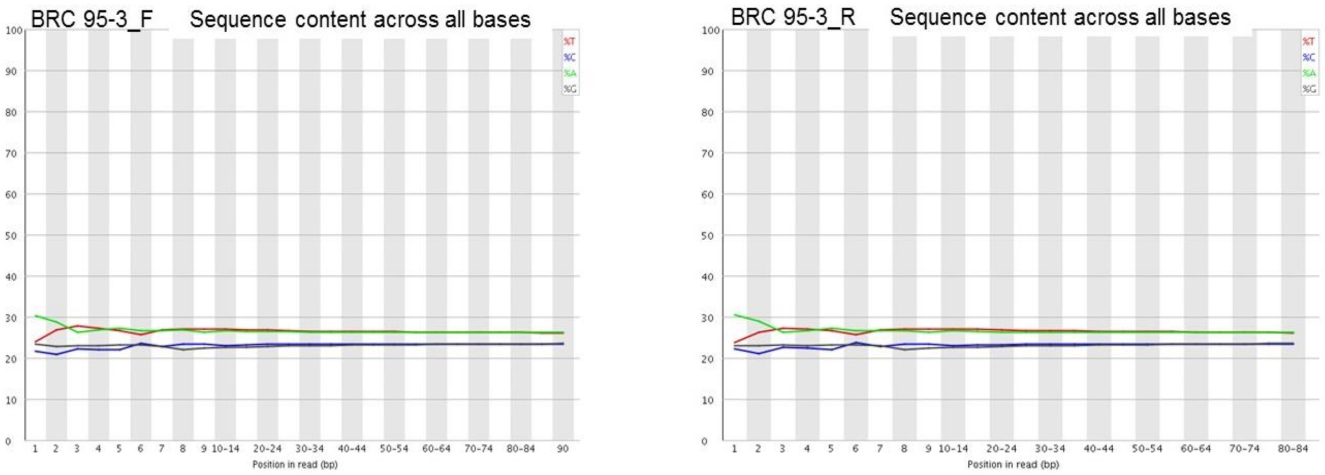
- All heterozygous variants predicted as possibly deleterious were prioritised as follows;
- Gene variants were kept if: mutated in both cases on gene level where two individuals were sequenced OR variants are present in both index cases OR mutation occurs in at least one of the case samples at the gene level
- Various parameters were chosen to retain variants that are associated with DNA damage signalling and repair pathways, regulation of cell proliferation and apoptosis and have been characterised in literature studies. We also included any variant, found in somatic breast/ovarian cancer databases at allele frequencies above 0.01%
- This pipeline generated a list of genes which we investigated to prioritise for the most significant genes with a possible link to our biological and clinical interest. Genes were stratified by first selecting those containing potential truncating variants (frameshift & stop-gain). Splice-site variants (SC), in-frame insertion/deletions (IF-indels) and finally damaging missense mutations followed
- We prioritised for genes with a more significant biological contribution to maintaining genomic stability and overall cellular homeostasis

Annexure 4:

Raw base call quality scores reviewed with FastQC.

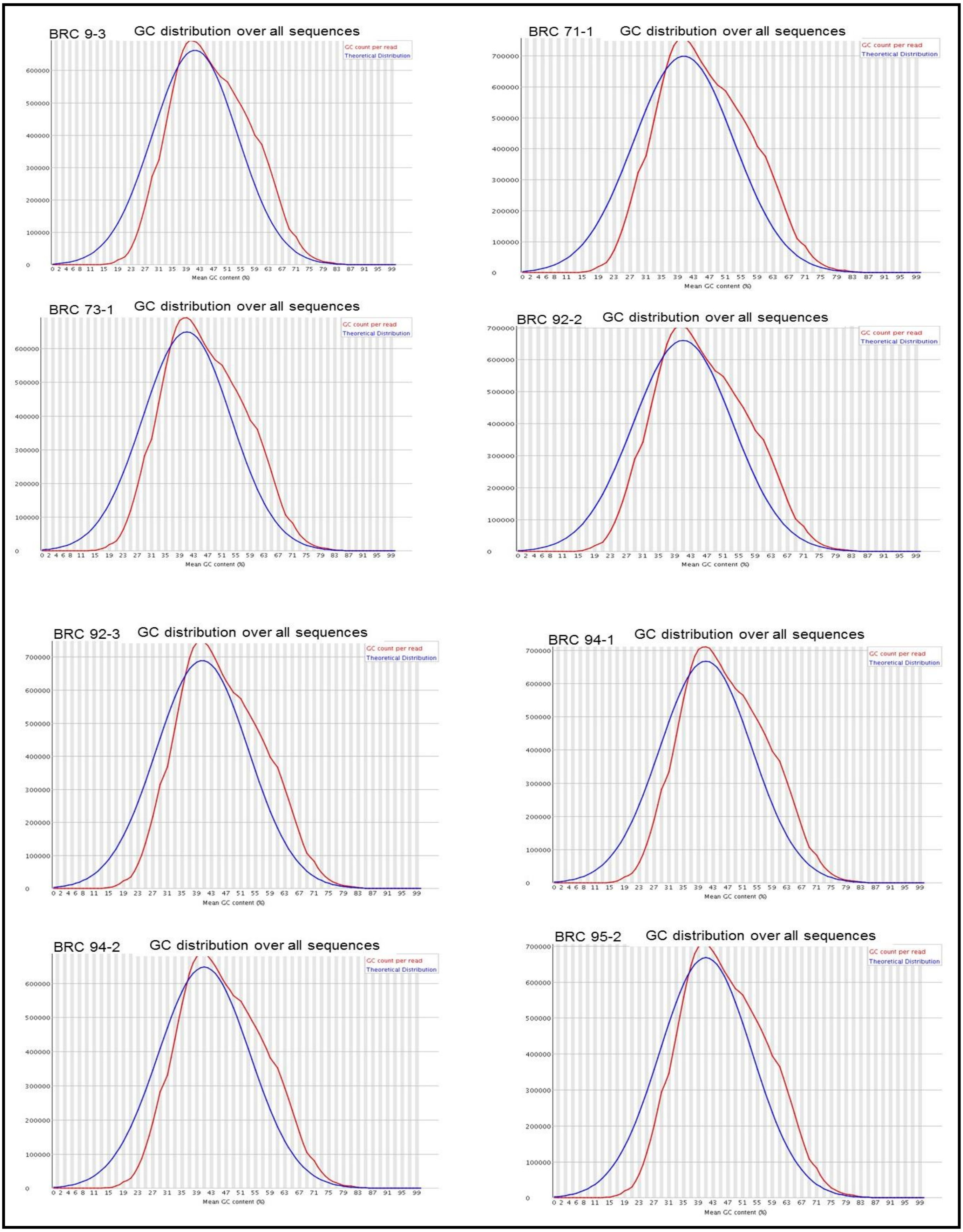

BRC 9-3_F
Sequence content across all bases
BRC 9-3_R
Sequence content across all bases
BRC 71-1_F
Sequence content across all bases
BRC 71-1_R
Sequence content across all bases
BRC 73-1_F
Sequence content across all bases
BRC 73-1_R
Sequence content across all bases
BRC 92-2_F
Sequence content across all bases
BRC 92-2_R
Sequence content across all bases
%T
%C
%A
%G
0
10
20
30
40
50
60
70
80
90
100
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)

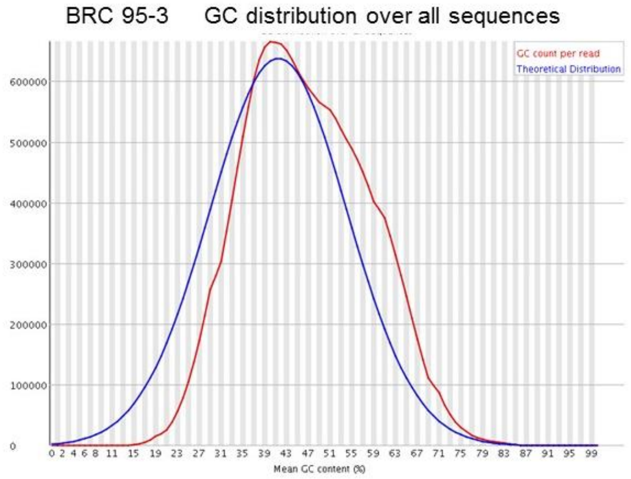

BRC 92-3_F Sequence content across all bases
BRC 92-3_R Sequence content across all bases
BRC 94-1_F Sequence content across all bases
BRC 94-1_R Sequence content across all bases
BRC 94-2_F Sequence content across all bases
BRC 94-2_R Sequence content across all bases
BRC 95-2_F Sequence content across all bases
BRC 95-2_R Sequence content across all bases
%T
%C
%A
%G
0
10
20
30
40
50
60
70
80
90
100
1 2 3 4 5 6 7 8 9 10-14 20-24 30-34 40-44 50-54 60-64 70-74 80-84 90
Position in read (bp)



Representations of per_base_sequence content.

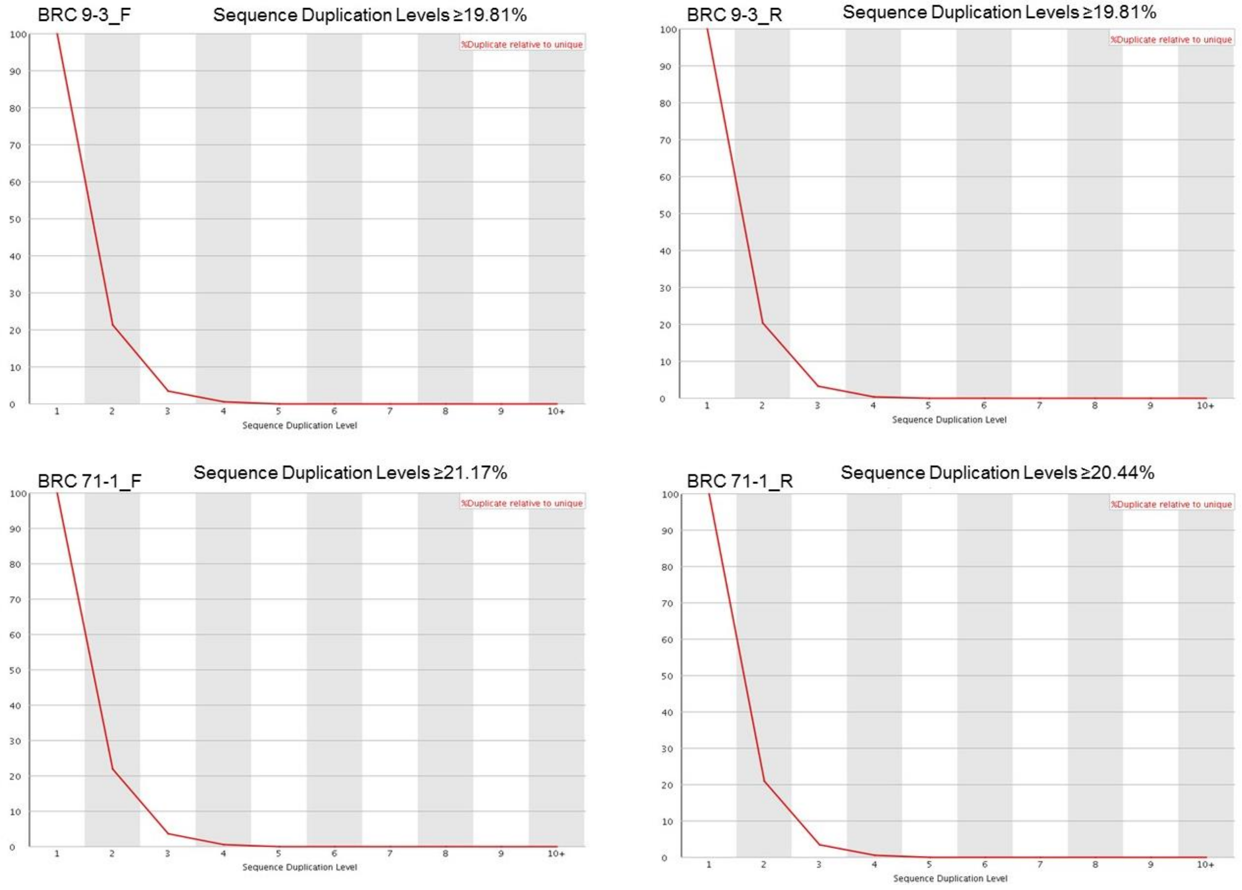
Plots illustrating the proportion of bases across all forward (F) and reverse (R) reads. This describes the percentage of guanine (red), adenine (blue), thymine (green) and cytosine (black) nucleotide bases incorporated across the ~84bp sequence read.

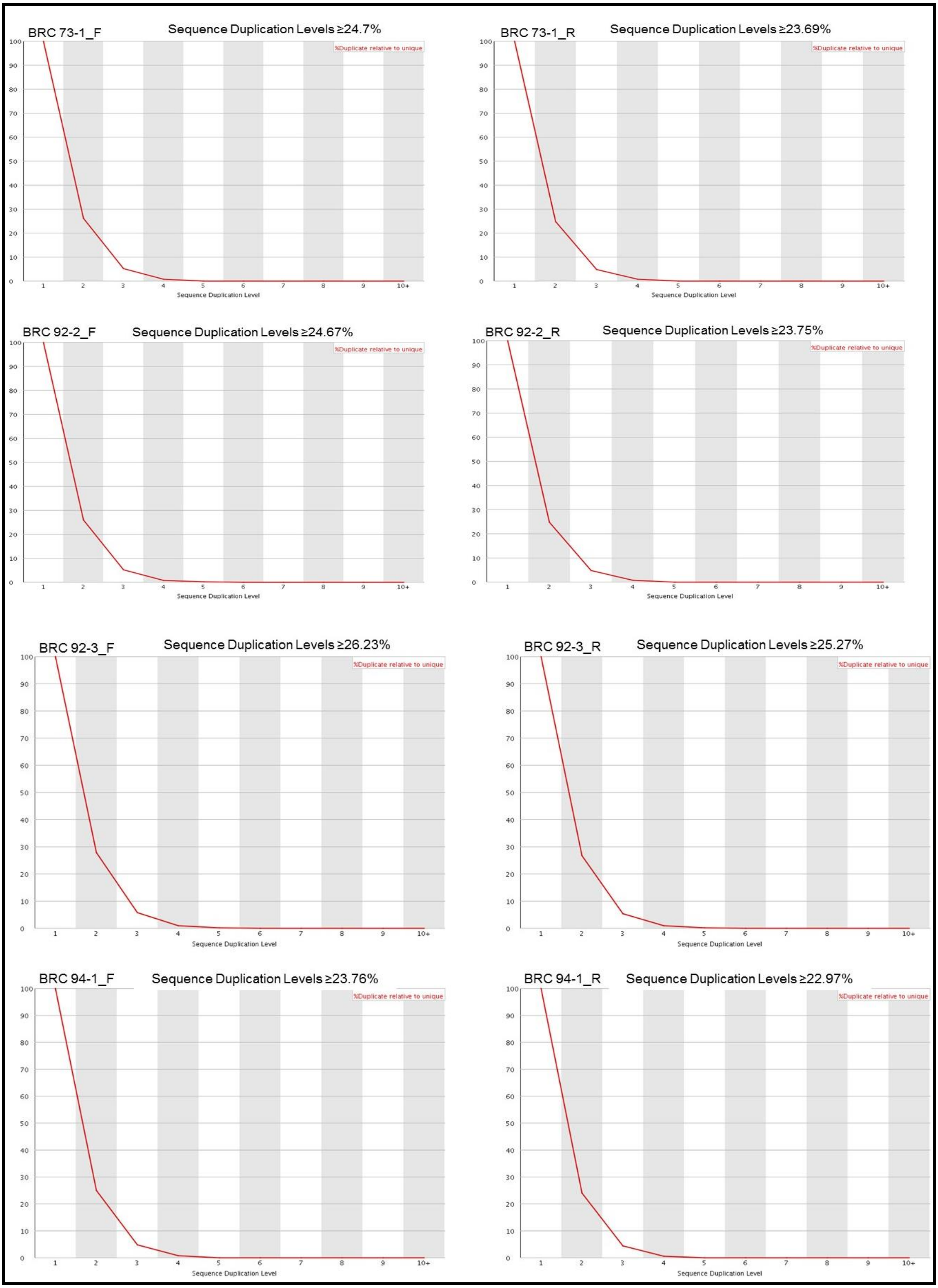

BRC 9-3
GC distribution over all sequences
BRC 71-1
GC distribution over all sequences
BRC 73-1
GC distribution over all sequences
BRC 92-2
GC distribution over all sequences
BRC 92-3
GC distribution over all sequences
BRC 94-1
GC distribution over all sequences
BRC 94-2
GC distribution over all sequences
BRC 95-2
GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)
0
100000
200000
300000
400000
500000
600000
700000
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99

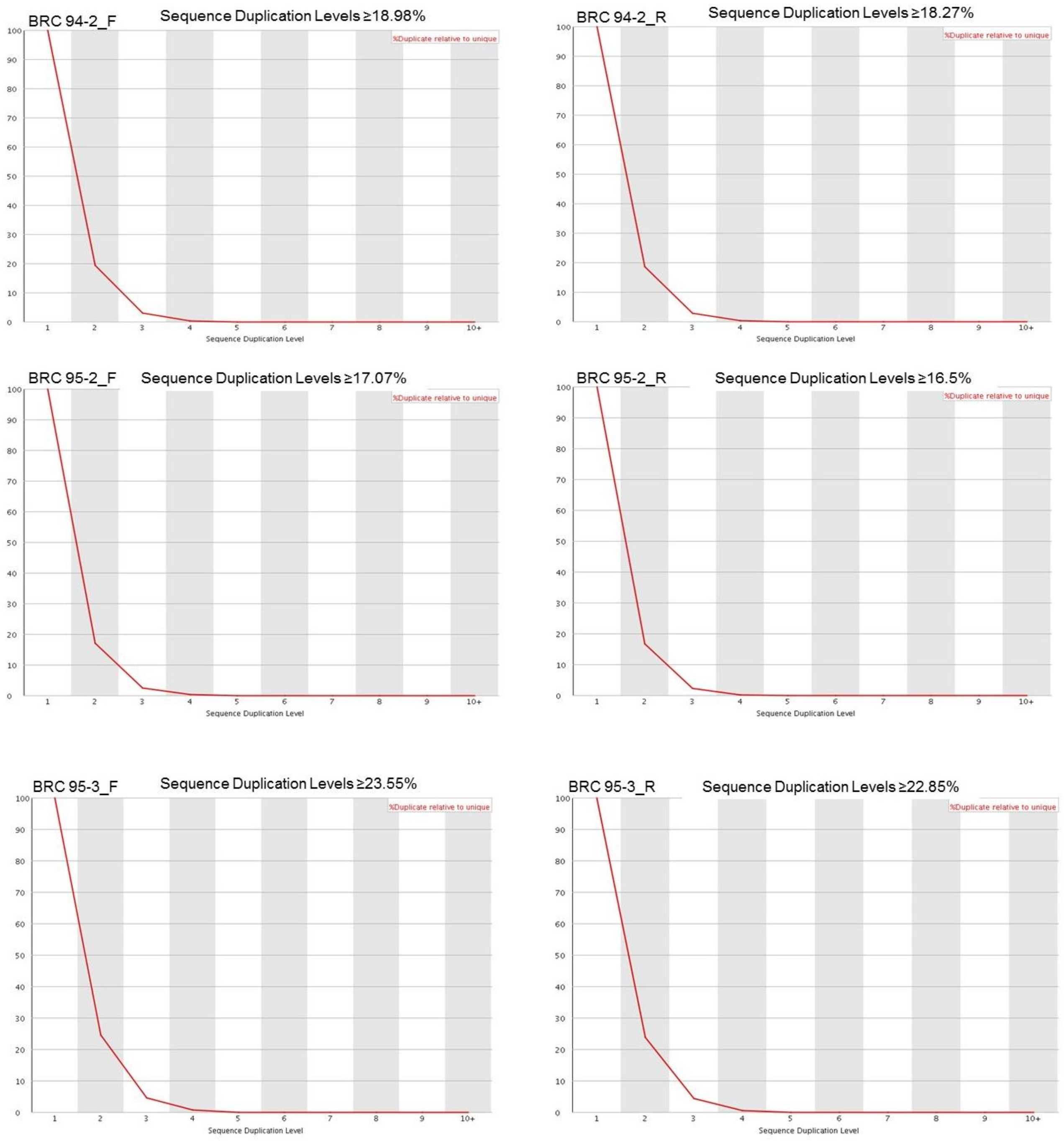


Representations of per_sequence_GC_content.

The GC content across ~84 bases in all sequence reads. This was quantified for each file (index case, red) and compared to the modelled normal (a.k.a theoretical) distribution representing the GC content that underlies the human genome (blue).




BRC 73-1_F
Sequence Duplication Levels ≥24.7%
BRC 73-1_R
Sequence Duplication Levels ≥23.69%
BRC 92-2_F
Sequence Duplication Levels ≥24.67%
BRC 92-2_R
Sequence Duplication Levels ≥23.75%
BRC 92-3_F
Sequence Duplication Levels ≥26.23%
BRC 92-3_R
Sequence Duplication Levels ≥25.27%
BRC 94-1_F
Sequence Duplication Levels ≥23.76%
BRC 94-1_R
Sequence Duplication Levels ≥22.97%
%Duplicate relative to unique
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10+
Sequence Duplication Level



BRC 94-2_F Sequence Duplication Levels ≥18.98%

BRC 94-2_R Sequence Duplication Levels ≥18.27%

BRC 95-2_F Sequence Duplication Levels ≥17.07%

BRC 95-2_R Sequence Duplication Levels ≥16.5%

BRC 95-3_F Sequence Duplication Levels ≥23.55%

BRC 95-3_R Sequence Duplication Levels ≥22.85%

Representations of sequence_duplication_levels.

The percentage duplicated sequences that appear in the first 100 bases. Sequence duplication levels stabilised beyond the first 100 read bases, which indicates that there is no bias in the represented sequences.

Annexure 5:

(A) - 516 Genes representative of major DNA damage recognition and repair processes.

(B) - Rare SNVs (missense and in-frame insertions/deletions) detected in DNA damage response associated genes.

(A) - 516 Genes representative of major DNA damage recognition and repair processes.

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
AATF	26574	*CASP3*	836
ABL1	25	*CCNA1*	8900
ACTR5	79913	*CCNA2*	890
ADA	100	*CCNB1*	891
AKT1	207	*CCND1*	595
ALKA	947371	*CCNE1*	898
ALKB	946708	*CCNH*	902
ALKBH1	8846	*CCNO*	10309
ALKBH2	121642	*CDC14B*	8555
ALKBH3	221120	*CDC25A*	993
APEX1	328	*CDC25B*	994
APEX2	27301	*CDC25C*	995
APITD1	378708	*CDC45*	8318
APLF	200558	*CDC6*	990
APTX	54840	*CDH13*	1012
ASCC3	10973	*CDK1*	983
ASF1A	25842	*CDK2*	1017
ASTE1	28990	*CDK4*	1019
ATF2	1386	*CDK7*	1022
ATM	472	*CDKN1A*	1026
ATMIN	23300	*CDKN1B*	1027
ATR	545	*CDKN2A*	1029
ATRIP	84126	*CDKN2D*	1032
ATRX	546	*CEBPG*	1054
ATXN3	4287	*CEP164*	22897
AXIN2	8313	*CEP170*	9859
BACH1	571	*CETN2*	1069
BAP1	8314	*CHAF1A*	10036
BARD1	580	*CHAF1B*	8208
BAX	581	*CHD1L*	9557
BAZ1B	9031	*CHD4*	1108
BCCIP	56647	*CHEK1*	1111
BLM	641	*CHEK2*	11200
BRAP	8315	*CHRNA4*	1137
BRCA1	672	*CIB1*	10519
BRCA2	675	*CINP*	51550
BRCC3	79184	*CLSPN*	63967
BRE	9577	*COPS5*	10987
BRIP1	83990	*COPS6*	10980
BTG2	7832	*CRB2*	286204
BUB1	699	*CREB1*	1385
BUB1B	701	*CREBBP*	1387
C11orf30	56946	*CRY1*	1407
C17orf70	80233	*CRY2*	1408
C19ORF40	91442	*CSA*	1161

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
CSB	2074	*EYA1*	2138
CSNK1D	1453	*EYA2*	2139
CSNK1E	1454	*EYA3*	2140
CUL4A	8451	*EYA4*	2070
CUL4B	8450	*FAM175A*	84142
CYP19A1	1588	*FANCA*	2175
CYP1A1	1543	*FANCB*	2187
DAPK1	1612	*FANCC*	2176
DBF4	10926	*FANCD2*	2177
DCLRE1A	9937	*FANCE*	2178
DCLRE1B	64858	*FANCF*	2188
DCLRE1C	64421	*FANCG*	2189
DDB1	1642	*FANCI*	55215
DDB2	1643	*FANCL*	55120
DDR1	780	*FANCM*	57697
DDX1	1653	*FBXO18*	84893
DEK	7913	*FBXO6*	26270
DHX9	1660	*FEN1*	2237
DMAP1	55929	*FGF10*	2255
DMC1	11144	*FHIT*	2272
DNA2	1763	*FIGN*	55137
DOT1L	84444	*FIGNL1*	63979
DTL	51514	*FOS*	2353
DTX3L	151636	*FOXM1*	2305
DUSP3	1845	*FTO*	79068
DYRK2	8445	*FZR1*	51343
E2F1	1869	*GADD45A*	1647
E2F2	1870	*GADD45G*	10912
E2F4	1874	*GEN1*	348654
E2F6	1876	*GPS1*	2873
EEPD1	80820	*GSTP1*	2950
EGFR	1956	*GTF2H1*	2965
EME1	146956	*GTF2H2*	2966
EME2	197342	*GTF2H2C*	653238
EP300	2033	*GTF2H3*	2967
EPC2	26122	*GTF2H4*	2968
ERBB2	2064	*GTF2H5*	404672
ERCC1	2067	*H2AFX*	3014
ERCC2	2068	*HDAC1*	3065
ERCC3	2071	*HDAC2*	3066
ERCC4	2072	*HELB*	92797
ERCC5	2073	*HELQ*	113510
ERCC6	2074	*HERC2*	8924
ERCC8	1161	*HIC1*	3090
ESCO1	114799	*HINFP*	25988
ESCO2	157570	*HIST3H2A*	92815
ESR1	2099	*HMGB1*	100130561
ETS1	2113	*HMGB2*	3148
EXO1	9156	*HUS1*	3364

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
HUS1B	135458	*MUM1*	84939
HUWE1	10075	*MUS81*	80198
IFI16	3428	*MUTYH*	4595
IGF1	3479	*MYC*	4609
IGHMBP2	3508	*NBN*	4683
IKBKG	8517	*NCOA6*	23054
INO80	54617	*NEIL1*	79661
INO80D	54891	*NEIL2*	252969
INO80E	283899	*NEIL3*	55247
INTS3	65123	*NEK1*	4750
IRS1	3667	*NEK11*	79858
JMY	133746	*NFKB1*	4790
JUN	3725	*NHEJ1*	79840
KAT5	10524	*NINL*	22981
KDM2A	22992	*NME1*	654364
KIAA0101	9768	*NME1*	4830
KIAA0430	9665	*NME1*	4831
KIAA2022	340533	*NONO*	4841
KIF22	3835	*NSMCE1*	197370
KIN	22944	*NSMCE2*	286053
KPNA2	3838	*NTH*	4913
LIG1	3978	*NTHL1*	4913
LIG3	3980	*NUDT1*	4521
LIG4	3981	*OGG1*	4968
MAD2L2	10459	*OGT*	8473
MBD4	8930	*OTUB1*	55611
MC1R	10381	*PALB2*	79728
MCM9	254394	*PAPD7*	11044
MCPH1	79648	*PARG*	8505
MDC1	9656	*PARP1*	142
MDM2	4193	*PARP2*	10038
MDM4	4194	*PARP3*	10039
MED17	9440	*PARP4*	143
MEN1	4221	*PARP9*	83666
MGMT	4255	*PCNA*	5111
MLH1	4292	*PLK1*	5347
MLH3	27030	*PLK3*	1263
MMS19	64210	*PMS1*	5378
MNAT1	4331	*PMS2*	5395
MORF4L1	10933	*PMS6*	5382
MORF4L2	9643	*PNKP*	11284
MPG	4350	*POLA1*	5422
MRE11A	4361	*POLB*	5423
MSH2	4436	*POLD1*	5424
MSH3	4437	*POLD2*	5425
MSH4	4438	*POLD3*	10714
MSH5	4439	*POLD4*	57804
MSH6	2956	*POLDIP2*	26073
MTA1	9112	*POLDIP3*	84271

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
POLE	5426	*RAD51AP1*	10635
POLE2	5427	*RAD51C*	5889
POLE3	54107	*RAD51L1*	5890
POLE4	56655	*RAD51L3*	5892
POLG	5428	*RAD52*	5893
POLG2	11232	*RAD54B*	25788
POLH	5429	*RAD54L*	8438
POLI	11201	*RAD9A*	5883
POLK	51426	*RAD9B*	144715
POLL	27343	*RASSF1*	11186
POLM	27434	*RB1*	5925
POLN	353497	*RBBP4*	5928
POLQ	10721	*RBBP4*	642954
POLR2A	5430	*RBBP7*	5931
POLR2B	5431	*RBBP8*	5932
POLR2C	5432	*RBM14*	5936
POLR2D	5433	*RBX1*	9978
POLR2E	5434	*RCHY1*	25898
POLR2F	5435	*RDM1*	201299
POLR2G	5436	*REC8*	9985
POLR2H	5437	*RECQL*	5965
POLR2I	5438	*RECQL2*	840026
POLR2J	5439	*RECQL4*	9401
POLR2K	5440	*RECQL5*	9400
POLR2L	5441	*RELA*	5970
PPM1D	8493	*REV1*	51455
PPP1CA	5499	*REV3L*	5980
PPP2R2A	5520	*RFC1*	5981
PPP2R5A	5525	*RFC2*	5982
PPP2R5B	5526	*RFC3*	5983
PPP2R5C	5527	*RFC4*	5984
PPP2R5D	5528	*RFC5*	5985
PPP2R5E	5529	*RFWD2*	64326
PPP4C	5531	*RFWD3*	55159
PPP4R2	151987	*RNASEH2A*	10535
PRKDC	5591	*RNF168*	165918
PRMT6	55170	*RNF169*	254225
PRPF19	27339	*RNF8*	9025
PSMD3	5709	*RPA1*	6117
PTTG1	9232	*RPA2*	6118
PTTG1	10744	*RPA3*	6119
RAD1	5810	*RPA4*	29935
RAD17	5884	*RPAIN*	84268
RAD18	56852	*RPS27A*	728590
RAD21	5885	*RPS27L*	51065
RAD23A	5886	*RPS3*	440991
RAD23B	5887	*RRM2B*	50484
RAD50	10111	*RTEL1*	8771
RAD51	5888	*RUVBL1*	8607

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
RUVBL2	10856	*TDG*	6996
SETD2	29072	*TDP1*	55775
SETMAR	6419	*TELO2*	9894
SETX	23064	*TERF1*	646359
SFPQ	6421	*TERF1*	646127
SHFM1	7979	*TERF1*	283523
SHPRH	257218	*TERF1*	7013
SIRT1	23411	*TERF1*	646316
SIRT6	51548	*TERF2*	7014
SLC30A9	10463	*TERF2IP*	54386
SMAD2	4087	*TEX12*	56158
SMAD3	4088	*TEX15*	56154
SMAD4	4089	*TMEM161A*	54929
SMAD7	4092	*TNP1*	7141
SMARCA1	6594	*TOP1*	7150
SMARCA2	6595	*TOP2A*	7153
SMARCA4	6597	*TOP3A*	7156
SMARCA5	8467	*TOPBP1*	11073
SMARCAD1	56916	*TP53*	7157
SMARCB1	6598	*TP53BP1*	7158
SMARCC2	6601	*TP73*	7161
SMARCD1	6602	*TREX1*	11277
SMARCD2	6603	*TREX2*	11219
SMC1A	8243	*TRIP12*	9320
SMC2	10592	*TRIP13*	9319
SMC3	9126	*TTC5*	91875
SMC4	10051	*TWIST1*	7291
SMC5	23137	*TYMS*	7298
SMC6	79677	*UBA1*	7317
SMG1	23049	*UBA52*	7311
SMUG1	23583	*UBB*	7314
SMURF2	64750	*UBC*	7316
SOD1	6647	*UBE2A*	7319
SP1	6667	*UBE2B*	7320
SPATA22	84690	*UBE2D3*	7323
SPO11	23626	*UBE2D3*	100037280
SPP1	6696	*UBE2I*	7329
SSB	6741	*UBE2N*	7334
SSRP1	6749	*UBE2NL*	389898
STAT1	6772	*UBE2T*	29089
STRA13	201254	*UBE2U*	148581
SUMO1	474338	*UBE2V2*	7336
SUPT16H	11198	*UBE4B*	10277
SUPT16H	400011	*UHRF1*	29128
SYCP1	6847	*UIMC1*	51720
TAOK1	57551	*UNG*	7374
TAOK2	9344	*UPF1*	5976
TAOK3	51347	*USP1*	7398
TCEA1	399511	*USP28*	57646

HGNC Gene Symbol	Entrez Gene ID	HGNC Gene Symbol	Entrez Gene ID
USP3	9960	*XPF*	2072
USP47	55031	*XPG*	2073
USP7	7874	*XRCC1*	7515
UVRAG	7405	*XRCC2*	7516
VCP	7415	*XRCC3*	7517
WDR16	146845	*XRCC4*	7518
WDR33	55339	*XRCC5*	7520
WDR48	57599	*XRCC6*	2547
WEE1	7465	*XRCC6BP1*	91419
WHSC1	7468	*YY1*	7528
WRN	7486	*ZBTB32*	27033
WRNIP1	56897	*ZFYVE26*	23503
WWP1	11059	*ZNF350*	59348
WWP2	11060	*ZRANB3*	84083
XAB2	56949	*ZSWIM7*	125150
XPA	7507		
XPB	2071		
XPC	7508		
XPD	2068		

This list of genes was obtained through literature searches and from the recent study by Smith, *et al.* (Smith, et al. 2016)

(B) Rare SNVs (missense and in-frame insertions/deletions) detected in DNA damage response associated genes.

Function[a]	Gene	Variant[b]	AA change[b]	*BRCAx* cases n=8[c]	*BRCA1/2*$^{+}$ cases n=4[c]	1000G EUR[d]	ESP EA[e]	ExAC EUR [f]
BER	*NEIL2*	c.22C>G	p.P8A	1	-	-	-	0.0000
	DNA2	c.2996G>A	p.R999H	1	-	-	-	0.0000
	NEIL1	c.421G>C	p.E141Q	-	1(*BRCA2)*	-	-	-
	LIG3	c.2861G>A	p.R954H	1	-	-	0.00046	0.0006
DDS	*ATR*	c.2776T>C	p.F926L	-	1(*BRCA2)*	0.00	0.00116	0.0020
	UIMC1	c.1138T>C	p.S380P	1	-	0.00	0.00170	0.0018
	ATM	c.1229T>C	p.V410A	1	-	0.00	0.00221	0.0033
		c.5417T>C	p.I1806T	2	-	-	-	0.0000
	FOXM1	c.425G>A	p.G142E	1	-	0.00	0.00035	0.0005
	TOP2A	c.3041C>T	p.T1014M	1	-	-	0.00010	0.0000
DNAP	*POLG*	c.1550G>T	p.G517V	1	-	0.01	0.00837	0.0070
		c.1174C>G	p.L392V	-	1(*BRCA1)*	0.00	0.00160	0.0020
EPN	*EXO1*	c.809A>T	p.D270V	-	1(*BRCA2)*	-	0.00046	0.0006
	SPO11	c.433A>G	p.R145G	1	-	-	0.00090	0.0007
FA	*FANCC*	c.1081C>T	p.R361W	1	-	-	-	-
		c.632C>G	p.P211R	-	*1(BRCA2)*	0.00	0.00090	0.0011
	BRCA2	c.7828G>A	p.V2610M	1	-	-	-	-
	FANCA	c.2859C>G	p.D953E	-	1(*BRCA2)*	0.00	0.00070	0.0014
HRR	*RAD50*	c.1153C>T	p.R385C	1	-	-	-	0.0000
	MCM9	c.3286A>G	p.M1096V	-	1(*BRCA2)*	-	-	0.0035
		c.1915C>G	p.L639V	1	-	0.00	-	0.0006
		c.302C>T	p.S101L	1	-	-	-	-
	NBN	c.643C>T	p.R215W	1	-	0.00	0.00372	0.0047
	RAD54B	c.1343A>G	p.N448S	-	*1(BRCA2)*	-	-	0.0000
	MUS81	c.896C>T	p.T299M	1	-	0.00	0.00384	0.0029
	BLM	c.2480T>C	p.M827T	2	-	-	-	0.0000
	EME2	c.772C>T	p.P258S	1	-	-	-	0.0000
		c.804_806del	p.S268_269del	-	1(*BRCA2)*	-	-	0.0000

Function[a]	Gene	Variant[b]	AA change[b]	*BRCAx* cases n=8[c]	*BRCA1/2*$^+$ cases n=4[c]	1000G EUR[d]	ESP EA[e]	ExAC EUR [f]
HRR	*BRCA1*	c.4394G>T	p.S1465I	-	1(*BRCA2*)	0.00	0.00395	0.0032
	EME1	c.1636C>T	p.R546C	1	-	0.00	0.00058	0.0003
MMR	*MSH6*	c.2827C>T	p.P943S	-	1(*BRCA2*)	-	0.00090	0.0006
		c.3335G>A	p.R1112H	1	-	-	-	0.0000
		c.3461C>T	p.T1154M	1	-	-	-	0.0000
	MLH1	c.376T>A	p.Y126N	1	-	0.00	-	0.0001
	MSH3	c.187C>G	p.P63A	1	-	-	-	0.0011
		c.2041C>T	p.P681S	1	-	0.00	0.00186	0.0013
NER	*CHD1L*	c.1486G>A	p.G496R	1	-	0.00	0.00546	0.0057
	ERCC3	c.2111C>T	p.S704L	-	1(*BRCA1*)	0.00	0.00256	0.0024
	ERCC4	c.1563C>G	p.S521R	-	1(*BRCA2*)	0.00	0.00050	0.0009
	ERCC2	c.1187G>A	p.S396N	1	-	-	-	0.0000
NHEJ	*WRN*	c.1066A>G	p.K356E	1	-	-	-	-
PARP	*PARP1*	c.1148C>A	p.S383Y	1	-	0.00	0.00314	0.0022
		c.450G>T	p.Q150H	1	-	-	0.00030	0.0004
TLS	*POLQ*	c.7393G>A	p.E2465K	-	1(*BRCA2*)	-	0.00080	0.0005
		c.4635C>A	p.H1545Q	1	-	-	-	0.0000
		c.673C>T	p.H225Y	1	-	-	-	0.0000
	POLN	c.270G>T	p.Q90H	1	-	-	-	-
	POLK	c.85G>A	p.E29K	1	-	0.00	0.00198	0.0025
Other	*SYCP1*	c.260A>G	p.Y87C	1	-	0.00	0.00560	0.0043
	FBXO6	c.151A>G	p.M51V	-	1(*BRCA2*)	0.00	-	0.0020
		c.820C>A	p.Q274K	-	1(*BRCA2*)	0.01	0.00988	0.0097
	IFI16	c.682C>T	p.P228S	2	-	0.01	0.00988	0.0082
		c.1868G>A	p.C623Y	-	1(*BRCA2*)	0.00	0.00546	0.0031
	MDM4	c.56T>G	p.I19S	1	-	-	-	-
	E2F2	c.794C>T	p.T265I	2	-	0.00	0.00314	0.0032
	HDAC1	c.1148C>T	p.A383V	-	1(*BRCA2*)	-	-	0.0000
	WDR33	c.3049G>C	p.D1017H	-	1(*BRCA2*)	-	0.00030	0.0002
	ZRANB3	c.1451A>C	p.K484T	1	-	-	0.00010	-
	SSB	c.1100_1108del	p.DEH367_370del	-	1(*BRCA2*)	0.01	0.00410	0.0047

Function[a]	Gene	Variant[b]	AA change[b]	*BRCAx* cases n=8[c]	*BRCA1/2*$^{+}$ cases n=4[c]	1000G EUR[d]	ESP EA[e]	ExAC EUR [f]
Other	*TNP1*	c.100C>T	p.R34C	1	-	-	-	0.0000
	ATF2	c.803C>T	p.P268L	1	-	0.00	0.00140	0.0013
	IRS1	c.3116C>T	p.A1039V	1	-	-	-	-
	SETD2	c.3601G>C	p.E1201Q	-	1(*BRCA2)*	-	-	0.0000
	JMY	c.134C>G	p.T45S	-	1(*BRCA2)*	0.00	-	0.0007
	WRNIP1	c.757C>T	p.R253C	1	-	0.00	0.00326	0.0047
		c.1882G>A	p.D628N	1	-	-	-	-
	EEPD1	c.95A>G	p.N32S	1	-	-	0.00058	0.0004
	EGFR	c.2051G>A	p.G684D	1	-	-	-	0.0000
	BAZ1B	c.3932C>A	p.S1311Y	1	-	-	-	-
	ESCO2	c.1094G>A	p.R365K	2	-	0.00	0.00302	0.0038
	SMC2	c.3334C>T	p.L1112F	-	1(*BRCA2)*	-	-	-
	CRB2	c.5C>T	p.A2V	1	-	0.00	0.00148	0.0037
	MMS19	c.1522C>T	p.R508W	1	-	-	0.00020	0.0001
	CEP164	c.1220C>T	p.S407F	-	1(*BRCA1)*	-	0.00163	0.0014
		c.3673G>A	p.D1225N	-	1(*BRCA2)*	-	-	0.0000
		c.3931G>T	p.G1311C	-	1(*BRCA2)*	-	-	-
	HINFP	c.1055A>G	p.K352R	1	-	0.01	0.00710	0.0053
	USP47	c.1162T>G	p.S388A	1	-	-	-	0.0000
	RECQL	c.1483G>C	p.D495H	-	1(*BRCA1)*	0.01	0.00488	0.0052
		c.156T>G	p.D52E	1	-	-	0.00010	0.0000
	INO80	c.3323T>G	p.V1108G	-	1(*BRCA2)*	0.01	0.00802	0.0100
		c.3194G>A	p.R1065Q	1	-	0.00	-	0.0001
	TP53BP1	c.895T>C	p.S299P	1	-	0.01	0.00896	0.0076
	CYP1A1	c.1318G>A	p.D440N	-	1(*BRCA2)*	0.01	0.00663	0.0057
	TELO2	c.578T>C	p.V193A	1	-	0.00	0.00523	0.0061
	PLK1	c.781C>T	p.L261F	-	1(*BRCA2)*	0.00	0.00419	0.0031
	CREBBP	c.984G>A	p.M328I	1	-	-	-	-
	WWP2	c.1205G>A	p.R402H	1	-	-	-	-
	CDH13	c.373G>A	p.D125N	1	-	-	-	0.0001
	TOP3A	c.604G>A	p.D202N	1	-	0.00	0.00151	0.0018

Function[a]	Gene	Variant[b]	AA change[b]	*BRCAx* cases n=8[c]	*BRCA1/2*$^+$ cases n=4[c]	1000G EUR[d]	ESP EA[e]	ExAC EUR [f]
Other	*AXIN2*	c.2272G>A	p.A758T	1	-	0.00	0.00372	0.0040
		c.1985T>C	p.L662P	1	-	0.00	0.00170	0.0015
		c.1615G>A	p.V539M	-	1(*BRCA2)*	-	-	0.0001
		c.1250C>T	p.A417V	-	1(*BRCA2)*	-	-	0.0002
	ERBB2	c.1960A>G	p.I654V	1	-	0.01	0.00942	0.0073
	SMURF2	c.1715C>T	p.P572L	-	1(*BRCA2)*	-	-	-
	RECQL5	c.2146G>A	p.V716I	-	1(*BRCA2)*	-	-	0.0000
	POLR2E	c.626T>G	p.V209G	1	-	-	-	0.0033
	DOT1L	c.2560G>A	p.A854T	-	*1(BRCA2)*	-	0.00048	0.0035
	NINL	c.3664C>G	p.L1222V	1	-	0.00	0.00198	0.0048
		c.877G>C	p.G293R	-	1(*BRCA2)*	-	-	0.0000
	NCOA6	c.2794A>C	p.T932P	-	1(*BRCA2)*	-	-	0.0008
	CHAF1B	c.47T>C	p.V16A	1	-	0.01	0.00535	0.0039
	POLA1	c.3035G>A	p.S1012N	-	1(*BRCA2)*	-	-	0.0000
	ATRX	c.3086C>T	p.S1029L	1	-	-	-	-

[a] BER= base excision repair, NER= nucleotide excision repair, MMR= mismatch repair, NHEJ= non-homologous end joining, HRR= homologous recombination repair, DDS= DNA damage signalling, TLS= translesion synthesis, FA= fanconi anemia, EPN= editing and processing nucleases, DNAP= DNA Polymerases, PARP=PARP Enzymes, Other=putative DNA repair genes not classified as part of a specific mechanism

[b] HGVS nomenclature was used

[c] Number of mutation-positive index cases

[d] Minor allele frequencies of European individuals from the 1000 Genomes browser, release date 16.10.2014 (http://browser.1000genomes.org)

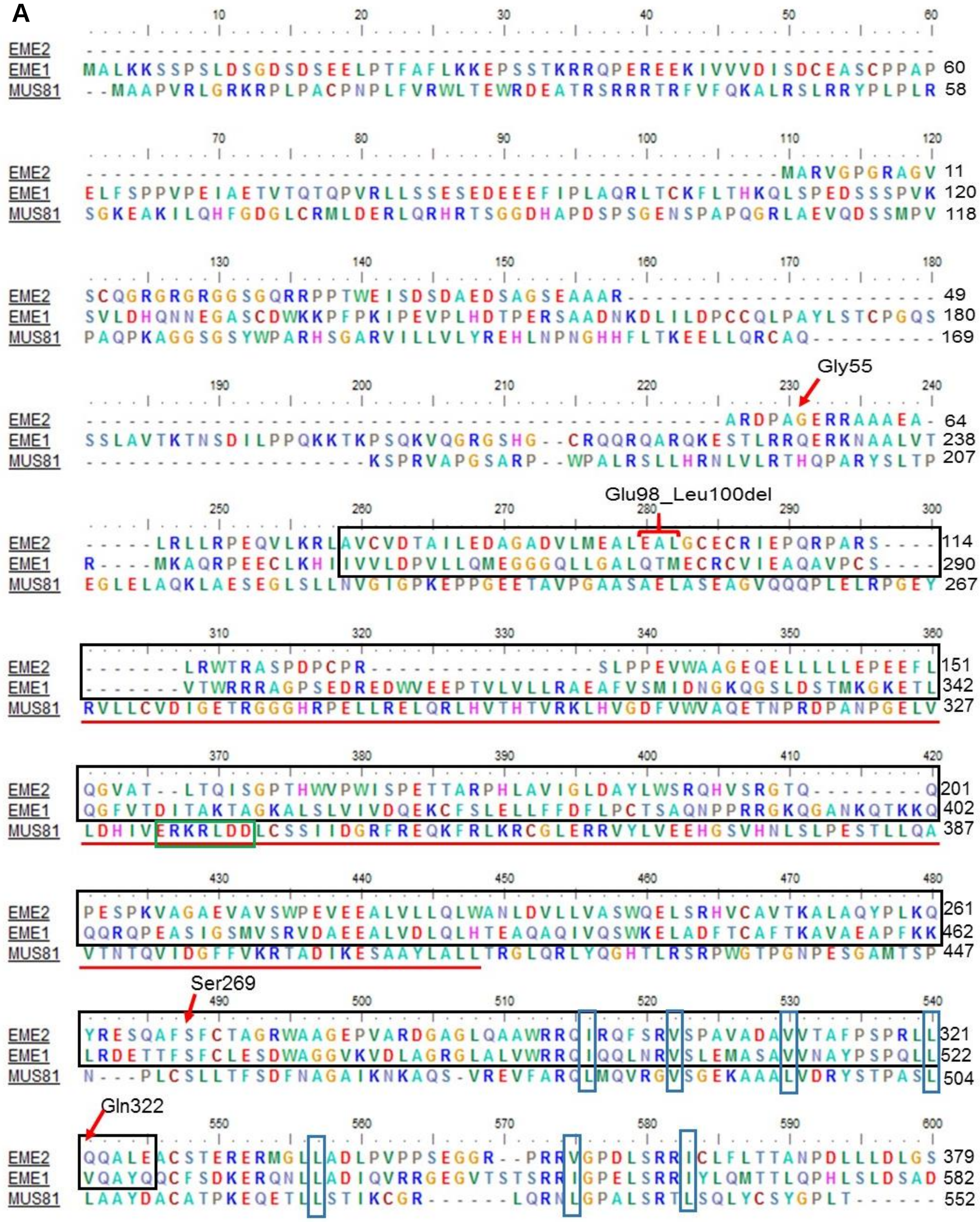
[e] Minor allele frequencies of European-American individuals from the NHLBI Exome sequencing project (ESP) 6500, release version v.0.0.30, date 03.01.2014 (http://evs.gs.washington.edu/EVS/)

[f] The Exome Aggregation Consortium (ExAC), release version v.0.3.1 (http://exac.broadinstitute.org/)

Annexure 6:

(A) - Alignments of EME1, EME2 and MUS81 amino acid sequences.

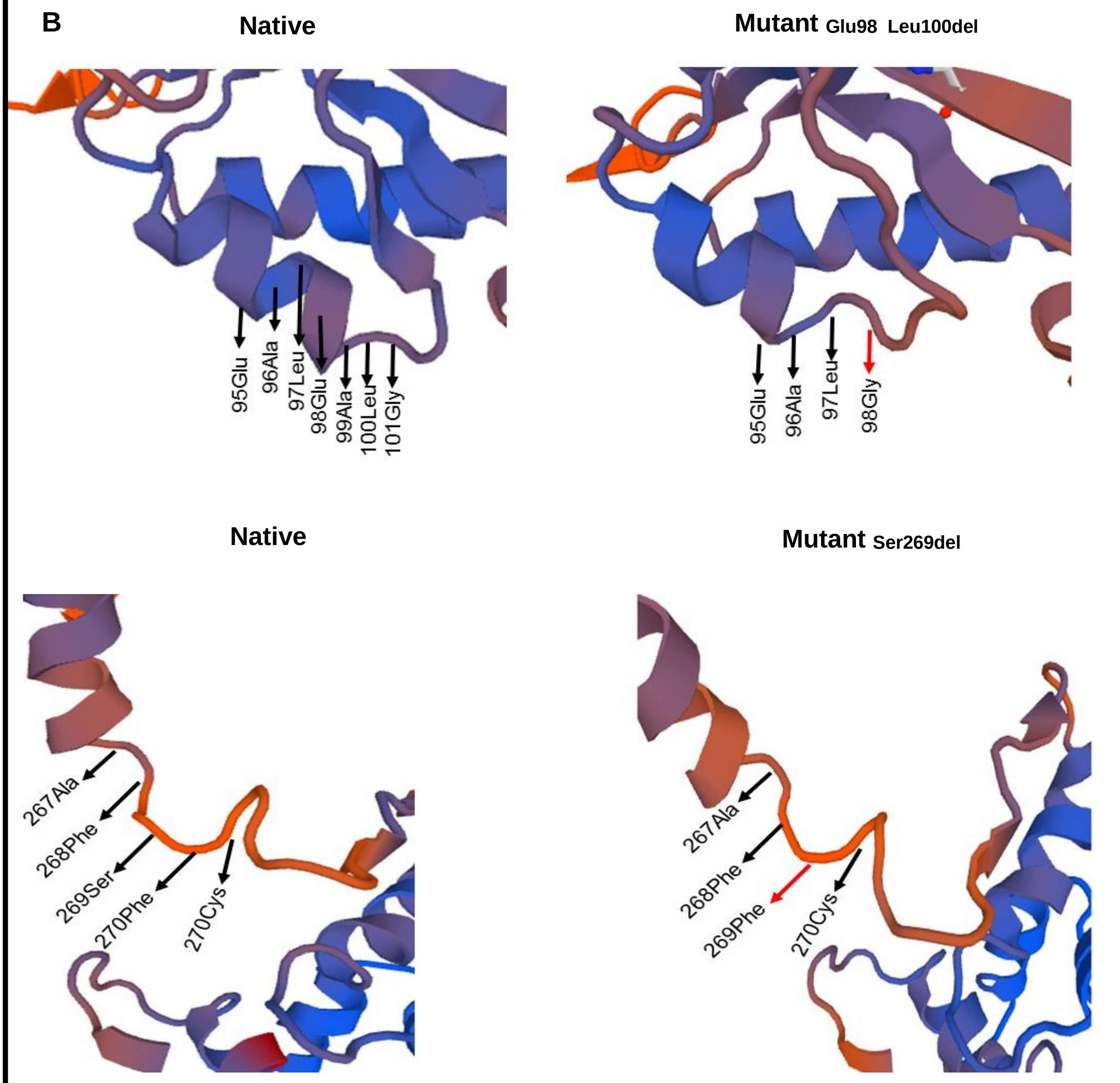
(B) - Partial view of 3D models of *EME2* p.Glu98_Leu100del and p.Ser269del mutations.



```
              10        20        30        40        50        60
EME2   ------------------------------------------------------------
EME1   MALKKSSPSLDSGDSDSEELPTFAFLKKEPSSTKRRQPEREEKIVVVDISDCEASCPPAP 60
MUS81  --MAAPVRLGRKRPLPACPNPLFVRWLTEWRDEATRSRRRTRFVFQKALRSLRRYPLPLR 58

              70        80        90       100       110       120
EME2   -------------------------------------------------MARVGPGRAGV 11
EME1   ELFSPPVPEIAETVTQTQPVRLLSSESEDEEEFIPLAQRLTCKFLTHKQLSPEDSSSPVK 120
MUS81  SGKEAKILQHFGDGLCRMLDERLQRHRTSGGDHAPDSPSGENSPAPQGRLAEVQDSSMPV 118

             130       140       150       160       170       180
EME2   SCQGRGRGRGGSGQRRPPTWEISDSDAEDSAGSEAAAR---------------------- 49
EME1   SVLDHQNNEGASCDWKKPFPKIPEVPLHDTPERSAADNKDLILDPCCQLPAYLSTCPGQS 180
MUS81  PAQPKAGGSGSYWPARHSGARVILLVLYREHLNPNGHHFLTKEELLQRCAQ--------- 169

             190       200       210       220       230       240
EME2   ---------------------------------------------ARDPAGERRAAAEA- 64
EME1   SSLAVTKTNSDILPPQKKTKPSQKVQGRGSHG--CRQQRQARQKESTLRRQERKNAALVT 238
MUS81  --------------------KSPRVAPGSARP--WPALRSLLHRNLVLRTHQPARYSLTP 207

             250       260       270       280       290       300
EME2   -----LRLLRPEQVLKRLAVCVDTAILEDAGADVLMEALEALGCECRIEPQRPARS---- 114
EME1   R----MKAQRPEECLKHIIVVLDPVLLQMEGGGQLLGALQTMECRCVIEAQAVPCS---- 290
MUS81  EGLELAQKLAESEGLSLLNVGIGPKEPPGEETAVPGAASAELASEAGVQQQPLELRPGEY 267

             310       320       330       340       350       360
EME2   -------LRWTRASPDPCPR----------------SLPPEVWAAGEQELLLLLEPEEFL 151
EME1   -------VTWRRRAGPSEDREDWVEEPTVLVLLRAEAFVSMIDNGKQGSLDSTMKGKETL 342
MUS81  RVLLCVDIGETRGGGHRPELLRELQRLHVTHTVRKLHVGDFVWVAQETNPRDPANPGELV 327

             370       380       390       400       410       420
EME2   QGVAT--LTQISGPTHWVPWISPETTARPHLAVIGLDAYLWSRQHVSRGTQ--------Q 201
EME1   QGFVTDITAKTAGKALSLVIVDQEKCFSLELLFFDFLPCTSAQNPPRRGKQGANKQTKKQ 402
MUS81  LDHIVERKRLDDLCSSIIDGRFREQKFRLKRCGLERRVYLVEEHGSVHNLSLPESTLLQA 387

             430       440       450       460       470       480
EME2   PESPKVAGAEVAVSWPEVEEALVLLQLWANLDVLLVASWQELSRHVCAVTKALAQYPLKQ 261
EME1   QQRQPEASIGSMVSRVDAEEALVDLQLHTEAQAQIVQSWKELADFTCAFTKAVAEAPFKK 462
MUS81  VTNTQVIDGFFVKRTADIKESAAYLALLTRGLQRLYQGHTLRSRPWGTPGNPESGAMTSP 447

             490       500       510       520       530       540
EME2   YRESQAFSFCTAGRWAAGEPVARDGAGLQAAWRRQIRQFSRVSPAVADAVVTAFPSPRLL 321
EME1   LRDETTFSFCLESDWAGGVKVDLAGRGLALVWRRQIQQLNRVSLEMASAVVNAYPSPQLL 522
MUS81  N---PLCSLLTFSDFNAGAIKNKAQS-VREVFARQLMQVRGVSGEKAAALVDRYSTPASL 504

             550       560       570       580       590       600
EME2   QQALEACSTERERMGLLADLPVPPSEGGR--PRRVGPDLSRRICLFLTTANPDLLLDLGS 379
EME1   VQAYQQCFSDKERQNLLADIQVRRGEGVTSTSRRIGPELSRRIYLQMTTLQPHLSLDSAD 582
MUS81  LAAYDACATPKEQETLLSTIKCGR------LQRNLGPALSRTLSQLYCSYGPLT------ 552
```

ClustalX v1.82 alignments of EME2, EME1 and MUS81 amino acid sequences.

The locations for all four truncation variants have been indicated. The amino acid region corresponding to the nuclease domain (ERKXXXD) in MUS81 is indicated by a green box. Inactive ERCC4 regions for both EME2, EME1 (black box) and the active MUS81 (red line) are shown as well. Blue boxes have been drawn around important hydrophobic residues that form part of the c-terminal $(HhH)_2$ domains.



3D models of the wild type EME2 protein sequence was constructed using SWISS-MODEL.

The site of in-frame variants p.Glu98_Leu100del in an alpha-helix structure and p.Ser269del located in a coiled-coil region have been emphasized (Biasini, et al. 2014).

Annexure 7:

Conference presentations and manuscript submission.

CONFERENCE PRESENTATIONS.

National Conferences

Mentoor, J.L.D., Joubert, F and van Rensburg, E. J. Whole exome sequencing of South African breast cancer families - Preliminary Results (Poster Presentation) – 15th Biennial Congress of the Southern African Society for Human Genetics , The Maslow Hotel, Sandton, Johannesburg, South Africa. 6-9 October 2013.

Mentoor, J.L.D., Joubert, F and van Rensburg, E. J. Whole exome sequencing of selected *BRCAx* South African high-risk breast cancer families: Variants in genes involved in maintenance of genomic stability (Poster presentation) – 16th Biennial Congress of the Southern African Society for Human Genetics (SASHG), KleinKaap Boutique Hotel, Centurion, Pretoria, South Africa. 16-19 August 2015.

International Conferences

Mentoor, J.L.D., Joubert, F and van Rensburg, E. J. Exome and directed analysis of South African breast cancer patients (Poster presentation) - EMBL Cancer Genomics Conference, EMBL Heidelberg, Germany. 3-5 November 2013.

Mentoor, J.L.D., Joubert, F and van Rensburg, E. J. Whole exome sequencing of South African breast cancer families (Poster Presentation) – Functional genomics and Systems biology 2013 Conference, The Welcome Trust Conference Centre, Hinxton, Cambridge, UK. 21-23 November 2013.

Mentoor, J.L.D., Joubert, F and van Rensburg, E. J. Whole exome sequencing reveals that DNA repair and apoptosis pathways are affected in hereditary breast cancer cases (Poster presentation) - 64th Annual Meeting of the American Society of Human Genetics, San Diego Convention Center (SDCC), San Diego, California. 18-22 October 2014.

MANUSCRIPT SUBMISSION.

The research described in Chapter 6 is detailed in a manuscript that has been submitted for publication to the journal : *Familial Cancer*Manuscript number: FAME-D-17-00122

Mentoor J., Dorfling CM, Joubert F, and van Rensburg EJ. *HELQ* sequence variants in high-risk non-*BRCA1/2* South-African breast cancer families

References

1. Align-Grantham Variation and Grantham Deviation (A-GVGD). Available from: http://agvgd.iarc.fr/agvgd_input.php. Web. 30 Nov. 2016.
2. Breast Cancer Association Consortium (BCAC). Available from: http://apps.ccge.medschl.cam.ac.uk/consortia/bcac/. Web. 30 Nov. 2016.
3. National Cancer Registry (2004). Cancer Statistics. Retrieved 2012.
4. National Cancer Registry (2009). Cancer Statistics. Retrieved 2015.
5. National Cancer Registry (2011). Cancer Statistics. Retrieved 2016.
6. The National Human Genome Research Institute (NHGRI): Breast Cancer Information Core (BIC) database. Available from: https://research.nhgri.nih.gov/projects/bic/. Web. 30 Nov. 2016.
7. Polymorphism Phenotyping v2 (PolyPhen-2). Available from: http://genetics.bwh.harvard.edu/pph2/. Web. 30 Nov. 2016.
8. Sorting Intolerant from Tolerant (SIFT). Available from: http://sift.bii.a-star.edu.sg/www/SIFT_BLink_submit.html. Web. 30 Nov. 2016.
9. New Putative Functional Polymorphisms at the 4q21 Locus Associated With Modification of Breast and Ovarian Cancer Risk in *BRCA2* Mutation Carriers. Fourth International Symposium on Hereditary Breast and Ovarian Cancer. 2012 Montreal, Quebec, Canada.
10. The National Center for Biotechnology Information (NCBI). U.S National Library of Medicine; 2015. Available from: http://www.ncbi.nlm.nih.gov/. Web. 21 Jul. 2017.
11. Abiatari I, Gillen S, DeOliveira T, et al. 2009. The microtubule-associated protein MAPRE2 is involved in perineural invasion of pancreatic cancer cells. *J Oncol* 35: p. 1111-1116.
12. Adelman CA, Boulton SJ 2010. Metabolism of postsynaptic recombination intermediates. *FEBS Letters* 584: p. 3709-3716.
13. Adelman CA, Lolo RL, Birkbak NJ, et al. 2013. HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. *Nature* 502: p. 381-384.
14. Adzhubei IA, Bork P, Gerasimova A, et al. 2010. A method and server for predicting damaging missense mutations. *Nature Methods* 7: p. 248-249.
15. Alberg AJ, Jorgensen TJ, Ruczinski I, et al. 2013. DNA repair gene variants in relation to overall cancer risk: a population-based study. *Carcinogenesis* 34: p. 86-92.
16. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* 500: p. 415-421.
17. Alexandrov Ludmil B, Nik-Zainal S, Wedge David C, et al. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* 3: p. 246-259.
18. Alexandrov LB, Stratton MR 2014. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr Opin Genetics Dev* 24: p. 52-60.
19. Altman DG. 1991. Practical statistics for medical research Chapman and Hall, London, 10: p. 1635-1636.

20. Amangyeld T, Shin Y-K, Lee M, et al. 2014. Human MUS81-EME2 can cleave a variety of DNA structures including intact Holliday junction and nicked duplex. *Nucl. Acids Res.* 42: p. 5846-5862.
21. Amemiya Y, Bacopulos S, Al-Shawarby M, et al. 2015. A comparative analysis of breast and ovarian cancer-related gene mutations in canadian and Saudi Arabian patients with breast cancer. *Anticancer Res* 35: p. 2601-2610.
22. Antoniou AC, Casadei S, Heikkinen T, et al. 2014. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med* 371: p. 497-506.
23. Apostolou P, Fostira F 2013. Hereditary breast cancer: the era of new susceptibility genes. *Biomed Res Int* 2013: p. 747318.
24. Arcand SL, Akbari MR, Mes-Masson A-M, et al. 2015. Germline TP53 mutational spectrum in French Canadians with breast cancer. *BMC Med Genet* 16: p. 1-11.
25. Arens J, Duong T-T, Dehmelt L 2013. A morphometric screen identifies specific roles for microtubule-regulating genes in neuronal development of P19 stem cells. *PLoS ONE* 8: p. e79796.
26. Autier P, Boniol M, Gavin A, et al. 2011. Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. *BMJ* 343: p. d4411.
27. Badve S, Dabbs DJ, Schnitt SJ, et al. 2011. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Mod Pathol* 24: p. 157-167.
28. Ball MP, Thakuria JV, Zaranek AW, et al. 2012. A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences* 109: p. 11920-11927.
29. Bamshad MJ, Ng SB, Bigham AW, et al. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: p. 745-755.
30. Barlow JH, Faryabi Robert B, Callén E, et al. 2013. Identification of early replicating fragile sites that contribute to genome instability. *Cell* 152: p. 620-632.
31. Batra J 2009. Biophysical Studies of Protein Folding and Binding Stability. Florida State University.
32. Beggs AD, Latchford AR, Vasen HFA, et al. 2010. Peutz–Jeghers syndrome: a systematic review and recommendations for management. *Gut* 59: p. 975-986.
33. Biasini M, Bienert S, Waterhouse A, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucl. Acids Res.* 42: p. W252-W258.
34. Blay P, Santamaria I, Pitiot AS, et al. 2013. Mutational analysis of BRCA1 and BRCA2 in hereditary breast and ovarian cancer families from Asturias (Northern Spain). *BMC Cancer* 13: p. 243.
35. Blute ML, Cerhan JR, Cunningham JM, et al. 2003. No association of germline alteration of MSR1 with prostate cancer risk. *Nat Genet* 35: p. 128 - 129.
36. Bodian DL, McCutcheon JN, Kothiyal P, et al. 2014. Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: Implications for individual genome sequencing. *PLoS ONE* 9: p. e94554.

37. Bodmer W, Tomlinson I 2010. Rare genetic variants and the risk of cancer. *Curr Opin Genetics Dev* 20: p. 262-267.

38. Bogdanova N, Helbig S, Dork T 2013. Hereditary breast cancer: ever more pieces to the polygenic puzzle. *Hereditary Cancer Clin Pract* 11: p. 12.

39. Bonifaci N, Berenguer A, Diez J, et al. 2008. Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes. *BMC Med Genomics* 1: p. 62.

40. Borg Å, Haile RW, Malone KE, et al. 2010. Characterization of BRCA1 and BRCA2 deleterious mutations and variants of unknown clinical significance in unilateral and bilateral breast cancer: the wecare study. *Hum. Mutat.* 31: p. E1200-E1240.

41. Bosch A, Eroles P, Zaragoza R, et al. 2010. Triple-negative breast cancer: Molecular features, pathogenesis, treatment and current lines of research. *Cancer Treat Rev* 36: p. 206-215.

42. Boyd J 2014. Genetic predisposition to breast cancer: The next chapters. *Cancer* 120: p. 932-934.

43. Brandt-Rauf P, Li Y, Long C, et al. 2013. The molecular epidemiology of DNA repair polymorphisms in carcinogenesis. *J Biomed Res* 27: p. 179–192.

44. Brentnall M, Rodriguez-Menocal L, De Guevara R, et al. 2013. Caspase-9, caspase-3 and caspase-7 have distinct roles during intrinsic apoptosis. *BMC Cell Biology* 14: p. 32.

45. Broeks A, Braaf L, Huseinovic A, et al. 2008. The spectrum of ATM missense variants and their contribution to contralateral breast cancer. *Breast Cancer Res Treat* 107: p. 243-248.

46. Broeks A, Schmidt MK, Sherman ME, et al. 2011. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet* 20: p. 3289-3303.

47. Brosh J, Robert M 2013. DNA helicases involved in DNA repair and their roles in cancer. *Nat Rev Cancer* 13: p. 542-558.

48. Broustas CG, Lieberman HB 2014. DNA damage response genes and the development of cancer metastasis. *Radiat. Res* 181: p. 111-130.

49. Casadei S, Norquist BM, Walsh T, et al. 2011. Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. *Cancer Res* 71: p. 2222-2229.

50. Cerami E, Gao J, Dogrusoz U, et al. 2012. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2: p. 401-404.

51. Cerqua C, Anesti V, Pyakurel A, et al. 2010. Trichoplein/mitostatin regulates endoplasmic reticulum–mitochondria juxtaposition. *EMBO reports* 11: p. 854-860.

52. Challis D, Yu J, Evani U, et al. 2012. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* 13: p. 8.

53. Chandler MR, Bilgili EP, Merner ND 2016. A Review of Whole-Exome Sequencing Efforts Toward Hereditary Breast Cancer Susceptibility Gene Discovery. *Human Mutation* 37: p. 835-846.

54. Chang X, Wang K 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet.* 49: p. 433–436.
55. Chen F, Dong M, Ge M, et al. 2013. The history and advances of reversible terminators used in new generations of sequencing technology. *Genomics, Proteomics & Bioinformatics* 11: p. 34-40.
56. Chen W, Moore MJ 2014. The spliceosome: disorder and dynamics defined. *Curr. Opin. Struct. Biol.* 24: p. 141-149.
57. Chong Jessica X, Ouwenga R, Anderson Rebecca L, et al. 2012. A population-based study of autosomal-recessive disease-causing mutations in a founder population. *Am J Hum Genet* 91: p. 608-620.
58. Chow AY 2010. Cell cycle control by oncogenes and tumor suppressors: driving the transformation of normal cells into cancerous cells. *Nature Education* 3: p. 7.
59. Chun J, Buechelmaier ES, Powell SN 2013. Rad51 paralog complexes BCDX2 and CX3 act at different stages in the BRCA1-BRCA2-dependent homologous recombination pathway. *Mol. Cell. Biol* 33: p. 387-395.
60. Ciccia A, Ling C, Coulthard R, et al. 2007. Identification of FAAP24, a Fanconi Anemia Core Complex Protein that Interacts with FANCM. *Molecular Cell* 25: p. 331-343.
61. Clauson C, Schärer OD, Niedernhofer L 2013. Advances in understanding the complex mechanisms of DNA interstrand cross-link repair. *Cold Spring Harb Perspect Biol* 5: p. a012732.
62. Cline MS, Craft B, Swatloski T, et al. 2013. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci. Rep.* 3: p. 2652.
63. Collins A, Politopoulos I 2011. The genetics of breast cancer: risk factors for disease. *Appl Clin Genet.* 4: p. 11-19.
64. Complexo, Southey M, Park D, et al. 2013. COMPLEXO: identifying the missing heritability of breast cancer via next generation collaboration. *Breast Cancer Research* 15: p. 402.
65. Consortium TGP 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: p. 1061-1073.
66. UniProt: A hub for protein information. 2015 Release 43. Available from http://nar.oxfordjournals.org/content/43/D1/D204.abstract. Web. 30 Nov 2016.
67. Couch FJ, Nathanson KL, Offit K 2014. Two decades after BRCA: Setting paradigms in personalized cancer care and prevention. *Science* 343: p. 1466-1470.
68. Croteau DL, Popuri V, Opresko PL, et al. 2014. Human RECQ helicases in DNA repair, recombination, and replication. *Annu. Rev. Biochem* 83: p. 519-552.
69. Curtis C, Shah SP, Chin S-F, et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486: p. 346-352.
70. Cybulski C, Carrot-Zhang J, Kluzniak W, et al. 2015. Germline RECQL mutations are associated with breast cancer susceptibility. *Nat Genet* advance online publication.
71. D'Andrea AD 2010. Susceptibility pathways in fanconi's anemia and breast cancer. *N Engl J Med* 362: p. 1909-1919.
72. Deans AJ, West SC 2011. DNA interstrand crosslink repair and cancer. *Nat Rev Cancer* 11: p. 467-480.

73. Delbridge ARD, Valente LJ, Strasser A 2012. The role of the apoptotic machinery in tumor suppression. *Cold Spring Harb Perspect Biol* 4: p. 1-14.

74. DePristo MA, Banks E, Poplin R, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: p. 491-498.

75. DeSantis C, Ma J, Bryan L, et al. 2014. Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians* 64: p. 52-62.

76. Desmet F-O, Hamroun D, Lalande M, et al. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucl. Acids Res.* 37: p. e67-e67.

77. Dexheimer TS. 2013. DNA repair pathways and mechanisms. In: Mathews LA, et al., editors. DNA repair of cancer stem cells. USA: Springer science and bussiness media Dordrecht. 178: p. 10.

78. di Masi A, Gullotta F, Cappadonna V, et al. 2011. Cancer predisposing mutations in BRCT domains. *IUBMB Life* 63: p. 503-512.

79. Di Tommaso P, Moretti S, Xenarios I, et al. 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucl. Acids Res.* 39: p. W13-W17.

80. Dietlein F, Reinhardt HC 2014. Molecular pathways: exploiting tumor-specific molecular defects in DNA repair pathways for precision cancer therapy. *Clin Cancer Res* 20: p. 5882-5887.

81. Ding L, Wendl MC, Koboldt DC, et al. 2010. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum. Mol. Genet* 19: p. R188-R196.

82. Dong X, Wang L, Taniguchi K, et al. 2003. Mutations in CHEK2 associated with prostate cancer risk. *Am J Hum Genet* 72: p. 270-280.

83. Dorfling CM, Eloff J, J van Rensburg E. 2016 *RAD51C* and *RAD51D* mutation screening in South African breast and ovarian cancer families. (Unpublished).

84. Dunnen JT, Antonarakis SE 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum. Mutat.* 15: p. 7-12.

85. Easton DF 1999. How many more breast cancer predisposition genes are there? *Breast Cancer Res* 1: p. 14 - 17.

86. Eccles S, Aboagye E, Ali S, et al. 2013. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Research* 15: p. R92.

87. Economopoulou P, Mountzios G, Kotsantis I, et al. 2013. The role of genetic instability in familial cancer syndromes. *J Genet Syndr Gene Ther* 4: p. 1-7.

88. Ellsworth RE, Henning JD, Oakes N, et al. 2015. Abstract P4-10-01: Evaluation of the role of EBV in breast cancer. *Cancer Res* 75: p. P4-10-01-P14-10-01.

89. Emmert-Streib F, De Matos Simoes R, Mullan P, et al. 2014. The gene regulatory network for breast cancer: Integrated regulatory landscape of cancer hallmarks. *Front Genet* 5: p. 1-12.

90. Enderling H, Hahnfeldt P 2011. Cancer stem cells in solid tumors: Is ‘evading apoptosis’ a hallmark of cancer? *Prog Biophys Mol Biol* 106: p. 391-399.

91. Ernani FP, LeProust EM. 2009. Agilent's SureSelect Target Enrichment System: Bringing Cost and Process Efficiency to Next-Generation Sequencing. In. U.S.A: Agilent technologies.

92. Eroles P, Bosch A, Alejandro Pérez-Fidalgo J, et al. 2012. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treat Rev* 38: p. 698-707.

93. Farooq A, Walker LJ, Bowling J, et al. 2012. Cowden syndrome. *Cancer Treat Rev* 36: p. 577-583.

94. Farooq U, Joshi M, Nookala V, et al. 2010. Self-reported exposure to pesticides in residential settings and risk of breast cancer: a case-control study. *Environ. Health* 9: p. 1-9.

95. Fassan M, D'Arca D, Letko J, et al. 2011. Mitostatin is down-regulated in human prostate cancer and suppresses the invasive phenotype of prostate cancer cells. *PLoS ONE* 6: p. e19771.

96. Feng B-J, Tavtigian SV, Southey MC, et al. 2011. Design Considerations for Massively Parallel Sequencing Studies of Complex Human Disease. *PLoS ONE* 6: p. e23221.

97. Feng S, Pistis G, Zhang H, et al. 2015. Methods for association analysis and meta-analysis of rare variants in families. *Genet Epidemiol* 39: p. 227-238.

98. Ferlay J, Soerjomataram I, Dikshit R, et al. 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136: p. E359-E386.

99. Fernandez-Cuesta L, Oakman C, Falagan-Lotsch P, et al. 2012. Prognostic and predictive value of TP53 mutations in node-positive breast cancer patients treated with anthracycline- or anthracycline/taxane-based adjuvant therapy: results from the BIG 02-98 phase III trial. *Breast Cancer Research* 14: p. R70.

100. Ferro R, Parvathaneni, A., Patel, S. and Cheriyath, P 2012. Pesticides and breast cancer. *Adv Breast Cancer Res* 1: p. 30-35.

101. Filippini S, Vega A 2013. Breast cancer genes: beyond BRCA1 and BRCA2. *Front Biosci (Landmark Ed)* 18: p. 1358-1372.

102. Fischer M, Snajder R, Pabinger S, et al. 2012. SIMPLEX: Cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. *PLoS ONE* 7: p. e41948.

103. Fletcher O, Houlston RS 2010. Architecture of inherited susceptibility to common cancer. *Nat Rev Cancer* 10: p. 353-361.

104. Fletcher O, Johnson N, dos Santos Silva I, et al. 2010. Missense variants in ATM in 26,101 breast cancer cases and 29,842 controls. *Cancer Epidem Biomar* 19: p. 2143-2151.

105. Floor SL, Dumont JE, Maenhaut C, et al. 2012. Hallmarks of cancer: of all cancer cells, all the time? *Trends Mol Med* 18: p. 509-515.

106. Foley SB, Rios JJ, Mgbemena VE, et al. 2015. Use of whole genome sequencing for diagnosis and discovery in the cancer genetics clinic. *EBioMedicine* 2: p. 74-81.

107. Forbes SA, Bindal N, Bamford S, et al. 2011. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucl. Acids Res.* 39: p. D945-D950.

108. Foulkes WD 2008. Inherited susceptibility to common cancers. *N Engl J Med* 359: p. 2143-2153.

109. Foulkes WD, Smith IE, Reis-Filho JS 2010. Triple-negative breast cancer. *N Engl J Med* 363: p. 1938-1948.

110. Fu W, O'Connor TD, Jun G, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493: p. 4.

111. Fukui K 2010. DNA Mismatch Repair in Eukaryotes and Bacteria. *J Nucleic Acids.* 2010: p. 1-16.

112. Gao J, Aksoy BA, Dogrusoz U, et al. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.* 6: p. pl1-pl1.

113. García-Alcalde F, Okonechnikov K, Carbonell J, et al. 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 28: p. 2678-2679.

114. Garner E, Smogorzewska A 2011. Ubiquitylation and the Fanconi anemia pathway. *FEBS Letters* 585: p. 2853-2860.

115. Garte S, Gaspari L, Alexandrie A-K, et al. 2001. Metabolic gene polymorphism frequencies in control populations. *Cancer Epidemiol Biomarkers Prev* 10: p. 1239-1248.

116. Gasco M, Shami S, Crook T 2002. The p53 pathway in breast cancer. *Breast Cancer Res* 4: p. 70 - 76.

117. Gerhardus A, Schleberger H, Schlegelberger B, et al. 2007. Diagnostic accuracy of methods for the detection of BRCA1 and BRCA2 mutations: a systematic review. *Eur J Hum Genet* 15: p. 619-627.

118. Ghoussaini M, Pharoah PDP, Easton DF 2013. Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? *Am. J. Pathol* 183: p. 1038-1051.

119. Gibson G 2012. Rare and common variants: twenty arguments. *Nat Rev Genet* 13: p. 135-145.

120. Goecks J, Nekrutenko A, Taylor J, et al. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11: p. R86.

121. Goldspink DA, Gadsby JR, Bellett G, et al. 2013. The microtubule end-binding protein EB2 is a central regulator of microtubule reorganisation in apico-basal epithelial differentiation. *J. Cell Sci* 126: p. 4000-4014.

122. Goodman MF, Woodgate R 2013. Translesion DNA polymerases. *Cold Spring Harb Perspect Biol* 5: p. a010363.

123. Goyal G, Fan T, Silberstein PT 2016. Hereditary cancer syndromes: utilizing DNA repair deficiency as therapeutic target. *Familial Cancer* 15: p. 359-366.

124. Grabarz A, Barascu A, Guirouilh-Barbat J, et al. 2012. Initiation of DNA double strand break repair: signaling and single-stranded resection dictates the choice between homologous recombination, non-homologous end-joining and alternative end-joining. *Am J Cancer Res* 2: p. 249-268.

125. Gracia-Aznarez FJ, Fernandez V, Pita G, et al. 2013. Whole exome sequencing suggests much of non-*BRCA1/BRCA2* familial breast cancer is due to moderate and low penetrance susceptibility alleles. *PLoS ONE* 8: p. e55681.
126. Guler GD, Liu H, Vaithiyalingam S, et al. 2012. Human DNA helicase b (HDHB) binds to replication protein a and facilitates cellular recovery from replication stress. *J. Biol. Chem.* 287: p. 6469-6481.
127. Hall MJ, Reid JE, Burbidge LA, et al. 2009. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer* 115: p. 2222-2233.
128. Hamdi Y, Soucy P, Adoue V, et al. 2016. Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. *Oncotarget* 7: p. 80140-80163.
129. Hamdi Y, Soucy P, Goldgar D New putative functional polymorphisms at the 4q21 locus associated with modification of breast and ovarian cancer risk in *BRCA2* mutation carriers. The Fourth International Symposium on Hereditary Breast and Ovarian Cancer. 2012 Montreal, Quebec, Canada.
130. Hammer GD, Else T, Zambetti G, et al. 2011. TP53 Molecular Genetics. In. Adrenocortical Carcinoma: Springer New York: p. 193-205.
131. Hanahan D, Weinberg R 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144: p. 646-674.
132. Hansford S, Kaurah P, Li-Chang H, et al. 2015. Hereditary diffuse gastric cancer syndrome: Cdh1 mutations and beyond. *JAMA Oncol.* 1: p. 23-32.
133. Hedges DJ, Guettouche T, Yang S, et al. 2011. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* 6: p. e18595.
134. Hellebrand H, Sutter C, Honisch E, et al. 2011. Germline mutations in the PALB2 gene are population specific and occur with low frequencies in familial breast cancer. *Hum. Mutat.* 32: p. E2176-E2188.
135. Helleday T, Lo J, van Gent DC, et al. 2007. DNA double-strand break repair: From mechanistic understanding to cancer treatment. *DNA Repair* 6: p. 923-935.
136. Henzler Wildman KA, Lee D-K, Ramamoorthy A 2002. Determination of α-helix and β-sheet stability in the solid state: A solid-state NMR investigation of poly(L-alanine). *Biopolymers* 64: p. 246-254.
137. Hicks S, Wheeler DA, Plon SE, et al. 2011. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.* 32: p. 661-668.
138. Hilbers FS, Meijers CM, Laros JFJ, et al. 2013. Exome sequencing of germline DNA from non-*BRCA1/2* familial breast cancer cases selected on the basis of aCGH tumor profiling. *PLoS ONE* 8: p. e55734.
139. Hollestelle A, Nagel J, Smid M, et al. 2010. Distinct gene mutation profiles among luminal-type and basal-type breast cancer cell lines. *Breast Cancer Res Treat* 121: p. 53-64.
140. Holthausen JT, Wyman C, Kanaar R 2010. Regulation of DNA strand exchange in homologous recombination. *DNA Repair* 9: p. 1264-1272.

141. Houdayer C, Caux-Moncoutier V, Krieger S, et al. 2012. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined *in silico/in vitro* studies on *BRCA1* and *BRCA2* variants. *Hum. Mutat.* 33: p. 1228-1238.

142. Hucl T, Gallmeier E 2011. DNA repair: Exploiting the fanconi anemia pathway as a potential therapeutic target. *Physiol. Res.* 60: p. 453-465.

143. Ibi M, Zou P, Inoko A, et al. 2011. Trichoplein controls microtubule anchoring at the centrosome by binding to Odf2 and ninein. *J. Cell Sci* 124: p. 857-864.

144. Inoko A, Matsuyama M, Goto H, et al. 2012. Trichoplein and Aurora A block aberrant primary cilia assembly in proliferating cells. *J. Cell Biol.* 197: p. 391-405.

145. Ishikawa T, Toyoda Y, Yoshiura K-i, et al. 2012. Pharmacogenetics of human ABC transporter ABCC11: new insights into apocrine gland growth and metabolite secretion. *Front Genet.* 3: p. 306.

146. J van Rensburg E, van der Merwe N, Sluiter M, et al. Impact of the BRCA-genes on the burden of familial breast/ovarian cancer in South Africa. [Abstract 382]. Presented at the annual meeting of the American Society of Human Genetics. 2007 San Diego, California.

147. Janavičius R 2010. Founder BRCA1/2 mutations in the Europe: implications for hereditary breast-ovarian cancer prevention and control. *The EPMA Journal* 1: p. 397-412.

148. Jardines L, Goyal S, Fisher P, et al. 2011. Breast cancer overview: Risk factors, screening, genetic testing, and prevention. *Oncology* 14: p. 1-23.

149. Jasin M, Rothstein R 2013. Repair of strand breaks by homologous recombination. *Cold Spring Harb Perspect Biol* 5: p. a012740.

150. Jemal A, Bray F, Forman D, et al. 2012. Cancer burden in Africa and opportunities for prevention. *Cancer* 118: p. 4372-4384.

151. Jemal A, Center MM, DeSantis C, et al. 2010. Global Patterns of Cancer Incidence and Mortality Rates and Trends. *Cancer Epidemiol Biomarkers Prev* 19: p. 1893-1907.

152. Ji H 2012. Improving bioinformatic pipelines for exome variant calling. *Genome Med* 4: p. 7.

153. Johansson S, Irgens H, Chudasama KK, et al. 2012. Exome sequencing and genetic testing for MODY. *PLoS ONE* 7: p. e38050.

154. Johns Jr MB, Paulus-Thomas JE 1989. Purification of human genomic DNA from whole blood using sodium perchlorate in place of phenol. *Anal. Chem.* 180: p. 276-278.

155. Johnson KC, Miller AB, Collishaw NE, et al. 2011. Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009). *Tob. Control.* 20: p. e2.

156. Johnson N, Fletcher O, Palles C, et al. 2007. Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. *Hum. Mol. Genet* 16: p. 1051-1057.

157. Joosse SA 2012. BRCA1 and BRCA2: a common pathway of genome protection but different breast cancer subtypes. *Nat Rev Cancer* 12: p. 372-372.

158. Karami F, Mehdipour P 2013. A comprehensive focus on global spectrum of *BRCA1* and *BRCA2* mutations in breast cancer. *BioMed Res. Int* 2013: p. 21.

159. Kast K, Rhiem K, Wappenschmidt B, et al. 2016. Prevalence of *BRCA1/2* germline mutations in 21 401 families with breast and ovarian cancer. *Journal of Medical Genetics* 53: p. 465-471.

160. Kee Y, D'Andrea AD 2010. Expanded roles of the Fanconi anemia pathway in preserving genomic stability. *Genes Dev* 24: p. 1680-1694.

161. Khan I, Sommers JA, Brosh Jr RM 2015. Close encounters for the first time: Helicase interactions with DNA damage. *DNA Repair* 33: p. 43-59.

162. Kiiski JI, Pelttari LM, Khan S, et al. 2014. Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer. *Proceedings of the National Academy of Sciences* 111: p. 15172-15177.

163. Kim H, D'Andrea AD 2012. Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. *Genes Dev* 26: p. 1393-1408.

164. Kim Y, Wilson D 2012. Overview of base excision repair biochemistry. *Curr Mol Pharmacol* 5: p. 3-13.

165. Kim YR, Kim SS, Yoo NJ, et al. 2010. Mutational analysis of MITOSTATIN, a candidate tumor-suppressor gene, at a mononucleotide repeat in gastric and colorectal carcinoma. *Gut Liver* 4: p. 149-150.

166. Kircher M, Kelso J 2010. High-throughput DNA sequencing – concepts and limitations. *BioEssays* 32: p. 524-536.

167. Kircher M, Witten DM, Jain P, et al. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: p. 310-315.

168. Kobayashi H, Ohno S, Sasaki Y, et al. 2013. Hereditary breast and ovarian cancer susceptibility genes (Review) *Oncology Rep* 30: p. 1019-1029.

169. Koboldt Daniel C, Steinberg Karyn M, Larson David E, et al. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* 155: p. 27-38.

170. Korde LA, Shiovitz SA 2013. High-, moderate-, and low-penetrance genes involved in the pathogenesis of a hereditary predisposition to breast cancer. p. 1-6.

171. Kottemann MC, Smogorzewska A 2013. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. *Nature* 493: p. 356-363.

172. Krumm N, Sudmant PH, Ko A, et al. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 22: p. 1525–1532.

173. Kunz C, Focke F, Saito Y, et al. 2009. Base excision by thymine DNA glycosylase mediates DNA-directed cytotoxicity of 5-fluorouracil. *PLoS Biol* 7: p. e1000091.

174. Kwei KA, Kung Y, Salari K, et al. 2010. Genomic instability in breast cancer: Pathogenesis and clinical implications. *Molecular Oncology* 4: p. 255-266.

175. Landau-Ossondo M, Rabia N, Jos-Pelage J, et al. 2009. Why pesticides could be a common cause of prostate and breast cancers in the French Caribbean Island, Martinique. An overview on key mechanisms of pesticide-induced cancer. *Biomed Pharmacother* 63: p. 383-395.

176. Landrum MJ, Lee JM, Benson M, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucl. Acids Res.* 44: p. D862-D868.

177. Landrum MJ, Lee JM, Riley GR, et al. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucl. Acids Res.* 42: p. D980-D985.

178. Lange SS, Takata K-i, Wood RD 2011. DNA polymerases and cancer. *Nat Rev Cancer* 11: p. 96.

179. Latif A, Hadfield KD, Roberts SA, et al. 2010. Breast cancer susceptibility variants alter risks in familial disease. *J Med Genet.* 47: p. 126-131.

180. Laura B 2010. Whole-genome sequencing breaks the cost barrier. *Cell* 141: p. 917-919.

181. Lawson J. 2009. Do Viruses Cause Breast Cancer? In: Verma M, editor. Cancer Epidemiology: *Humana Press*. 471: p. 421-438.

182. Le Guen T, Ragu S, Guirouilh-Barbat J, et al. 2014. Role of the double-strand break repair pathway in the maintenance of genomic stability. *Mol Cell Oncol* 2: p. e968020.

183. Lee AJ, Cunningham AP, Kuchenbaecker KB, et al. 2014a. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *British Journal of Cancer* 110: p. 535-545.

184. Lee S, Abecasis Gonçalo R, Boehnke M, et al. 2014b. Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet* 95: p. 5-23.

185. Lek M, Karczewski KJ, Minikel EV, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536: p. 285-291.

186. Leroy B, Fournier JL, Ishioka C, et al. 2013. The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis. *Nucl. Acids Res.* 41: p. D962-D969.

187. Lhota F, Zemankova P, Kleiblova P, et al. 2016. Hereditary truncating mutations of DNA repair and other genes in BRCA1/BRCA2/PALB2-negatively tested breast cancer patients. *Clin. Genet.* 90: p. 324-333.

188. Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: p. 1754-1760.

189. Li H, Handsaker B, Wysoker A, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: p. 2078-2079.

190. Li H, Homer N 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform* 11: p. 473-483.

191. Li M-X, Gui H-S, Kwan JSH, et al. 2012. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucl. Acids Res.* 7: p. e53.

192. Lim S, Kaldis P 2013. Cdks, cyclins and CKIs: roles beyond cell cycle regulation. *Development* 140: p. 3079-3093.

193. Lin DDM, Barker PB, Lederman HM, et al. 2014. Cerebral abnormalities in adults with ataxia-telangiectasia. *AJNR Am J Neuroradiol* 35: p. 119-123.

194. Lin H-Y, Sun M, Lin C, et al. 2009. Androgen-induced human breast cancer cell proliferation is mediated by discrete mechanisms in estrogen receptor +positive and -negative breast cancer cells. *J Steroid Biochem Mol Biol* 113: p. 182-188.

195. Liu C 2012. The CHEK2 I157T variant and colorectal cancer susceptibility: A systematic review and meta-analysis. *Asian Pacific J Cancer Prev* 13: p. 2051-2055.

196. Liu C, Cao K-AL, Chenevix-Trench G, et al. 2014. A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. *Nucl. Acids Res* 10: p. 6106-6127.

197. Liu X, Jian X, Boerwinkle E 2011. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32: p. 894-899.

198. Loke J, Pearlman A, Upadhyay K, et al. 2015. Functional variant analyses (FVAs) predict pathogenicity in the BRCA1 DNA double-strand break repair pathway. *Hum. Mol. Genet* 24: p. 3030-3037.

199. Long J, Zhang B, Signorello LB, et al. 2013. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS ONE* 8: p. e58350.

200. Loubser F, de Villiers JNP, van der Merwe NC 2012. Two double heterozygotes in a South African Afrikaner family: implications for BRCA1 and BRCA2 predictive testing. *Clin. Genet.* 82: p. 599-600.

201. Lu C, Xie M, Wendl MC, et al. 2015. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun* 6: p. 10086.

202. Luebben SW, Kawabata T, Akre MK, et al. 2013. Helq acts in parallel to Fancc to suppress replication-associated genome instability. *Nucl. Acids Res.* 41: p. 10283-10297.

203. Lynch H, Wen H, Kim YC, et al. 2013. Can Unknown Predisposition in Familial Breast Cancer be Family-Specific? *The Breast Journal* 19: p. 520-528.

204. Maby P, Tougeron D, Hamieh M, et al. 2015. Correlation between density of CD8Þ T-cell infiltrate in microsatellite unstable colorectal cancers and frameshift mutations: a rationale for personalized immunotherapy. *Cancer Res* 75: p. 3446.

205. MacArthur DG, Manolio TA, Dimmock DP, et al. 2014. Guidelines for investigating causality of sequence variants in human disease. *Nature* 504: p. 7.

206. Machackova E, Damborsky J, Valik D, et al. 2001. Novel germline BRCA1 and BRCA2 mutations in breast and breast/ovarian cancer families from the Czech Republic. *Hum. Mutat.* 18: p. 545-545.

207. Mai PL, Garceau AO, Graubard BI, et al. 2011. Confirmation of family cancer history reported in a population-based survey. *J. Natl. Cancer Inst.* 103: p. 788-797.

208. Marceau N, Loranger A, Gilbert S 2014. Intermediate Filaments. *Colloquium Series on Building Blocks of the Cell: Cell Structure and Function* 2: p. 1-112.

209. Marini F, Kim N, Schuffert A, et al. 2003. POLN, a nuclear pola family DNA polymerase homologous to the DNA cross-link sensitivity protein MUS308. *J. Biol. Chem.* 278: p. 32014-32019.

210. Marini F, Wood RD 2002. A Human DNA Helicase Homologous to the DNA Cross-link Sensitivity Protein Mus308. *J. Biol. Chem.* 277: p. 8716-8723.

211. Masciari S, Dillon D, Rath M, et al. 2012. Breast cancer phenotype in women with *TP53* germline mutations: a Li-Fraumeni syndrome consortium effort. *Breast Cancer Res Treat* 133: p. 1125-1130.

212. Mathe E, Olivier M, Kato S, et al. 2006. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucl. Acids Res.* 34: p. 1317-1325.

213. Matta J, Echenique M, Negron E, et al. 2012. The association of DNA Repair with breast cancer risk in women. A comparative observational study. *BMC Cancer* 12: p. 490.

214. Mavaddat N, Antoniou AC, Easton DF, et al. 2010a. Genetic susceptibility to breast cancer. *Molecular Oncology* 4: p. 174-191.

215. Mavaddat N, Pharoah PD, Blows F, et al. 2010b. Familial relative risks for breast cancer by pathological subtype: a population-based cohort study. *Breast Cancer Res* 12: p. 1–2.

216. Mealiffe ME, Stokowski RP, Rhees BK, et al. 2010. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natl. Cancer Inst.* 102: p. 1618-1627.

217. Mehta A, Haber JE 2014. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol* 6: p. a016428.

218. Meienberg J, Zerjavic K, Keller I, et al. 2015. New insights into the performance of human whole-exome capture platforms. *Nucl. Acids Res.* 43: p. e76-e76.

219. Mena S, Ortega A, Estrela JM 2009. Oxidative stress in environmental-induced carcinogenesis. *Mutat Res Genet Toxicol Environ Mutagen* 674: p. 36-44.

220. Merdad A, Gari M, Hussein S, et al. 2015. Characterization of familial breast cancer in Saudi Arabia. *BMC Genomics* 16: p. S3.

221. Metzker ML 2010. Sequencing technologies-the next generation. *Nat Rev Genet* 11: p. 31-46.

222. Miki Y, Swensen J, Shattuck-Eidens D, et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266: p. 66-71.

223. Milanowska K, Krwawicz J, Papaj G, et al. 2011. REPAIRtoire—a database of DNA repair pathways. *Nucl. Acids Res.* 39: p. D788-D792.

224. Minocherhomji S, Hickson ID 2014. Structure-specific endonucleases: guardians of fragile site stability. *Trends Cell Biol* 24: p. 321-327.

225. Mnif W, Hassine AIH, Bouaziz A, et al. 2011. Effect of endocrine disruptor pesticides: a review. *Int J Environ Res Public Health* 8: p. 2265-2303.

226. Moldovan G-L, D'Andrea AD 2009. How the fanconi anemia pathway guards the genome. *Annu Rev Genet* 43: p. 223-249.

227. Moldovan G-L, Dejsuphong D, Petalcorin Mark IR, et al. 2012. Inhibition of homologous recombination by the PCNA-interacting protein PARI. *Molecular Cell* 45: p. 75-86.

228. Moldovan G-L, Madhavan MV, Mirchandani KD, et al. 2010. DNA polymerase POLN participates in cross-link repair and homologous recombination. *Mol. Cell. Biochem* 30: p. 1088-1096.

229. Moncunill V, Gonzalez S, Bea S, et al. 2014. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat Biotech* 32: p. 1106-1112.

230. Montemurro F, Di Cosimo S, Arpino G 2013. Human epidermal growth factor receptor 2 (HER2)-positive and hormone receptor-positive breast cancer: new insights into molecular interactions and clinical implications. *Ann Oncol.* 24: p. 2715-2724.

231. Moorthie S, Mattocks C, Wright C 2011. Review of massively parallel DNA sequencing technologies. *The HUGO Journal* 5: p. 1-12.

232. Moravcikova E, Krepela E, Prochazka J, et al. 2012. Down-regulated expression of apoptosis-associated genes APIP and UACA in non-small cell lung carcinoma. *Int. J. Oncol.* 40: p. 2111-2121.

233. Moreau Y, Tranchevent L-C 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 13: p. 523-536.

234. Muranen TA, Greco D, Blomqvist C, et al. 2017. Genetic modifiers of CHEK2*1100delC associated breast cancer risk. *Genetics in medicine : official journal of the American College of Medical Genetics* 19: p. 599-603.

235. Nagel ZD, Chaim IA, Samson LD 2014. Inter-individual variation in DNA repair capacity: a need for multi-pathway functional assays to promote translational DNA repair research. *DNA Repair* 19: p. 199-213.

236. Need AC, Shashi V, Hitomi Y, et al. 2012a. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet.* 49: p. 353-361.

237. Need EF, Selth LA, Harris TJ, et al. 2012b. Research resource: interplay between the genomic and transcriptional networks of androgen receptor and estrogen receptor in luminal breast cancer cells. *J. Mol. Endocrinol* 26: p. 1941-1952.

238. Negrini S, Gorgoulis VG, Halazonetis TD 2010. Genomic instability-an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 11: p. 220-228.

239. Nehrt N, Peterson T, Park D, et al. 2012. Domain landscapes of somatic mutations in cancer. *BMC Genomics* 13: p. S9.

240. Neill T, Torres A, Buraschi S, et al. 2014. Decorin induces mitophagy in breast carcinoma cells via peroxisome proliferator-activated receptor γ coactivator-1α (PGC-1α) and MITOSTATIN. *J. Biol. Chem.* 289: p. 4952-4968.

241. Nelson AC, Holt JT 2010. Impact of RING and BRCT domain mutations on BRCA1 protein stability, localization, and recruitment to DNA damage. *Radiat. Res* 174: p. 1-13.

242. Nevanlinna H, Bartek J 2006. The *CHEK2* gene and inherited breast cancer susceptibility. *Oncogene* 25: p. 5912-5919.

243. Neyen C, Plüddemann A, Mukhopadhyay S, et al. 2013. Macrophage Scavenger Receptor A Promotes Tumor Progression in Murine Models of Ovarian and Pancreatic Cancer. *J. Immunol* 190: p. 3798-3805.

244. Ng PC, Henikoff S 2003. SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* 31: p. 3812-3814.

245. Nhung D. 2014. Helq works in parallel to Brca2 to suppress chromosome instability. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/163254.

246. Nicolay NH, Helleday T, Sharma AR 2012. Biological relevance of DNA polymerase beta and translesion synthesis polymerases to cancer and its treatment. *Curr Mol Pharmacol* 5: p. 54-67.

247. Nishizawa M, Izawa I, Inoko A, et al. 2005. Identification of trichoplein, a novel keratin filament-binding protein. *J. Cell Sci* 118: p. 1081-1090.

248. Nitsch D, Goncalves J, Ojeda F, et al. 2010. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 11: p. 460.

249. Njiaju UO, Olopade OI 2012. Genetic determinants of breast cancer risk: a review of current literature and issues pertaining to clinical application. *Breast J* 18: p. 436-442.

250. Noh JM, Kim J, Cho DY, et al. 2015. Exome sequencing in a breast cancer family without BRCA mutation. *Radiat Oncol J* 33: p. 149-154.

251. O'Hayre M, Degese MS, Gutkind JS 2014. Novel insights into G protein and G protein-coupled receptor signaling in cancer. *Curr Opin Cell Biol* 27: p. 126-135.

252. Olivier M, Hollstein M, Hainaut P 2010. TP53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* 2.

253. Olsson N, Carlsson P, James P, et al. 2013. Grading breast cancer tissues using molecular portraits. *Mol Cell Proteomics* 12: p. 3612-3623.

254. Osborne C, Wilson P, Tripathy D 2004. Oncogenes and tumor suppressor genes in breast cancer: potential diagnostic and therapeutic applications. *The Oncologist* 9: p. 361-377.

255. Osher DJ, De Leeneer K, Michils G, et al. 2012. Mutation analysis of *RAD51D* in non-*BRCA1/2* ovarian and breast cancer families. *Br J Cancer* 106: p. 1460-1463.

256. Pabinger S, Dander A, Fischer M, et al. 2013. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform* 2: p. 256-278.

257. Panoutsopoulou K, Tachmazidou I, Zeggini E 2013. In search of low-frequency and rare variants affecting complex traits. *Hum. Mol. Genet* 22: p. R16-R21.

258. Park D, Odefrey F, Hammet F, et al. 2011. FAN1 variants identified in multiple-case early-onset breast cancer families via exome sequencing: no evidence for association with risk for breast cancer. *Breast Cancer Research and Treatment* 130: p. 1043-1049.

259. Park DJ, Lesueur F, Nguyen-Dumont T, et al. 2012. Rare mutations in *XRCC2* increase the risk of breast cancer. *Am J Hum Genet* 90: p. 734-739.

260. Park DJ, Tao K, Le Calvez-Kelm F, et al. 2014. Rare mutations in RINT1 predispose carriers to breast and Lynch Syndrome-spectrum cancers. *Cancer Discov* 4: p. 804-815.

261. Paszkiewicz K, Studholme D. 2012. High-throughput sequencing data analysis software: current state and future developments. In: Rodríguez-Ezpeleta N, et al., editors. Bioinformatics for High Throughput Sequencing: Springer New York: p. 231-248.

262. Patel RK, Jain M 2012. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* 7: p. e30619.

263. Patrono C, Sterpone S, Testa A, et al. 2014. Polymorphisms in base excision repair genes: Breast cancer risk and individual radiosensitivity. *World J Clin Oncol* 5: p. 874-882.

264. Patterson RE, Cadmus LA, Emond JA, et al. 2010. Physical activity, diet, adiposity and female breast cancer prognosis: A review of the epidemiologic literature. *Maturitas* 66: p. 5-15.

265. Pelttari L, Kinnunen L, Kiiski J, et al. 2015. Screening of *HELQ* in breast and ovarian cancer families. *Fam. Cancer* 1: p. 19-23.

266. Pepe A, West SC 2013. Substrate specificity of the MUS81-EME2 structure selective endonuclease. *Nucl. Acids Res.* 42: p. 3833-3845.

267. Pepe A, West Stephen C 2014. MUS81-EME2 promotes replication fork restart. *Cell Rep* 7: p. 1048-1055.

268. Perou CM, Sorlie T, Eisen MB, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406: p. 747-752.

269. Petitjean A, Mathe E, Kato S, et al. 2007. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* 28: p. 622-629.

270. Petrucelli N, Daly M, Feldman G. 2013. *BRCA1* and *BRCA2* Hereditary Breast and Ovarian Cancer. In: Pagon R, et al., editors. Gene Reviews. Seattle University of Washington, Seattle: p. 1993-2014.

271. Plati J, Bucur O, Khosravi-Far R 2011. Apoptotic cell signaling in cancer progression and therapy. *Integr. Biol* 3: p. 279-296.

272. Polo SE, Jackson SP 2011. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev* 25: p. 409-433.

273. Pop M, Salzberg SL 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet* 24: p. 142-149.

274. Potter JD 2011. Development and the environment: clues to carcinogenesis. *Cancer Epidemiol Biomarkers Prev* 20: p. 574-577.

275. Poulogiannis G, Frayling IM, Arends MJ 2010. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. *Histopathology* 56: p. 167-179.

276. Prat A, Adamo B, Fan C, et al. 2013. Genomic analyses across six cancer types identify basal-like breast cancer as a unique molecular entity. *Sci. Rep.* 3: p. 1-12.

277. Prat A, Parker J, Karginova O, et al. 2010. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research* 12: p. R68.

278. Prat A, Perou CM 2011. Deconstructing the molecular portraits of breast cancer. *Molecular Oncology* 5: p. 5-23.

279. Quintáns B, Ordóñez-Ugalde A, Cacheiro P, et al. 2014. Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Appl Transl Genom.* 3: p. 60-67.

280. Rahman N 2014. Realizing the promise of cancer predisposition genes. *Nature* 505: p. 302-308.

281. Rahman N, Stratton MR 2008. The emerging landscape of breast cancer susceptibility. *Nat Genet* 40: p. 17-22.
282. Reese MG, Eeckman FH, Kulp D, et al. 1997. Improved splice site detection in genie. *J. Comp. Biol.* 4: p. 311-323.
283. Reeves GK, Travis RC, Green J, et al. 2010. Incidence of breast cancer and its subtypes in relation to individual and multiple low-penetrance genetic susceptibility loci. *JAMA* 304: p. 426-434.
284. Reeves MD, Yawitch TM, van der Merwe NC, et al. 2004. BRCA1 mutations in South African breast and/or ovarian cancer families: Evidence of a novel founder mutation in Afrikaner families. *Int. J. Cancer* 110: p. 677-682.
285. Resta N, Pierannunzio D, Lenato GM, et al. 2013. Cancer risk associated with *STK11/LKB1* germline mutations in Peutz-Jeghers syndrome patients: Results of an Italian multicenter study. *Dig Liver Dis* 45: p. 606-611.
286. Reva B, Antipin Y, Sander C 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucl. Acids Res.* 17: p. e118.
287. Robinson PN, Köhler S, Oellrich A, et al. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 24: p. 340-348.
288. Robinson PN, Krawitz P, Mundlos S 2011. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clin. Genet.* 80: p. 127-132.
289. Rotunno M, Sun X, Figueroa J, et al. 2014. Parity-related molecular signatures and breast cancer subtypes by estrogen receptor status. *Breast Cancer Research* 16: p. R74.
290. Roy R, Chun J, Powell SN 2012. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat Rev Cancer* 12: p. 68-78.
291. Ruijs MWG, Verhoef S, Rookus MA, et al. 2010. TP53 germline mutation testing in 180 families suspected of Li-Fraumeni syndrome: mutation detection rate and relative frequency of cancers in different familial phenotypes. *J Med Genet.* 47: p. 421-428.
292. Sanborn JZ, Benz SC, Craft B, et al. 2011. The UCSC cancer genomics browser: update 2011. *Nucl. Acids Res.* 39: p. D951-D959.
293. Sandhu R, Parker JS, Jones WD, et al. 2010. Microarray-based gene expression profiling for molecular classification of breast cancer and identification of new targets for therapy. *Lab Medicine* 41: p. 364-372.
294. Santarpia L, Iwamoto T, Di Leo A, et al. 2013. DNA repair gene patterns as prognostic and predictive factors in molecular breast cancer subtypes. *Oncologist* 18: p. 1063-1073.
295. Sapkota Y, Yasui Y, Lai R, et al. 2013. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. *PLoS ONE* 8: p. e62550.
296. Schlebusch CM, Dreyer G, Sluiter MD, et al. 2010. Cancer prevalence in 129 breast-ovarian cancer families tested for BRCA1 and BRCA2 mutations. *S Afr Med J.* 100: p. 113-117.

297. Schweingruber C, Rufener SC, Zünd D, et al. 2013. Nonsense-mediated mRNA decay — Mechanisms of substrate mRNA recognition and degradation in mammalian cells. *BBA – Gene Regulatory Mechanisms* 1829: p. 612-623.
298. Sehgal M, Singh TR 2014. DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems. *DNA Repair* 16: p. 97-103.
299. Shakeel MK, George PS, Jose J, et al. 2010. Pesticides and breast cancer risk: a comparison between developed and developing countries. *Asian Pacific J Cancer Prev* 11: p. 173-180.
300. Shao X, Grishin NV 2000. Common fold in helix–hairpin–helix proteins. *Nucl. Acids Res.* 28: p. 2643-2650.
301. Shen C, Sun H, Sun D, et al. 2011. Polymorphisms of tumor necrosis factor-alpha and breast cancer risk: a meta-analysis. *Breast Cancer Res Treat* 126: p. 763-770.
302. Shin Y-K, Amangyeld T, Nguyen TA, et al. 2012. Human MUS81 complexes stimulate flap endonuclease 1. *FEBS Journal* 279: p. 2412-2430.
303. Shrivastav M, De Haro LP, Nickoloff JA 2008. Regulation of DNA double-strand break repair pathway choice. *Cell Res* 18: p. 134-147.
304. Shuen A, Foulkes W 2011. Inherited mutations in breast cancer genes risk and response. *J Mammary Gland Biol Neoplasia.* 16: p. 3-15.
305. Signore M, Ricci-Vitiani L, De Maria R 2013. Targeting apoptosis pathways in cancer stem cells. *Cancer Lett* 332: p. 374-382.
306. Sluiter M, Mew S, van Rensburg E 2009. PALB2 sequence variants in young South African breast cancer patients. *Fam. Cancer* 8: p. 347-353.
307. Sluiter M, van Rensburg E 2011. Large genomic rearrangements of the *BRCA1* & *BRCA2* genes: review of the literature and report of a novel *BRCA1* mutation. *Breast Cancer Res Treat* 125: p. 325-349.
308. Smith AL, Alirezaie N, Connor A, et al. 2016. Candidate DNA repair susceptibility genes identified by exome sequencing in high-risk pancreatic cancer. *Cancer Lett* 370: p. 302-312.
309. Smith J, Tho L, Xu N, et al. 2010. The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv Cancer Res.* 108: p. 73-112.
310. Smith TR, Levine EA, Freimanis RI, et al. 2008. Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis* 29: p. 2132-2138.
311. Snape K, Ruark E, Tarpey P, et al. 2012. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res Treat* 134: p. 429-433.
312. Sonnenschein C, Soto AM 2010. Environmental causes of cancer: endocrine disruptors as carcinogens. *Nat Rev Endocrinol* 6: p. 363-370.
313. Sotiriou C, Pusztai L 2009. Gene-expression signatures in breast cancer. *N Engl J Med* 360: p. 790-800.
314. Spencer SL, Berryman MJ, García JA, et al. 2004. An ordinary differential equation model for the multistep transformation to cancer. *J. Theor. Biol.* 231: p. 515-524.
315. Spurrell C 2013. Identifying New Genes for Inherited Breast Cancer by Exome Sequencing. University of Washington.

316. Stephens PJ, McBride DJ, Meng-Lay L, et al. 2009. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 462: p. 1005-1010.
317. Stephens PJ, Tarpey PS, Davies H, et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486: p. 400-404.
318. Stitziel N, Kiezun A, Sunyaev S 2011. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol* 12: p. 227.
319. Stratton MR, Rahman N 2008. The emerging landscape of breast cancer susceptibility. *Nat Genet* 40: p. 17-22.
320. Subauste MC, Sansom OJ, Porecha N, et al. 2010. Fem1b, a proapoptotic protein, mediates proteasome inhibitor-induced apoptosis of human colon cancer cells. *Mol Carcinog* 49: p. 105-113.
321. Suhasini AN, Brosh Jr RM 2013. Disease-causing missense mutations in human DNA helicase disorders. *Mutat Res* 752: p. 138-152.
322. Sun J, Hsu F-C, Turner AR, et al. 2006. Meta-analysis of association of rare mutations and common sequence variants in the MSR1 gene and prostate cancer risk. *The Prostate* 66: p. 728-737.
323. Szabo C, Masiello A, Ryan JF, et al. 2000. The breast cancer information core: database design, structure, and scope. *Hum. Mutat.* 16: p. 123-131.
324. Tafel AA, Wu L, McHugh PJ 2011. Human HEL308 localizes to damaged replication forks and unwinds lagging strand structures. *J. Biol. Chem.* 286: p. 15832-15840.
325. Takata K-i, Reh S, Tomida J, et al. 2013. Human DNA helicase HELQ participates in DNA interstrand crosslink tolerance with ATR and RAD51 paralogs. *Nat Commun* 4: p. 1-11.
326. Takata K-i, Tomida J, Reh S, et al. 2015. Conserved overlapping gene arrangement, restricted expression, and biochemical activities of DNA polymerase ν (POLN). *J. Biol. Chem.* 290: p. 24278-24293.
327. Tan R, Wang Y, Kleinstein SE, et al. 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum. Mutat.* 35: p. 899-907.
328. Taube JH, Herschkowitz JI, Komurov K, et al. 2010. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences* 107: p. 15449-15454.
329. The Cancer Genome Atlas Network T 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490: p. 61-70.
330. Thiagalingam S. 2015. Systems biology of cancer progression. In: Thiagalingam S, editor. Systems Biology of Cancer. United Kingdom: Cambridge University Press: p. 1-531.
331. Thomas D 2010. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet* 11: p. 259-272.
332. Thomas S, Bonchev D 2010. A survey of current software for network analysis in molecular biology. *Hum Genomics* 4: p. 353 - 360.

333. Thompson ER, Doyle MA, Ryland GL, et al. 2012. Exome sequencing identifies rare deleterious mutations in DNA repair genes *FANCC* and *BLM* as potential breast cancer susceptibility alleles. *PLoS Genet* 8: p. e1002894.

334. Tough IC, Carter DC, Fraser J, et al. 1969. Histological grading in breast cancer. *Br. J. Cancer, BJC* 23: p. 294-301.

335. Tranchevent L-C, Capdevila FB, Nitsch D, et al. 2011. A guide to web tools to prioritize candidate genes. *Brief. Bioinform* 12: p. 22-32.

336. The Human Protein Atlas. 2015. Available from: http://www.proteinatlas.org/ENSG00000163312-HELQ. Web. 30 Nov. 2016.

337. Uhlén M, Fagerberg L, Hallström BM, et al. 2015. Tissue-based map of the human proteome. *Science* 347.

338. Usary J, Zhao W, Darr D, et al. 2013. Predicting drug responsiveness in human cancers using genetically engineered mice. *Clin. Cancer Res* 19: p. 4889-4899.

339. Van der Auwera GA, Carneiro MO, Hartl C, et al. 2013. From FASTQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics.* 43: p. 1-33.

340. van der Merwe NC, Hamel N, Schneider SR, et al. 2012. A founder BRCA2 mutation in non-Afrikaner breast cancer patients of the Western Cape of South Africa. *Clin. Genet.* 81: p. 179-184.

341. Vandeweyer G, Van Laer L, Loeys B, et al. 2014. VariantDB: a flexible annotation and filtering portal for next generation sequencing data. *Genome Med* 6: p. 74.

342. Vazquez M, de la Torre V, Valencia A 2012. Chapter 14: Cancer Genome Analysis. *PLoS Comput Biol* 8: p. e1002824.

343. Vecchione A, Fassan M, Anesti V, et al. 2008. MITOSTATIN, a putative tumor suppressor on chromosome 12q24.1, is downregulated in human bladder and breast cancer. *Oncogene* 28: p. 257 - 269.

344. Vogelstein B, Papadopoulos N, Velculescu VE, et al. 2013. Cancer Genome Landscapes. *Science* 339: p. 1546-1558.

345. Vorobiof DA, Sitas F, Vorobiof G 2001. Breast cancer incidence in South Africa. *J Clin Oncol* 19: p. 125-127.

346. Vuorela M, Pylkäs K, Hartikainen J, et al. 2011. Further evidence for the contribution of the RAD51C gene in hereditary breast and ovarian cancer susceptibility. *Breast Cancer Res Treat* 130: p. 1003-1010.

347. Walsh T, Casadei S, Coats KH, et al. 2006. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* 295: p. 1379-1388.

348. Walsh T, Casadei S, Lee MK, et al. 2011. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proceedings of the National Academy of Sciences* 108: p. 18032-18037.

349. Wang Anderson T, Kim T, Wagner John E, et al. 2015a. A Dominant Mutation in Human RAD51 Reveals Its Function in DNA Interstrand Crosslink Repair Independent of Homologous Recombination. *Molecular Cell* 59: p. 478-490.

350. Wang K, Li M, Hakonarson H 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.* 38: p. e164-e164.

351. Wang R-A, Li Q-L, Li Z-S, et al. 2013. Apoptosis drives cancer cells proliferate and metastasize. *J Cell Mol Med.* 17: p. 205-211.

352. Wang W, Zhao S, Zhuang L, et al. 2015b. The screening of *HELQ* gene in Chinese patients with premature ovarian failure. *Reprod Biomed Online* 31: p. 573-576.

353. Weigelt B, Geyer FC, Reis-Filho JS 2010. Histological types of breast cancer: How special are they? *Molecular Oncology* 4: p. 192-208.

354. Weitzel JN, Blazer KR, MacDonald DJ, et al. 2011. Genetics, genomics, and cancer risk assessment. *CA Cancer J Clin.* 61: p. 327-359.

355. Wen H, Kim Y, Snyder C, et al. 2014. Family-specific, novel, deleterious germline variants provide a rich resource to identify genetic predispositions for *BRCAx* familial breast cancer. *BMC Cancer* 14: p. 470.

356. Whiley PJ, Guidugli L, Walker LC, et al. 2011. Splicing and multifactorial analysis of intronic *BRCA1* and *BRCA2* sequence variants identifies clinically significant splicing aberrations up to 12 nucleotides from the intron/exon boundary. *Hum. Mutat.* 32: p. 678-687.

357. Williams AB, Michael WM 2010. Eviction notice: new insights into RAD51 removal from DNA during homologous recombination. *Molecular Cell* 37: p. 157-158.

358. Witsch E, Sela M, Yarden Y 2010. Roles for growth factors in cancer progression. *Physiology* 25: p. 85-101.

359. Wolters S, Schumacher B 2013. Genome maintenance and transcription integrity in ageing and disease. *Front Genet.* 4: p. 1-10.

360. Woodman IL, Brammer K, Bolt EL 2011. Physical interaction between archaeal DNA repair helicase Hel308 and Replication Protein A (RPA). *DNA Repair* 10: p. 306-313.

361. Wooster R, Bignell G, Lancaster J, et al. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378: p. 789-792.

362. Xiang H-p, Geng X-p, Ge W-w, et al. 2011. Meta-analysis of CHEK2 1100delC variant and colorectal cancer susceptibility. *Eur. J. Cancer* 47: p. 2546-2551.

363. Yajima H, Isomoto H, Nishioka H, et al. 2013. Novel serine/threonine kinase 11 gene mutations in Peutz-Jeghers syndrome patients and endoscopic management. *World J Gastrointest Endosc* 16: p. 102-110.

364. Yang P, Du CW, Kwan M, et al. 2013. The impact of p53 in predicting clinical outcome of breast cancer patients with visceral metastasis. *Sci. Rep.* 3: p. 1-6.

365. Yang XR, Chang-Claude J, Goode EL, et al. 2011. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the breast cancer association consortium studies. *J. Natl. Cancer Inst.* 103: p. 250-263.

366. Yao Y, Dai W 2014. Genomic Instability and Cancer. *J Carcinog Mutagen.* 5: p. 1000165.

367. Yeo G, Burge CB 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comp. Biol.* 11: p. 377-394.

368. Yoshida K, Miki Y 2004. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci.* 95: p. 866-871.

369. Zhan H, Suzuki T, Aizawa K, et al. 2010. Ataxia telangiectasia mutated (ATM)-mediated DNA damage response in oxidative stress-induced vascular endothelial cell senescence. *J. Biol. Chem.* 285: p. 29662-29670.

370. Zhang J, Fackenthal JD, Zheng Y, et al. 2012. Recurrent BRCA1 and BRCA2 mutations in breast cancer patients of African ancestry. *Breast Cancer Res Treat* 134: p. 889+.

371. Ziebarth JD, Bhattacharya A, Cui Y 2012. Integrative analysis of somatic mutations altering microRNA targeting in cancer genomes. *PLoS ONE* 7: p. e47137.