# Germline sequence variants contributing to cancer susceptibility in South African breast cancer patients

**by**

**Dewald Eygelaar**

**Submitted in partial fulfilment of the requirements for the degree**

**PhD Bioinformatics**

**In the Faculty of Natural & Agricultural Sciences**

**University of Pretoria**

**Pretoria**

November 2022

# DEDICATION

I dedicate this thesis to my loving wife, who literally walked this journey with me from start to finish. Without you I would have never been able to climb this mountain, you were always there with words of encouragement and a hot cup of coffee.

Thank you, Monique

# DECLARATION

---

**I, Dewald Eygelaar declare that the thesis/dissertation, which I hereby submit for the degree PhD in Bioinformatics 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.**

**SIGNATURE: Dewald Eygelaar**

**DATE: November 2022**

# ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the following people, organizations, and institutions for supporting me throughout the course of my Doctoral studies. Their valuable inputs have contributed to the completion of this dissertation.

First and foremost, a great shout out to Prof. Fourie Joubert for all his help, mentorship, patience throughout this project. He played a fundamental role in guiding me into and through the field of Bioinformatics, which I grow to love and now work in. Each day was a blessing to work with you. I would also like to extend my appreciation to my co-supervisor, Prof. Lizette Jansen van Rensburg for sharing their scientific knowledge, countless corrections, and help. I would also like to thank her for the acquisition and sharing of samples that was used for this project.

This work was funded by the South African National Research Foundation and University of Pretoria Research Development Programme.

# RESEARCH OUTPUTS

**Published papers:**

1. Eygelaar D, Jansen van Rensburg E, Joubert F. Germline sequence variants contributing to cancer susceptibility in South African breast cancer patients of African ancestry. Nature, Scientific Reports, 2022

**International conferences:**

1. Eygelaar D, Jansen van Rensburg E, Joubert F. Understanding breast cancer in a subset of South African patients by identifying single nucleotide variants, insertions and deletions, copy number variations and gene expression profiles. ISMB, Orlando, Florida, 8-12 July 2016. (Poster presentation)
2. Eygelaar D, Jansen van Rensburg E, Joubert F. A Comparative study to identify single nucleotide variants, insertions and deletions, copy number variations and gene expression profiles in a subset of South African patients. H3Africa Consortium Meeting, Dakar, Senegal ,12-17 May 2016. (Oral presentation)
3. Eygelaar D, Jansen van Rensburg E, Joubert F. Germline variants in cancer-related genes of South African breast cancer patients. AACR-KCA, Seoul, South Korea, 2018. (Poster presentation)

**National conferences:**

1. Eygelaar D, Jansen van Rensburg E, Joubert F. The analysis of breast cancer germline variants in a South African black population using next generation sequencing. South African Genetics Society and SA Society for Bioinformatics, Durban, South Africa, 2016. (Oral presentation)
2. WITS Cancer Research symposium. Johannesburg, South Africa , 2017. (Attended)
3. Eygelaar D, Jansen van Rensburg E, Joubert F. Germline sequence variants in cancer susceptibility genes of South African women with breast cancer. South African Genetics Society and SA Society for Bioinformatics, Free State, South Africa, 2018. (Oral presentation)

# TABLE OF CONTENTS

DEDICATION....2
DECLARATION....3
ACKNOWLEDGEMENTS....4
RESEARCH OUTPUTS....5
TABLE OF CONTENTS....6
ABBREVIATIONS....9
PROJECT SUMMARY....14

**1. CHAPTER 1 Literature Review....16**
1.1 What is cancer?....16
1.2 History of cancer....16
1.3 Cancer genomics....17
1.4 Types of cancer....19
1.5 Gender-based studies....19
1.6 Ethnical importance of different cancers....20
1.7 What is breast cancer....20
1.8 Different types of breast cancer....21
1.8.1 Histological breast cancer types....22
1.8.2 Histological tumour grades....24
1.8.3 Intrinsic breast cancer tumours....24
1.8.4 Known markers used in the identification of breast cancer....26
1.9 Specialized invasive carcinomas....28
1.10 Gender based studies for breast cancer....28
1.11 Risk factors for breast cancers....29
1.12 Enriched pathways....32
1.13 Breast cancer in Africa....33
1.14 Breast cancer genomics....35
1.15 Researching cancer....36
1.15.1 Variant analysis and identification....37
1.15.2 Gene expression profiling....38
1.15.3 Cancer genome sequencing....42

1.15.4 Copy number variance.....45

1.16 Aims and objectives.....47

1.17 References.....47

**2. CHAPTER 2 Materials and methods.....60**

2.1 Sampling, patient information and consent.....60

2.2 DNA extractions.....65

2.3 Ethics approval.....65

2.4 Analysed cancer genes.....65

2.5 Library preparation and sequencing.....66

2.6 Sequencing data analysis.....66

2.6.1 Execution of analysis.....66

2.6.2 Variant calling.....67

2.6.3 Variant annotation.....67

2.6.4 Variant filtration and in silico evaluation of variants.....67

2.6.5 Enriched pathways.....68

2.7 Nature and form of results.....69

2.8 Data management.....69

2.9 Authors' list of agreement for publications and presentations.....69

2.10 References.....70

**3. CHAPTER 3 General results and pathogenic/likely pathogenic variants in known breast cancer susceptibility genes.....72**

3.1 Introduction.....72

3.2 Aim.....84

3.3 Results.....84

3.3.1 General results.....86

3.3.2 Variants of interest.....86

3.3.3 Pathogenic/likely pathogenic variants in known breast cancer susceptibility genes.....90

3.3.4 Pathway enrichement.....95

3.3.5 Comparison between GATK 3.8 and GATK 4.0.....96

3.4 Discussion.....97

3.5 Supplementary tables.....102
3.6 References.....103
**4. CHAPTER 4 Pathogenic variants in hereditary cancer predisposition genes exclusively investigated for truncating variants, and variants of unknown significance.....108**
4.1 Introduction.....108
4.2 Other cancer susceptibility variants .....109
4.2.1 Pathogenic variants in hereditary cancer predisposition genes exclusively investigated for truncating variants.....109
4.2.2 Lesser-known hereditary variants identified: intronic variants .....116
4.3 Variants of unknown significance.....119
4.4 Discussion.....122
4.5 References.....123

**5. CHAPTER 5 Analysis of non-coding variants.....130**
5.1 Introduction.....130
5.2 Materials and methods.....134
5.3 Aim.....134
5.4 Results.....134
5.5 Discussion.....138
5.6 References.....143

**6. CHAPTER 6 Concluding discussion.....150**

**APPENDIX I: Copy of published paper.....155**

# ABBREVIATIONS

A.D.: anno domini

A: adenine

Ala: alanine

ALCL: Anaplastic Large Cell Lymphoma

AML: Acute Myeloid Leukaemia

APC/P: anaphase-promoting complex/cyclosome

Arg: arginine

Asn: asparagine

ASCO: American society of clinical oncology

B.C.: before Christ

BRCT: BRCA1 C-terminal

BQSR: base quality score recalibration

BWA: Burrows–Wheeler Aligner

C: cytosine

CA: cancer antigen

CADD: Combined Annotation Dependent Depletion

CAGE: cap analysis gene expression

CCV: credible causal variants

cDNA: complementary  deoxyribonucleic acid

CEA: carcinoembryonic antigen

CD: Cathepsin D

CDS: coding sequences

ChAM: chromatin association motif

CIN: chromosome instability

CM: cutaneous melanoma

CNV: copy number variant

DAC: data access committee

dbNSFP: database for nonsynonymous SNPs' functional predictions

dbSNP: Single Nucleotide Polymorphism Database

del: deletion

DES: diethylstilbestrol

DNA: deoxyribonucleic acid

dup: duplication

Dx: diagnosis

EDTA: Ethylenediaminetetraacetic acid

ELISA: Enzyme-linked immunosorbent assay

EMBL-EBI: European Bioinformatics Institute

ENCODE: Encyclopaedia of DNA Elements

ER: oestrogen receptor

EST: expressed sequence tag

ExAC: Exome Aggregation Consortium

FA: Fanconi anaemia

FATHMM-MKL: Functional Analysis through Hidden Markov Model

FFPE: formalin-fixed paraffin-embedded

FISH: fluorescence *in situ* hybridization

G: guanine

GATK: Genome Analysis Toolkit

GB: gigabyte

GBM: glioblastoma multiforme

GLEBS: Gle2-binding-sequence

Gln: glutamate

Glu: glutamic acid

gnomAD: Genome Aggregation Database

GRCh37: Genome Reference Consortium Human genome build 37

gVCF: genomic variant call format

GWAS: genome-wide association studies

HER2: epidermal growth factor receptor 2

hg19: human genome 19

HGVS: Human Genome Variation Society

His: histidine

IARC: International Agency for Research on Cancer

ICGC: International Cancer Genome Consortium

Ile: Isoleucine

InDel: insertion/deletion

Inf duct: infiltrating ductal carcinoma

inf lobular: infiltrating lobular carcinoma

kDa: kilodalton

Leu: leucine

lncRNA: long non-coding RNA

LOH: loss of heterozygosity

Lys: lysine

Met: methionine

miRNA: micro ribonucleic acid

MLPA: multiplex ligation-dependent probe amplification

MM: mesothelioma

MsigDB: molecular signatures database

NES: nuclear export signal

NGS: next generation sequencing

NHGRI: National Human Genome Research Institute

PAI-1: plasminogen activator inhibitor

PCAWG: Pan-Cancer Analysis of Whole Genomes

PCR: polymerase chain reaction

Pfam: protein families

PKC: protein kinase

PR: progesterone receptor

Pro: proline

qPCR: quantitative real time PCR

RAM: random-access memory

RBP: RNA binding protein

RCC: renal cell carcinoma

RefSNP: reference SNP

RNA: ribonucleic acid

ROS: reactive oxygen species

RTK: receptor tyrosine kinase

SAGE: serial analysis of gene expression

Ser: Serine

SNP: single nucleotide polymorphisms

T: thymine

TBNC: triple negative breast cancer

TCGA: The Cancer Genome Atlas

Ter: termination

TF: transcription factor

Thr: Threonine

TPR: tetratrico-peptide repeat

Trp: tryptophan

Tyr: tyrosine

UCSC: University of California, Santa Cruz

UM: uveal melanoma

uPA: urokinase plasminogen activator

USA: United States of America

UTR: untranslated region

Val: valine

VEP: Variant Effect Predictor

VUS: variant of uncertain significance

WES: whole exome sequencing

WGS: whole genome sequencing

# PROJECT SUMMARY

Breast cancer is increasingly a public health problem worldwide. It is the most commonly diagnosed cancer and the leading cause of cancer deaths in women. Breast cancer incidence and mortality rates are rising in transitioning countries in Africa, with some of the most rapid increases occurring in sub-Saharan Africa. Newly diagnosed breast cancer cases in South Africa accounts for 27.1% of female cancers in 2020, with age-standardized (World) incidence and mortality rates of 52.6 and 16 (per 100,000 women) respectively. Cancer results from a process of genetic changes, some inherited, some induced by environmental exposures and some occurring by chance. Early age of onset and a family history is a hallmark of hereditary breast cancer that is associated with germline variants in the high-penetrance genes, BRCA1 and BRCA2. An association with breast cancer susceptibility has also been reported for a further eleven high- to moderate-penetrance genes (TP53, PALB2, PTEN, STK11, CDH1, ATM, BRIP1, CHEK2, RAD51B, RAD51C, and RAD51D). In addition, pathogenic variants in genes from the mismatch repair pathway (MLH1, MSH2, MSH6 and PMS2) have been identified in breast cancer and ovarian cancer patients.

This study screened 165 South African breast cancer patients of African ancestry (self-identified) for the presence of deleterious germline sequence variants in 94 genes associated with hereditary cancer. The patients were unselected for age at diagnosis or family history of cancer. We identified pathogenic/likely pathogenic variants in thirteen patients from genes (ALK, ATM, BRCA1, BRCA2, BUB1B, CHEK2, FANCG,PALB2, RB1 and XPC ). Furthermore, a set of 27 variants of unknown significance was identified and reported that may play an important role in the future of pathogenic variants in the African population. Lastly, fourteen significant non-coding pathogenic variants from upstream, downstream and intergenic introns around the exons were identified using a combination of variant effect, CADD-PHRED and FATHMM-MKL predictions.

To our knowledge, only two studies in Africa, one on Nigerian women, and one on women from Uganda and Cameroon, have used multigene panel sequencing to test for germline variants in patients, unselected for family history or age at diagnosis. Although we investigated a relatively

small cohort of patients, our study provides some insights towards the genetic breast cancer risk factors in South African women of African ancestry. Further studies of a larger patient cohort is warranted to assess the distribution of variants in clinically relevant cancer susceptibility genes.

# Chapter 1: Literature review

## 1.1 WHAT IS CANCER?

Cancer is the occurrence of a number of different diseases in the same system that leads to the unmanaged/uncontrolled proliferation of cells, which have the ability to spread and affect different tissues/organs in that system (Rebbeck 2020, NCI 2021). Different names that have been attributed to this disease include: malignant tumours and malignant neoplasms, where malignant implies that the tumours are harmful, because benign tumours could also be present which are not detrimental to the system.

## 1.2 HISTORY OF CANCER

Cancer has been around longer than we think and medical practitioners, with no aid of the modern techniques available today, could describe cancer as early as Egyptian times. The first description of cancer has been noted in the old Edwin Smith Papyrus which forms part of the ancient Egyptian trauma surgery textbook (3000 B.C.) (ACS 2009). In that time, there was no treatment for it.

Only thousands of years later did the father of medicine, Hippocrates (460 B.C. – 370 B.C.) coin a term for this disease. Because dissection was not allowed in that time, he derived the name, *carcinos*, from how the tumours with veins seemingly spreading out of it looked like a crab (Greek: carcinos) (ACS 2009). The roman encyclopaedist, Aulus Cornelius Celsus (c. 25 B.C. – c. 50 A.D.) translated carcinos into the Latin term for crab, cancer. This is the term still being used to describe the disease today.

During the Renaissance (1500-1800), science was introduced into medicine which led to a greater understanding of circulation throughout the body through autopsies. In the late 1700's, John Hunter, a Scottish surgeon, speculated that some cancers may be removed surgically if it was "moveable" or has not invaded other tissues (ACS 2009).

In the 1800's, advances in research led to the birth of scientific oncology. Here, damage caused by cancer could be better understood. Pathologists could now remove body tissues and make a clear diagnosis of the disease as well as aid surgeons by informing them if the cancer was completely removed during a procedure (ACS 2009).

## 1.3 CANCER GENOMICS

Historically, the initiation of genomic research was in the 1900s when DNA sequences were generated for a variety of organisms. The first sequence to be determined by research was that of alanine transfer RNA in 1964 (Holley, Everett et al. 1965). The first cancer genome was sequenced in 2008, that of a typical Acute Myeloid Leukaemia (AML) genome and its counterpart normal genome (Strausberg and Simpson 2010). These researchers identified ten genes that contained mutations which brought on the cancer.

Research into cancer has led to the identification of different levels at which it could be studied:

- Protein level, altered levels of proteins that are translated or their abundance.
- Transcription level, altered gene expression.
- Epigenetic level, the physical alterations to DNA to control the transcription of specific genes, histone modifications and methylation.
- Genomic level, DNA mutations.

Oncogenomics, also known as cancer genomics, is the study of cancer using high throughput sequencing technologies. This field of study incorporates all the levels referred to above except the protein level in analysing cancer.

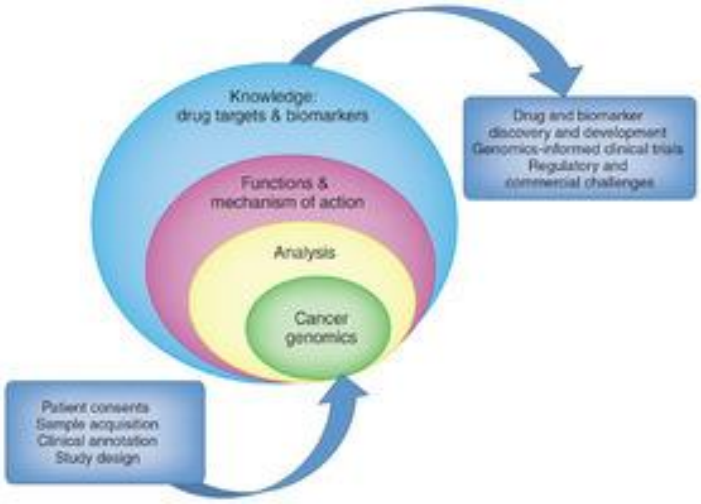


Figure 1.1: A flow diagram depicting the different levels of cancer genomic research and its outputs (Chin, Andersen et al. 2011).

Alterations in genomic and genetic cancers

These alterations can be broadly placed into two groups namely, somatic and germline mutations. Two fields that can be added below these groups include copy number variants (CNVs) and loss of

heterozygosity. Copy number variants have been identified to occur as germline constitutional CNVs as well as somatic CNVs (Shlien and Malkin 2009). A CNV can be defined as a deletion/duplication event that causes the repetition of a specific section in the genome where this repetition differs between individuals in the population (Shlien and Malkin 2009). Loss of heterozygosity refers to an event where a whole gene and its surrounding chromosomal material is lost on one chromosome arm or otherwise termed as the somatic loss of the wild-type allele in often-occurring cancer syndromes (Ryland, Doyle et al. 2015)

A mutation is defined as a change in the nucleotide sequence brought on by unrepaired damage to the DNA. Somatic mutations refer to mutations that occur after conception and in any cell of the body except germline cells. Offspring / children of the cancer patient cannot inherit these mutations. Germline mutations are found in the germline (sperms and eggs); the offspring/children can inherit these mutations. Loss of heterozygosity refers to the complete loss of a one copy of a gene and also the surrounding chromosomal regions (Joseph, Darrah et al. 2014).

Different types of mutations that may occur include:

- A missense is when one nucleotide is substituted by another causing the transcription of a different amino acid.
- A nonsense mutation is also a substitution of a nucleotide but the amino acid is changed into a premature stop codon or a nonsense codon. This leads to a much shorter translated protein; in most cases this protein is non-functional.
- Insertions and deletions are simply put, inserted nucleotide/nucleotides into an existing sequence and nucleotide/nucleotides removed from an existing sequence respectively. These changes may cause the resulting protein to not function properly.
- Duplications occur where several nucleotides are duplicated once or more and form part of the same larger sequence.

Insertions, deletions and duplications all may cause a frameshift mutation. The addition/loss of nucleotides causes a frameshift in the original reading frame of the gene leading to different amino acids to be transcribed. All these mutations may lead to a non-functional or impaired translated protein.

## 1.4 TYPES OF CANCER

A variety of different cancers have been identified which are located throughout the body. More than a hundred different cancers which affect the human body have been identified thus far. Because of the diversity of cancer, they are classified by the organ that is afflicted as well as the type of cell from which the tumour cell originates from (NCI 2021). These include:

- Benign tumours, which are not classified as a cancer.
- Blastoma, cancers which develop from unipotent stem cells or precursor cells. Typically found in children rather than adults.
- Carcinoma, cancer of the epithelial cells. These cells can be found in the endodermal/ectodermal germ layers. Also, designated as the most common cancers, especially in adults.
- Germline tumours, cancer afflicting the testicular tissue and the ovaries. But these types of tumours have been identified outside of the gonads as well and are attributed to birth defects.
- Leukaemia/lymphoma, cancer affecting the blood cells, bone marrow and lymphoid system.
- Sarcoma, cancers that originate in cells from the mesenchymal/connective tissue (bone, cartilage, fat, hematopoietic tissue, muscle and vascular tissue).

## 1.5 GENDER-BASED STUDIES

Studies have been done to compare the occurrence rate of different cancers in males and females. With only a few exceptions, most cancers are more prominent in men (Dorak and Karpuzoglu 2012). Exceptions included anus, breast, gallbladder and thyroid cancer which had higher incidences in females (Dorak and Karpuzoglu 2012). Hypopharyngeal and laryngeal cancers were found to be six times more prevalent in males then females in these studies. The different cancers are affected by not only environmental but also by biological and occupational factors. Alcohol consumption is one example given for increased liver cancer activity in men (McCann 2000). The author speculates that biologically, women are more susceptible to thyroid and gallbladder cancers because of their predisposition to chronic inflammation of gallstones and autoimmune diseases (McCann 2000).

## 1.6 ETHNICAL IMPORTANCE OF DIFFERENT CANCERS

A few studies have been done to compare the presence of cancers in different ethnic groups around the world. A study in 1996 had a good coverage of different racial groups including: Alaskan natives, American Indians, African-American, Caucasian, Chinese, Filipino, Hawaiian, Hispanic, Japanese, Korean and Vietnamese individuals (Miller, Kolonel et al. 1996). The authors found that prostate cancer was most prominent in American Indians, Blacks, Caucasian, Filipino, Hispanic and Japanese men. In the remaining groups, lung cancer was the most observed cancer in men. Breast cancer in women was identified with the highest incidence level in all the groups except for the Vietnamese population where cervical cancer outweighed breast cancer (Miller, Kolonel et al. 1996). Overall, cancers were identified most in Blacks followed by Caucasian, Alaska natives and Hawaiian for men. In women, Caucasian had the highest levels of cancer followed by Alaska native and Blacks. Korean men and women proved to have the lowest levels of cancer identified (Miller, Kolonel et al. 1996).

## 1.7 WHAT IS BREAST CANCER?

Breast cancer is the development of cancer in the tissues of the breast (BreastCancer.org 2022). Anatomically, the breast consists out soft tissue, connective and fatty tissues, blood and lymph vessels, milk ducts and lobules, the areola and nipple (Figure 1.2). Cancers have been identified to develop in the ducts, lobules, nipple and soft tissues. Ductal and lobular cancers are the most common breast cancers while the other two are only seen in rare cases but are far more aggressive (BreastCancer.org 2022).

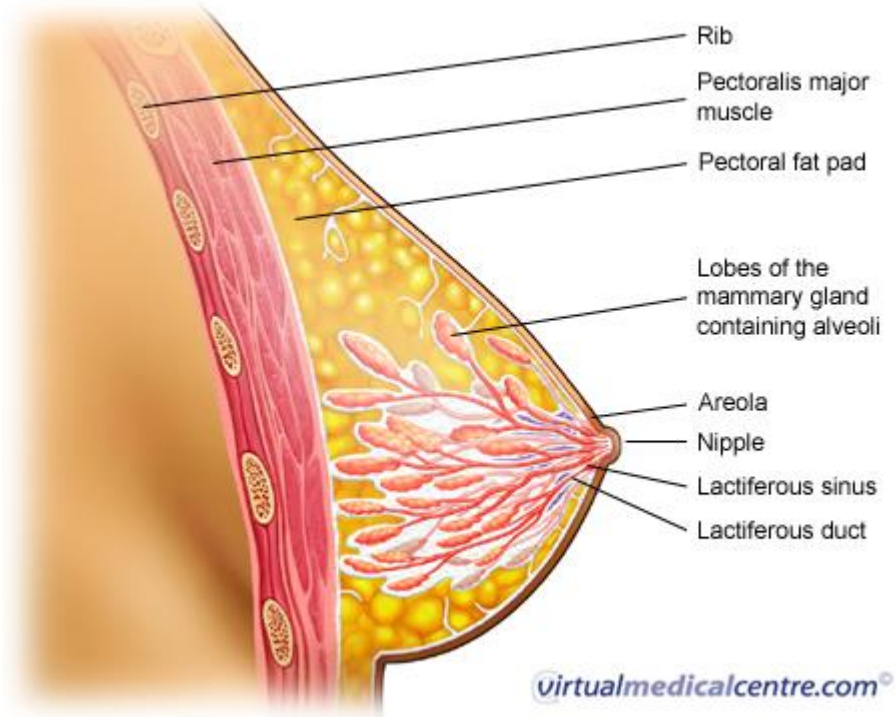


Figure 1.2: Anatomical representation of the breast (https://www.myvmc.com/anatomy/breast).

A worldwide statistical study identified breast cancer to be the most prominent cancer in women and the second most common cancer when compared to others (Ferlay, Soerjomataram et al. 2015). Also, the same study classified breast cancer as the fifth highest in mortality rates for humans but came first when only women are considered (Ferlay, Soerjomataram et al. 2015). A study in 2013 focusing on breast cancer specifically in the United States where a slightly different result was present. Here the authors found that breast cancer was also the second most active cancer but that lung cancer has a higher mortality rate than breast cancer which was placed second in women (DeSantis, Ma et al. 2014).

## 1.8 DIFFERENT TYPES OF BREAST CANCER

As research into breast cancer improved through the years, it has become more apparent how complex the disease truly is. It cannot simply be characterized by clinical parameters such as tumour size, lymph node involvement, histological grade and age (Eroles, Bosch et al. 2012). To further complicate its study, researchers found that breast cancer is not a single disease entity but rather a collection of different diseases. Breast cancer can also be classified according to the presence of certain biomarkers: oestrogen receptor (ER), progesterone receptor (PR) and epidermal growth

factor receptor 2 (HER2), which are routinely used in the diagnosis and treatment of patients (Eroles, Bosch et al. 2012).

Different attempts have been made to classify breast cancers according to their histological characteristics (Ronnov-Jessen, Petersen et al. 1996, Fabbri, Carcangiu et al. 2008) which has led to the following:

### 1.8.1 HISTOLOGICAL BREAST CANCER TYPES

Invasive ductal carcinoma

This cancer is classified as the most common type of breast cancer presently. Cells lining the milk ducts become cancerous. This is then followed by abnormal growth which extends into fatty tissues from where the cancer cells can spread through blood and lymph vessels to the rest of the body (ACS 2010).

Ductal carcinoma in situ (non-invasive)

Also named intraductal carcinoma. Like above, the cancer cells form in the linings of the milk ducts. The difference lies in that these cancer cells do not have the ability to invade/metastasize the rest of the body and are restricted to the breast where they developed originally (ACS 2010).

Invasive lobular carcinoma

Cells from the milk-producing glands (lobules) become cancerous. These cancer cells can also metastasize to the rest of the body (ACS 2010). This is one of the rarer cancers to our knowledge.

Lobular carcinoma in situ (non-invasive)

Similar to its invasive form in development but with the inability to metastasize (ACS 2010).

Inflammatory breast cancer (invasive)

The most aggressive breast cancer known, it is sometimes mistaken for an infection rather than cancer. It differs from other cancers in that a lump is not formed. A mammogram is inefficient at diagnosing it as a cancer in the early stages of development. Instead, cancerous cells block up the lymphatic system and causes swelling and other infection symptoms. This type of cancer usually starts its development in the soft tissue beneath the skin or in the skin cells (ACS 2010).

Paget's disease

This type is classified as cancer of the nipple and areola skin. But this cancer first develops in the ducts and spreads to the nipple and then to the areola. Symptoms include the affected areas appearing crusted, red and scaly almost like eczema. This type is accompanied by an invasive or non-invasive ductal carcinoma, where the invasive form is more aggressive than the non-invasive form. Mastectomy is usually required (ACS 2010).

Table 1.1: Different stages recognized in breast cancer.

| **Stage** | **Definition** |
|---|---|
| Stage 0 | Cancer located in the breast duct, no metastasis into normal adjacent breast tissue. |
| Stage I | Size of cancer tumour is 2 centimetres or less and is confined to the breast without affecting the lymph nodes. |
| Stage IIA | No tumour is present in the breast tissue, but cancer cells are located in the axillary lymph nodes.<br><br>The size of the tumour is 2 centimetres or smaller and has spread to the axillary lymph nodes or the tumour is larger than 2 but no larger than 5 centimetres without spreading to the axillary lymph nodes. |
| Stage IIB | The tumour size is between 2 and 5 centimetres and localized to the axillary lymph nodes or the tumour is larger than 5 centimetres without spreading to the axillary lymph nodes. |
| Stage IIIA | Tumours are not present in the breast tissue. Cancer is found in axillary lymph nodes that are sticking together or to other structures or the cancer is located in the lymph nodes near the breastbone with the tumour ranging in any size. |
| Stage IIIB | Tumours ranging in any size have spread to the chest wall and/or skin of the breast. There is a possibility of spreading to the axillary lymph nodes that are clumped together or sticking to other structures or to lymph nodes near the breastbone. Inflammatory breast cancers are at least stage IIIB. |
| Stage IIIC | There may either be no sign of cancer in the breast tissue or tumours in any size may have spread to the chest wall and/or the skin of the breast and the cancer has spread to lymph nodes either above or below the collarbone, as well as having |

| | |
|---|---|
| | spread to the axillary lymph nodes or those near the breastbone. |
| Stage IV | The cancer has metastasized to other parts of the body. |

In 2000, researchers attempted to use gene expression profiling in conjunction with hierarchical clustering and breast cancer tumour phenotypes to characterize and place them into different intrinsic groups (Perou, Sørlie et al. 2000). They could agree on four different groups: basal-like, ER+/luminal-like, Erb-B2+ and normal breast tissue. One year later, researchers from the same group released an article claiming that the luminal-like group could in fact be subdivided into two distinct groups, subtype A and B (Sorlie, Perou et al. 2001).

### 1.8.2 HISTOLOGICAL TUMOUR GRADES

A recent study in 2014 revealed that using the histological grade of tumours as a prognostic tool still has value when applied to breast cancer (Schwartz, Henson et al. 2014). Their study revealed that, even with changes in tumour size and number of cancer positive lymph nodes the grading system remained a relatively accurate means for prognosis. In 1925, researchers realized the benefit of using histological grading for prognosis for the first time (Greenough 1925), here they identified three factors that played a role in survival of patients: tubule formation, nuclear pleomorphism and hyperchromatism. In the same study, they also divided cancerous tumours into three groups: (I) low malignancy, (II) medium malignancy and (III) high malignancy, which is the grading we still use today. Currently, histological grading is attributed to the differences in mitotic activity, nuclear features and tubular formation found within the tumours (Schwartz, Henson et al. 2014).

### 1.8.3 INTRINSIC BREAST CANCER TUMOURS

Luminal-like

Initially, luminal-like tumours were identified to have a high expression of the luminal component of the breast as well as luminal cytokeratins 8/18 (Perou, Sørlie et al. 2000). They were also found to be positive for ER and PR receptors but negative for the HER2 receptor (Perou, Sørlie et al. 2000). But with the inclusion of the hierarchical clustering information (Sorlie, Perou et al. 2001) and the expression levels of a nuclear cell proliferation marker, Ki67, researchers were able to divide the luminal tumours into two groups, luminal A (low to no Ki67 expression) and luminal B (high Ki67 expression) (Cheang, Chia et al. 2009). A subset of luminal B tumours have also been identified to be

positive for the HER2 receptor (Cheang, Chia et al. 2009). Luminal A tumours tend to be graded as 1/2 while luminal B are more harmful and are graded as 2/3 (Dai, Li et al. 2015).

Basal-like (Triple negative)

This refers to breast cancer tumours which have a high expression of breast basal epithelial cells and keratins markers 5, 6, 14, 17 and EGFR, accompanied by a failure of these tumours to produce oestrogen receptors and co-expressed genes (Perou, Sørlie et al. 2000). These tumours have also been associated with the absence of ER, PR and HER2 and their receptors (Cheang, Chia et al. 2009). Other mutations that have been associated with triple negative have included those of TP53 (Sorlie, Perou et al. 2001, O'Brien, Cole et al. 2010) and BRCA1 (van 't Veer, Dai et al. 2002). The metastasis patterns of basal tumours also differ from other tumour types in that they migrate to visceral organs leaving lymph nodes relatively unaffected (Ho-Yen, Bowen et al. 2012) and have a histological grade of 3 (Sorlie, Perou et al. 2001, O'Brien, Cole et al. 2010).

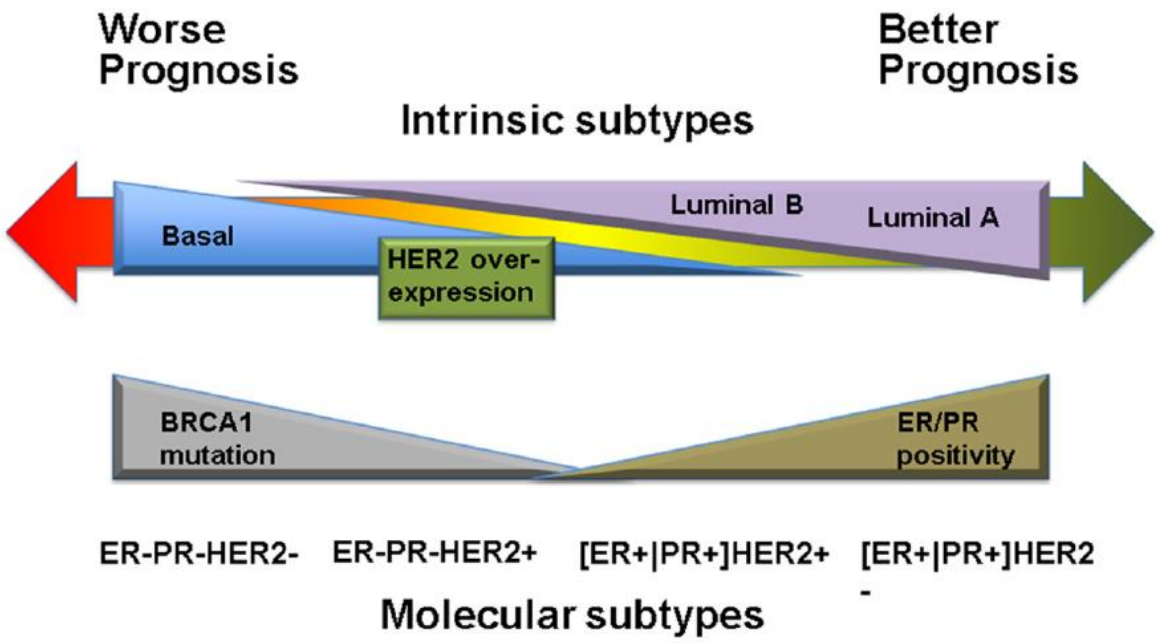


Figure 1.3: Prognosis of breast cancer according to the different molecular subtypes (Dai, Li et al. 2015).

HER2 enriched

Examination of gene expression profiles of HER2 tumours through fluorescence *in situ* hybridization (FISH) revealed it to be negative for ER and PR and positive for HER2 (Vallejos, Gomez et al. 2010). The over-expression of GRB7 (Perou, Sørlie et al. 2000) and PGAP3 (Dai, Chen et al. 2014) genes have been found to accompany these type of tumours. As with basal tumours, TP53 mutations are also

quite common and these tumours also share the same histological tumour grade (Dai, Li et al. 2015). Another similarity shared with basal tumours is the poor prognosis of HER2 tumours, this is linked to the increased probability of relapse if the tumour cells are not completely eradicated.

### 1.8.4 KNOWN MARKERS USED IN THE IDENTIFICATION OF BREAST CANCER

Cancer antigen

CA 15-3: This antigen (also known as MUC1) was classified as the most widely used serum marker in breast cancer. It is a large transmembrane glycoprotein which plays a role in cell adhesion that has been found to be overexpressed in patients with cancer (Bon, von Mensdorff-Pouilly et al. 1997).

CA 27.29: Antigen which is normally expressed at the surface of epithelial cells but is present in malignant (cancerous) epithelial cells of the breast, lung, ovary, pancreas and other tissues. It is more specific for tumour cells and is found to be less glycosylated in cancerous cells compared to the normal epithelial cell (Bon, von Mensdorff-Pouilly et al. 1997).

Patients with metastatic cancer tend to have elevated CA levels in 75% - 90% of cases (Harris, Fritsche et al. 2007). But the American society of clinical oncology (ASCO) warns against the use of serum CA's because of its inconsistency in sensitivity and specificity (Harris, Fritsche et al. 2007).

Carcinoembryonic antigen (CEA)

Another glycoprotein, but this one is produced in the gastrointestinal tract during embryonic development. It also plays a role in cell adhesion. After conception, this antigen is kept at very low levels except in the case of cancer, eg. colon cancer, where the cells lose their basal lamina and multiply uncontrollably. This causes the overproduction of CEA and the accumulation of it in the blood (Hammarstrom 1999). Patients with metastatic cancers tend to have elevated CEA levels in 50% - 60% of cases, which is lower than CA assays (Harris, Fritsche et al. 2007).

Oestrogen receptors (ERs) and progesterone receptors

These receptors are groups of proteins that are activated by the binding of oestrogen and progesterone. The receptors then translocate to a nucleus where they regulate the activity of different genes (Levin 2005).

With ERs, two different classes have been identified: ER and G protein-coupled oestrogen receptor 1 (GPER), the latter has been shown to be involved in tamoxifen resistance development in breast

cancer (Ignatov, Ignatov et al. 2010). ERs (DNA-binding transcription factors) have been shown to be associated with breast, colon, ovarian and prostate cancers.

Progesterone receptors have two forms, A and B, which only differ in a 165 amino acid deletion in A (Kastner, Krust et al. 1990). These receptors play a role in cell growth of breast and uterus tissues. Some research has shown that the +331G/A polymorphism seem to have no association with breast or endometrial cancers (Zhang, Zhang et al. 2003).

DNA flow cytometry-based parameters:

This entails the measurement of DNA levels in a cell to determine for one, the ploidy of the cell (Ormerod 2008). This becomes relevant when measuring tumour cells (S-phase fraction) where the DNA content of the tumour can then be compared to normal diploid cell DNA content.

Human epidermal growth factor receptor 2 (HER2)

Also known as receptor tyrosine-protein kinase erbB-2 (ERBB2), it is one of the more commonly used markers in breast cancer (Mitri, Constantine et al. 2012). The HER2 gene is located on the long arm of chromosome 17 and the protein produced plays an important role as a receptor on breast cells. Overexpression of HER2 leads to the unchecked growth and replication of breast cells which then give rise to cancer development (BreastCancer.org 2016). Currently there are four different tests for HER2: Fluorescence In Situ Hybridization, Immunohistochemistry, Inform Dual In Situ Hybridization and Subtraction Probe Technology Chromogenic In Situ Hybridization.

Tumour protein p53

This gene is also located on chromosome 17 but on the short arm (Isobe, Emanuel et al. 1986). It will give rise to an important protein that plays a role as a tumour suppressor by binding to DNA and regulating gene expression (Mraz, Malinova et al. 2009). When DNA gets damaged TP53 determines if the DNA should be repaired or if the cell containing the damaged DNA should undergo programmed cell death (apoptosis).

Urokinase plasminogen activator (uPA) and Plasminogen activator inhibitor (PAI-1)

Evidence show that increased levels of these tumour-associated proteins has been associated with aggressive forms of cancer. Cancer cells facilitate movement (metastasis) through these glycoproteins to reach the rest of the body (Reuning, Sperl et al. 2003).

Cathepsin D

Studies have shown that like uPA, the levels of CD in a cell may act as a prognostic factor in determining different cancers (Benes, Vetvicka et al. 2008).

Cyclin E

Cyclin E plays a role in regulating several different processes of the cell cycle by phosphorylating downstream proteins (Hinds, Mittnacht et al. 1992). In breast cancer cells, altered cleaved isoforms of cyclin E (33 kDa and 44 kDa) are expressed which have been successfully been used as a prognostic tool (Wingate, Puskas et al. 2009).

Ki67

This protein has been shown to be expressed in proliferating cells but it's true function is unclear, only that it is associated with ribosomal RNA (Bullwinkel, Baron-Lühr et al. 2006). Information on its expression can still be used to identify cells that are proliferating to fast like in the case of cancerous cells (Gerdes, Lemke et al. 1984). Controversy about the use of Ki67 as diagnostic tool still exists and more research is needed to evaluate its use.

Protein kinase C (PKC)

A multitude of mutations (400 different ones) in PKC has been identified in human cancers. PKC has an enzymatic function where it phosphorylates serine and threonine to control the functions of other proteins (Stabel and Parker 1991).

Several assays have been developed to identify different tumour markers, these include: chemiluminescence immunoassay, enzyme-linked immunosorbent assay (ELISA), radioimmunoassay, enzyme immunoassay and electrochemical immunosensors (Li, He et al. 2013).

## 1.9 SPECIALIZED INVASIVE CARCINOMAS

These cancers are carcinoma subtypes which are characterized by their features, eg. arrangement (ACS 2010). These include: adenocystic, adenosquamous, medullary, metaplastic, mixed (invasive ductal and lobular features), mucinous, papillary and tubular carcinomas.

## 1.10 GENDER BASED STUDIES FOR BREAST CANCER

At a young age (9 to 10 years) both boys and girls have similar breast anatomy. The changes come at puberty when the female ovaries produce hormones (oestrogen) that promote breast tissue, duct and lobule growth (ACS 2008). Men also produce small amounts of oestrogen but not enough to

have the same function/effect as in women; the function of oestrogen in men is to promote the maturation of sperm.

Because breast tissue in men also contains ducts and lobules (at far lower levels than women), men can still develop the same breast cancer disorders as women but at a very low probability. Statistics show that women are hundred times more likely to develop breast cancer. Studies have shown that one in a 1 000 men develops breast cancer over the entirety of a lifetime (ACS 2008).

## 1.11 RISK FACTORS FOR BREAST CANCERS

The increased likelihood of developing breast cancer can be directly and indirectly influenced by different factors which can be controlled while others cannot.

Factors that cannot be controlled (BreastCancer.org 2020, ACS 2022):

Gender: The most important risk factor of breast cancer is simply being a woman. Both men and women can develop breast cancer but the rate at which a woman's breast develops and changes puts them at a greater risk for the introduction of factors related to cancer. Oestrogen and progesterone control the development of breast tissue.

Aging: As the body gets older the risk for breast cancer increases. Comparing invasive breast cancers with age researchers found that 1 in 8 women under the age of 45 develop cancer where as women over the age of 55 have a 2/3 chance of developing breast cancer.

Inherited mutations: A variety of genes have been identified that play an important role in the development of breast cancer. Mutated copies of these genes that are inherited by the progeny increase their risk for the development of breast cancer. The most important genes known are *BRCA1* and *BRCA2* (Angeli, Salvi et al. 2020). Women with mutations in these genes have been shown to have an increased risk for cancer as high as 55-65% and 45% respectively. Studies have shown that women with these mutations seem to develop breast cancer at a younger age and in both breasts as well as having a greater disposition to developing other cancers, like ovarian cancer (Angeli, Salvi et al. 2020).

Other genes that also play a role in the risk for cancer include *ATM*, *TP53*, *CHEK2*, *PTEN*, *CDH1*, *STK11* and *PALB2*. These genes do not have such a marked effect as *BRCA1* and *BRCA2* but are still significant.

Family history of breast cancer: Work has shown that where a woman has a blood relative with breast cancer, she herself has an increased risk to develop breast cancer. If a first-degree relative

(mother, daughter, sister) developed breast cancer, studies have shown that a women's likelihood increases 2-fold to also develop breast cancer and 3-fold if the women has two first-degree relatives with breast cancer.

Personal history of breast cancer: Where a woman already developed cancer in one of her breasts, she has an increased risk of developing it in the other breast or in another part of the same breast (this is not metastasis/recurrence of the older cancer). This risk increases if the first cancer was diagnosed at a younger age.

Race and ethnicity: Worldwide studies have identified Caucasian women to be most prone to breast cancer followed by African American women but the latter are at a greater risk of dying from it. Other races including; Asian, Hispanic and Native American are far less at risk to develop breast cancer. The disparities in incidence levels and mortality rates between different ethnical groups can be tied to a variety of reasons, some of which include: stage of breast cancer at diagnosis, socio-economic status, health care availability, biologic and genetic differences in tumours (Ademuyiwa, Groman et al. 2011). For example, African American females are on average diagnosed at an earlier stage in their lives compared to other groups. Compared to the Caucasian population (63), the African American median age of diagnosis is 59 (Howlader, Noone et al. 2014).

Exposure to oestrogen: As mentioned before, oestrogen plays an important role in the development of breast tissue, but prolonged exposure to it without a break increases the risk for breast cancer. This is also true for when a woman starts menstruating at an early age (before the age of 12) or when they go through menopause late in life (55 and older). Oestrogen can also be encountered in the environment and not only be produced in the body. Examples include pesticides and hormones in meat, where oestrogen-like substances are produced when these compounds are broken down in the body.

Exposure to diethylstilbestrol (DES): This drug was administered to pregnant women during the period of 1940-1960 with the possibility to lower the chance of a miscarriage. These women and their progeny have been found to have a predisposition for developing breast cancer.

Dense breast tissue: Several factors influences the density of breast tissue; age, genetics, menopausal status, medications, and pregnancy. The tissue is compromised out of fatty, fibrous and glandular tissues. When the total breast is made up of mostly fibrous and glandular tissue, the breast is classified as denser. Women with dense breast tissue have a 1.2 - 2-fold increased risk of developing breast cancer compared to women with average density. Dense tissue also impairs the diagnosis made by mammograms resulting in false-negative calls.

Benign breast conditions: Divided into three general groups, these conditions might increase the risk for breast cancer;

- Non-proliferative lesions

  Not associated with overgrowth of breast tissue. Research has not been able to find any evidence for its activity in breast cancer risk, but if so, its extent is very little.

- Proliferative lesions without atypia

  Moderate to high overgrowth of duct/lobule cells. These lesions increase the risk for breast cancer by 1.5 - 2-fold.

- Proliferative lesions with atypia

  Also, overgrowth of duct/lobule cells except cells not appearing to be normal anymore. These lesions are more severe; they increase the risk for breast cancer by 3.5 - 5-fold.

Previous chest radiation: Patients who were treated for cancer (chest area) by radiation therapy in their adolescences/youth showed an extremely high risk for developing breast cancer later in their life. The risk is increased even more depending on how young the patients were when they received the radiation treatment. Radiation after the age of 40 did not seem to affect the breast cancer risk for those patients.

Factors that can be controlled (BreastCancer.org 2020, ACS 2022):

Number of pregnancies / offspring and breastfeeding: Overall, pregnancies and breastfeeding lower the risk for developing breast cancer, because it decreases the amount of menstrual cycles a woman has during her lifetime. Women that have children after the age of 30 have been shown to have a slight increase in susceptibility. There is an exception to this, the likelihood to develop a specific breast cancer, such as triple negative breast cancer, seems to increase with the number of pregnancies.

Birth control: Oral and injection (depot-medroxyprogesterone acetate) contraceptives are proven to raise the risk of developing breast cancer while in use. Respectively, after 5 and 10 years the effect of these contraceptives seems to diminish completely reverting the risk back to original levels before their use.

Hormone therapy after menopause: Also, classified as hormone replacement therapy, menopausal hormone therapy and post-menopausal hormone therapy. This therapy is believed to have many beneficial characteristics including relief of menopause symptoms and prevention of osteoporosis. Therapy usually includes administration of both oestrogen and progesterone or only oestrogen. Research implies that only using oestrogen will not increase the risk for breast cancer but a combination of the two hormones would (Mehta, Kling et al. 2021).

Alcohol consumption and smoking: Both these activities have been linked to increased susceptibility to breast cancer as well as other cancers. Even drinking one drink a day compared to non-drinking increases the risk. This risk increases as the amount of alcohol consumption goes up. Smoking has been linked with breast cancer in younger, premenopausal women and even second-hand smoking increases the risk for development of cancer.

Physical activity: Exercising weekly between 4 to 7 hours has been shown to decrease breast cancer risk.

Overweight/obesity: In women, the ovaries and fat tissue produce oestrogen, after menopause, oestrogen is only contributed by the fat tissue. More fat tissue leads to increased levels of oestrogen which has been shown to be linked with increased risk for breast cancer. Also, obesity leads to higher levels of blood insulin which has been linked to development of cancers.

Other factors that may play a role in breast cancer have been identified but better/more research is needed for a definite answer. These included: chemicals in objects and the environment, dietary and exposure to light during the evenings.

## 1.12 ENRICHED PATHWAYS

Pathway studies are based on biological systems of well-studied processes where interactions comprise biochemical reactions, regulation and signaling (Creixell, Reimand et al. 2015). These studies represent consensus systems based on decades of research and the complex cellular activities can be visualized as simplified linear diagrams. This analysis has a variety of benefits to research which include relative to analysing genomics data (Chi, Gribbin et al. 2014).

Because of the aggregation of molecular events through multiple genes, it increases the likelihood that any statistical detection threshold will be passed more easily and it decreases the need for

multiple hypothesis testing. Secondly, interpretation of results is often simpler because the genomic alterations are then coupled to more familiar concepts such as cell cycle or apoptosis. Causal mechanisms may potentially be identified, such as a up/down-regulated transcription factor. The use of pathway information makes research more comparable in a common feature space between different studies. Lastly, it allows for the integration of genomic, transcriptomic and proteomic data into a unified view of cancer which leads to an increased statistical and interpretative power.

Pathway analysis has previously been used to analyse cancer samples successfully. A variety of studies has been done to identify driver genes and cancer pathways (Akavia, Litvin et al. 2010, Danussi, Akavia et al. 2013), common patterns of network alteration (Hoadley, Yau et al. 2014), cancer mechanisms and biomarkers (Danussi, Akavia et al. 2013, Hoadley, Yau et al. 2014) and also, key regulators of cancer-related gene networks (Carro, Lim et al. 2010, Sonabend, Bansal et al. 2014).

## 1.13 BREAST CANCER IN AFRICA

Breast cancer has emerged as a very important disease in developing countries, including Africa (Farmer, Frenk et al. 2010). Such is the situation that steps need to be taken to increase the awareness of the public as well as professionals. More effort needs to be given on patterns of disease presentation, its epidemiology and treatment outcome in these countries. The general problems faced in Africa is that cancer does not always get diagnosed or treated and additionally there is a limited number of professionals working in often poor conditions.

Advanced techniques such as immunohistochemistry and molecular biology rarely show up in African studies (Clement, Famooto et al. 2008). Other problems include poor tumour specimens, inadequate fixation of tumour tissues and fixation materials and the absence of good control practices (Clement, Famooto et al. 2008). In most developing countries mammography services and biopsies are too expensive or completely absent, the use of biopsies has also shown to have adverse effects on the women who undergo them (Howard 1987, Kiguli-Malwadde, Gonzaga et al. 2010). Inadequate reporting and data has led to the misinterpretation of the breast cancer incidence levels as well as mortality rates in the different geographical areas of Africa (Fregene and Newman 2005).

African countries are currently accounting for more than a million new cancer cases a year, with Africa as whole, over 600 000 cancer deaths are reported annually. Statistics suggest that 70% of all new annual cancer cases will be in or from developing countries (Sankaranarayanan 2006). Many

Africans do not have access to cancer screening, early diagnosis, treatment or palliative care because of the distance of availability of these services, the lack of resources and basic infrastructure. Data compiled suggests that cancer currently kills more people than HIV/AIDS, tuberculosis and malaria combined, but even these numbers are underestimated because of cases not being reported.

Breast cancer is usually diagnosed in African females between the ages of 35 and 45 years compared to western woman who are diagnosed much later (fifteen years) in their lives. A study comparing the age of cancer diagnosis of different ethnic groups found that black patients averaged around 57.6 years compared to 62.6 years in white patients. The same study also identified black females to have a lower incidence level overall, but with patients that were younger than 40, the incidence level was 20% higher in black females (Anderson, Rosenberg et al. 2008). These values may be negatively affected by females not reporting or being examined for breast cancer (Parkin, Bray et al. 2005). The mortality rate in sub-Saharan Africa was still high when comparing aggressive tumours within a short period between the onset of symptoms and diagnosis (Fregene and Newman 2005).

Keeping this in mind, researchers speculate that the probability of a woman suffering from cancer, living in Kampala to reach the age of 65 is only 20% lower than that of Western women (Cancer 2003), furthermore the probability that a woman from a developing country, suffering from cancer would live to 65 is extremely low.

Several studies conducted on African breast cancers revealed that they are predominantly hormone receptor-poor (Mavaddat, Rebbeck et al. 2010). Studies show the breast cancer incidence levels in Morocco are estimated at around 22.3%, where cancer also only gets diagnosed at advanced stages (Chaouki and Nel Gueddari 1991). During 2009-2010, 22% of cancer cases in Sudan were accounted for by breast cancer (Saeed, Weng et al. 2014). A steady breast cancer incidence has been witnessed in Nigeria with approximately 27 000 in 2012 (Saibu, James et al. 2017) to approximately 26 000 in 2018 (Ferlay, Ervik et al. 2018).

Breast cancer by itself was responsible for 8.1% of all female cancers between the years of 1974 and 1987 in Tanzania, most of the patients were younger than 30 (Amir, Kitinya et al. 1994), these patients presented with advanced stages. Patients with stage I breast cancer were never diagnosed, rather stage IIIB cancers were the most prominent (Amir, Aziz et al. 1997). In Tunisia, patients suffering from breast cancer are associated with poor survival due to late diagnosis of the disease

(Gao, Shu et al. 2000). In Libya, 68.4% of breast cancer patients were premenopausal and only 31.6% were postmenopausal (Boder, Elmabrouk Abdalla et al. 2011).

Breast cancer has been identified to have a lifetime risk of 1 in 26 women in the different population groups of South Africa. Annually more than 3 000 women die from breast cancer in South Africa [www. cansa.org.za]. More than 60% of cases present with advanced breast cancer. Studies indicate that the highest age-standardized cancer death rates are found in the coloured population (212.5/100 000), followed by the white (198.9), African (126.0) and Asian (121.4) groups (Steyn, Fourie et al. 2006). In South Africa, 5% to 10% of all breast cancers seem to be linked to some form of inheritance. Well known genes, BRCA1 and BRCA2, play an important role in the aetiology of familial breast cancer and have been associated with 19% and 47% respectively of familial breast cancers (Reeves, Yawitch et al. 2004, van Rensburg, van der Merwe et al. 2007).

## 1.14 BREAST CANCER GENOMICS

These studies involve rapidly scanning a set of known markers across a whole genome of many individuals with the result of identifying variations/mutations between the genomes that may be disease-causing. By using this newly mined information, researchers can more accurately devise strategies/procedures to detect, treat and prevent the disease. Genome-wide studies have been shown to be very useful in mapping genetic variations linked to complex disease, such as asthma, cancer, diabetes, heart disease and mental illnesses.

When analysing cancer, researchers have found it to be most informative to sequence the whole genome or protein-coding exome of the cancerous and normal tissue and then compare these sequences (Ormerod 2008). Over time researchers started to recognize that the accumulation of cancerous mutations forms part of a more complex, dynamic process. Exposures to carcinogenic materials as well as DNA repair defects leads to increased levels of mutational rates; ex. chromothripsis and telomere attrition have led to massive genomic rearrangements (Stephens, Greenman et al. 2011).

With the incorporation of multiple genome-wide association studies (GWAS) several less relevant single nucleotide polymorphisms (SNP) were identified and the importance of these SNPs was severely overestimated in explaining the heritability (Turnbull, Ahmed et al. 2010). Recently, a variety of GWAS started to compare cancer susceptible alleles across multiple racial/ethnic groups

but most known SNPs were only validated in Caucasian women (Easton, Pooley et al. 2007, Li, Humphreys et al. 2010, Han, Long et al. 2016). In some cases, follow-up work was conducted to compare replication patterns of the original GWAS in African women (Huo, Zheng et al. 2012, Ruiz-Narvaez, Rosenberg et al. 2013).

As stated before, most cancer studies in GWAS have been conducted on Caucasian women but more specifically those originating from Europe. GWAS research comparing different populations/ethnic groups may lead to the discovery of weakly tagged or uncommon variants, that are not present in the European population. A variant of the H19 gene and oestrogen receptor 1 gene (ESR1) compared to that of European origin has been identified in the Chinese population which is speculated to also increase the risk for breast cancer (Zheng, Long et al. 2009). This variant was replicated in a very small European population but the authors state that a much larger sample size would be needed to determine the effect of the variant on the different populations (Zheng, Long et al. 2009). Studies like this further our knowledge and understanding of the effects of different variants on different populations/ethnic groups.

With the evolution of breast cancer genomics research, new variants and sub-classes of cancer tumours will be discovered. As an example, a different breast cancer intrinsic subtype, known as Claudin-low, was discovered not so long ago in human and mouse tumours (Herschkowitz, Simin et al. 2007) as well as in breast cancer cell line panels (Prat, Parker et al. 2010).

Another GWAS study focussed their work on Caucasian and African American women (O'Brien, Cole et al. 2014). These authors found SNPs for the fibroblast growth factor receptor 2gene (FGFR2) and the TOC high mobility group box family member 3 gene (TOX3) for both groups to be strongly associated with elevated risk for breast cancer. The following SNPs only affected breast cancer risk in Caucasian women: the maternally imprinted expressed transcript gene (H19), the mitochondrial ribosomal protein S30gene (MRPS30), the mitogen activated protein kinase 1 gene (MAP3K1) and the zinc finger, MIZ type-containing 1 gene (ZMIZ1) and SNPs in ESR1 elevated the risk for breast cancer in African Americans (O'Brien, Cole et al. 2014).

The Breast Cancer Association Consortium published a GWAS report on the evidence of genotypic polymorphisms at more than 40 different genomic loci (Bojesen, Pooley et al. 2013, Michailidou, Hall et al. 2013, Zheng, Zhang et al. 2013). Here they found Asian and African women to have very similar

SNPs to those of western women which were classified as high-risk SNPs, as well as SNPs that were specific to the different populations.

## 1.15 RESEARCHING CANCER

When cancer research started it was believed that cancer had to be some kind of alteration of the DNA, either through mutations (McCann and Ames 1976), rearrangements (Sager 1979) or methylation patterns (Holliday 1979). Researchers needed to test how cancerous tissues differed from normal tissues. The first study comparing normal vs. tumour cancer tissue was conducted in 1983 on normal colonic mucosa tissue and a colon adenocarcinoma respectively (Balinsky, Platz et al. 1983).

### 1.15.1 VARIANT ANALYSIS AND IDENTIFICATION

One of the most important ways to identify discrepancies in cancer is using variant analysis. As previously stated, variants may occur in somatic tissues or in the germline cells, which may have different consequences.

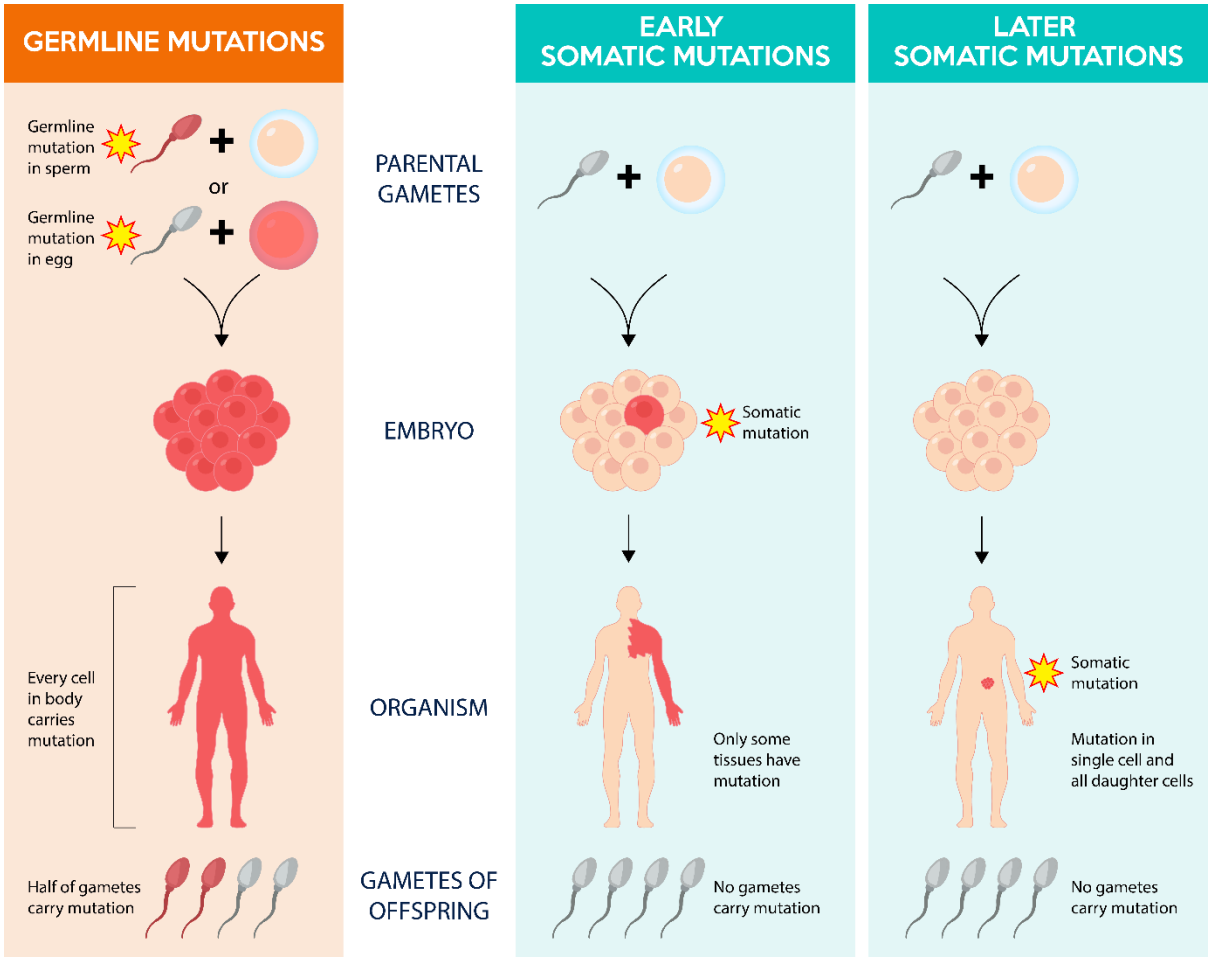


Figure 1.4: Acquisition of germline mutations versus somatic mutations (HHS 2012).

As illustrated in Figure 1.4, the origin of a somatic variant is localized to one tissue or cell which is then carried over to the daughter cells but cannot be inherited by the progeny of the variant carrier. In contrast, germline variants are inherited by progeny and the variant is carried in all the progeny's cells/tissues. Furthermore, for germline testing blood or saliva is mostly frequently used while tumour tissue is used for somatic testing.

Some consideration needs to be taken when comparing germline breast cancer versus somatic breast cancer, each having its own unique set of conditions. Germline variants in breast cancer-related genes ex. *TP53*, will predispose the patient to breast cancer from birth, because the protein produced by this gene plays an important role in suppressing tumours. Currently, 5-10% of breast cancer cases are attributed to inherited gene mutations (Apostolou and Fostira 2013). Somatic breast cancer mutations are sporadic and may in some cases only occur later in a life cycle and only increase the susceptibility to breast cancer from the time of acquisition.

### 1.15.2 GENE EXPRESSION PROFILING

The comparison of normal and cancerous tissue has been able to identify fluctuations in gene expression (Alon, Barkai et al. 1999) as well as miRNA expression (Volinia, Calin et al. 2006) but there was little success when messenger RNA was compared (Lu, Getz et al. 2005). Research has shown that most miRNA irrespective of the cell type had lower expressions in the tumour version when compared to the norm (Lu, Getz et al. 2005), making them very good candidates for breast cancer research. The researchers compiled a large number of miRNA that may be targets in humans. While this approach has shown a lot of promise, a few concerns have been introduced. One of them is the question of what should be classified as normal breast tissue as well as the availability of normal tissue in studies (Srour, Reymond et al. 2008).

More questions about the comparability of normal and cancerous tissue led to the development of the field, gene expression profiling. Basically, it is the study of the expression of many genes at once leading to biological conclusions. Gene expression profiling can broadly be placed into two groups: microarray-based and sequenced-based profiling.

Microarray technology can be defined as an array in the form of a solid support whether it be a glass, nylon, filter or silicon chip to which an arrangement of known short sequences are hybridized as probes. Each sequence usually represents a single specific gene or polymorphism depending on the test. A biological sample which is being tested will then be denatured, labelled and administered to the array. Hybridization between the sample and the probe leaves a fluorescent/radio-active signal that can be visualized. This technology can be used to test for different SNPs in a variety of diseases/phenotypes including cardiovascular disease (Keating, Tischfield et al. 2008), cancer (DeRisi, Penland et al. 1996), pathogens (Vora, Meador et al. 2004), ethnicity (Imai, Kricka et al. 2011) and genome-wide association study (GWAS) analysis (Jia, Wang et al. 2010). It could also be used to investigate copy number variations (Carter 2007), DNA-protein interactions (Buck and Lieb 2004), gene expression profiling (Liang, Morar et al. 2013) and structural variations (Marshall, Noor et al. 2008).

Microarray-based profiling dates back to 1983, when the first computerized image-based analysis was done on RNA (Taub, DeLeo et al. 1983). The concept of microarrays was already being applied in 1975, where researchers cloned DNA of interest into *E. coli* which were grown on agar plates. DNA was lysed from the *E. coli*, denatured and fixed to a filter, where after a radio-actively labelled

DNA/RNA strand would be used to hybridize to complementary strands on the filter for screening purposes (Grunstein and Hogness 1975).

Modern age research (late 90's and 2000's) gave rise to a lot of attention to microarray development. In this period three unique types of arrays came into play:

- Spotted arrays on glass

  A technique where DNA arrays are bound to glass microscope slides instead of filters (DeRisi, Penland et al. 1996). Also, radio-active labels have been replaced with fluorescent labels which is quite sensitive and proved to be more affordable (Bumgarner 2013).

- In situ synthesized arrays

  In 1991, a method was developed where light is used to synthesize an array of different peptides on a solid-phase support using photolabile protecting groups and photolithography (Fodor, Read et al. 1991). This research group in conjunction with the company, Affymetrix, developed microarrays even further to contain up to 256 different octa-nucleotides (Pease, Solas et al. 1994). This method is advantageous over spotted arrays because DNA sequences are synthesized on the solid-phase support during array construction instead of having to synthesize them beforehand (Bumgarner 2013).

- Self-assembled arrays

  The other type of array was developed by David Walt's group at the Tufts University (Michael, Taylor et al. 1998, Ferguson, Steemers et al. 2000, Steemers, Ferguson et al. 2000, Walt 2000). The solid stage phase used in this technique consisted out of a fibre optic array etched with small wells where specialized polystyrene microspheres/beads could bind. DNA would be synthesized on these beads which in the earlier studies were labelled with different fluorophore combinations (Steemers, Ferguson et al. 2000). This type of labelling limited the total number of unique beads that could be used so researchers switched to hybridization of short, fluorescently labelled oligos and detection of them in a sequential manner (Gunderson, Kruglyak et al. 2004). This is also the method adopted by the Illumina company.

Microarrays made it possible for researchers to analyse hundreds/thousands of genes/SNPs simultaneously. Researchers prefer the use of microarrays because of its affordability compared to next generation sequencing (NGS) (Zhao, Fung-Leung et al. 2014). Important limitations of this

technology include it is lack of clear standards for data collection, analysis and validation. Other problems include, variations in amount and quality of the RNA used for the experiment in different studies (Russo, Zegar et al. 2003) and cross-hybridization artefacts (Kukurba and Montgomery 2015).

RNA sequencing (RNA-seq) is a different technology which also has been successfully used to investigate gene expression profiling. This technology uses NGS instead of arrays for their comparisons. An initial NGS study into gene expression incorporated expressed sequence tags (ESTs) to discover a variety of unknown genes in the human brain in 1991 (Adams, Kelley et al. 1991), but was limited using low-throughput methods. To overcome this problem, serial analysis of gene expression (SAGE) and cap analysis gene expression (CAGE) were developed, both being tag-based methods, which enabled a more accurate quantification of expression levels as well as a higher throughput (Velculescu, Zhang et al. 1995). This first use of RNA-seq was documented in 2008 where researchers mapped the yeast transcriptome (Nagalakshmi, Wang et al. 2008).

The principle of an RNA-seq experiment consists out of: isolation of RNA from the target population/group tested, conversion of RNA to complementary DNA (cDNA) using a reverse transcriptase enzyme, preparation of sequencing libraries and sequencing with NGS (Kukurba and Montgomery 2015).

A more in depth look at the RNA-seq procedure reveals a variety of RNAs being isolated, some of which mask the detection of the RNA species in question: precursor messenger RNA (pre-mRNA), mRNA, ribosomal RNA (rRNA) and different noncoding RNA (ncRNA). rRNA alone makes up 95% of the total cellular RNA and if not removed will lower the overall coverage of the sequencing depth (Kukurba and Montgomery 2015). Various steps can be taken to increase the enrichment of the tested RNA species, some may include: inclusion of poly-T oligos in library construction to specifically select for poly-A tailed mRNAs or the use of kits (RiboMinus and RiboZero) to remove excess rRNAs (Kukurba and Montgomery 2015).

Very successful versions of RNA-seq include the MammaPrint and BluePrint assays. The MammaPrint 70-gene assays has been shown to be very efficient into grouping patient tumours into low and high risk of relapse. This helps patients make an important decision in if chemotherapy will be necessary/helpful or not (Glas, Floore et al. 2006). The BluePrint 80-gene assay has the ability to identify intrinsic molecular subtypes of early-stage breast cancers. It takes into account the pathway

genes that are active with different subtypes (basal-, luminal-, and HER2-type) (Krijgsman, Roepman et al. 2012, Whitworth, Stork-Sloots et al. 2014)

Compared to micro-arrays, RNA-seq delivers a far higher coverage of the transcriptome and a greater resolution of transcription complexity, also it has the ability to investigate new unknown transcripts where microarrays cannot (Kukurba and Montgomery 2015). This technology has made it possible to not only investigate mRNA but all the other RNA species already named earlier as well (Kukurba and Montgomery 2015). Limitations that have surfaced thus far include: using the wrong manipulation steps during cDNA library construction of different RNA species would result in erroneous data, it is cheaper to use microarrays than RNA-seq and with rare transcripts RNA-seq needs more sequencing depth for adequate coverage which increases the cost of the process (Wang, Gerstein et al. 2009).

## 1.15.3 CANCER GENOME SEQUENCING

Intensive research into breast cancer over the years has revealed that its far more complex than previously thought and might not be caused by a single mutation but rather an overlap of different gene mutations which increases the risk of cancer development (Buys, Sandbach et al. 2017). With this information, the presence of cancer might be misdiagnosed when only the familial history is considered to determine further testing (Buys, Sandbach et al. 2017).

### 1.15.3.1 Whole genome sequencing

This technique entails the sequencing of the complete DNA or near complete DNA sequence of an organism in one attempt (Ng and Kirkness 2010). In 2003, the first human draft genome was completed at an estimated $300 million (the final draft, technology and tools, estimated $3 billion) (Service 2006). Just three years later, researchers were able to bring down the cost of sequencing a whole genome even further to $20-25 million (Kris A. Wetterstrand 2016) and a more recent study illustrated that the sequencing of a whole genome roughly works out to $600 (Kris A. Wetterstrand 2016).

Whole genome sequencing (WGS) has led to the further development of personalized medicine. More specific, cancer research of an individual's whole genome can help with the proper therapeutic intervention, better chemotherapy treatment (if possible) and administration of medicine which the patient would react most positively to (Ng and Kirkness 2010). While WGS could benefit an

individual, most of the information eluded from the genome is still not understood and not useful until further research is done, also unwanted information (eg. mutations/disease) may be discovered while investigating some other condition (Ng and Kirkness 2010). The sequencing of the whole genome requires significant computational and computing power and more time than other technologies, which tends to cost a lot more to complete (Schneeberger 2014).

#### 1.15.3.2 Germline sequencing

Sequencing linked to the elucidation of inheritable DNA. This technique is focussed on DNA from the patient's blood (or other non-cancerous tissue) which is received from the parents. Germline sequencing has important application with many inheritable disorders including, Huntington's and Alzheimer's disease (Baylis 2017). Early onset cancer has been linked to germline variants numerous times (Schon and Tischkowitz 2018, Stoffel, Koeppe et al. 2018) and cancer research would vastly benefit from germline sequencing which may inform researchers regarding cancer susceptibility.

#### 1.15.3.3 Somatic sequencing

As previously stated, somatic mutations are accumulated in a single person's body during their lifetime and are usually sampled from a tumour (or blood in the case of non-solid tumours). The material is often preserved as formalin-fixed paraffin-embedded sections, which may complicate the sequencing process. The research of somatic variants is complicated by the fact that they are of a mosaic nature and may only be present in very low quantities (Freed, Stevens et al. 2014). With the improvement of sequencing techniques, sequencing accuracy and depth also increased, so it became far simpler to only sequence DNA that a researcher was interested in (Gerstung, Papaemmanuil et al. 2014). Somatic variants are mostly identified through whole genome and whole exome sequencing but this is a costly and time-consuming endeavour. Single-cell sequencing has come forth as a promising technique to study somatic mosaicism (Nawy 2014) and cancer research. Single-cell sequencing has also reliably been used to find somatic copy number variation and retro-transposition events (Baslan, Kendall et al. 2012, Evrony, Cai et al. 2012, Wang, Waters et al. 2014) and warrants further exploration.

#### 1.15.3.4 Whole exome sequencing

As an alternative to WGS, instead of the whole genome which is cluttered with non-coding regions and introns, only the exons could be sequenced (Ng, Turner et al. 2009). Exons are the part of the genome that harbour gene regions which are translated into functional proteins. This technique is called whole exome sequencing (WES) and in most cases researchers are more interested in the functional genome (Ng, Turner et al. 2009). Because the functional genome only takes up roughly 1%

of the whole genome, WES is far cheaper and could be done in a fraction of time compared to WGS. Exons are captured by means of chips containing complimentary sequence tags to the exon regions, followed by elution.

WES has been shown to be invaluable in the study of rare Mendelian diseases (Ng, Buckingham et al. 2010). These diseases in most cases are caused by a very rare genetic variants that only occur in a few individuals. Using a technique like SNP arrays would be ineffective with these variants because of their rarity, pre-selected SNP arrays would miss novel/rare variants (Yang, Bakshi et al. 2015).

1.15.3.5 Panel sequencing

Panel sequencing gives researchers the option to test for hundreds of different genes in one test. The concept behind gene panels is to only sequence genes that are specific to a certain disease/phenotype that is being tested (Saudi Mendeliome 2015). Currently, the most well understood germline breast cancer predisposition genes include: ATM, BRCA1, BRCA2, CDH1, CHEK2, PTEN, STK11 and TP53 but other less important predisposition genes like BARD1, BRIP1, NBN, PALB2 and RAD51C still have no guidelines for genetic testing (Tung, Battelli et al. 2015). This makes it so convenient and more affordable to aggregate all these genes into one test (Walsh, Lee et al. 2010, Tung, Battelli et al. 2015). At this stage a multitude of different panels are available for cancer research. There is a tendency to include the core, well known genes mentioned above but in these panels in most cases this is where panel consensus ends. A specific set of genes are then added to the core ones to make up the panel depending on what type of cancer and which group of patients are under research.

The use of panels can identify a much larger number of patients with an increased risk for breast cancer development than BRCA testing by itself (Tung, Battelli et al. 2015, Thompson, Rowley et al. 2016, Tung, Lin et al. 2016). Compared to whole exome and whole genome sequencing, panels have shown to be more cost effective, have a much shorter turnaround time and the ability to study rare/unique variants more effectively (Rehm 2013, Saudi Mendeliome 2015). Also, panels have increased the clinical sensitivity for many already existing tests (Rehm 2013)

When testing multiple genes, unforeseen results surface which are not anticipated and, in some instances, cannot be explained (Tung, Battelli et al. 2015). Another concern with using panels is the inclusion of the correct genes in the test and the uniformity of its design, there is a possibility that the gene of interest might not even be included (Saudi Mendeliome 2015). It has been speculated that a comprehensive cancer panel screen may need to contain up to 1 000 genes.

### 1.15.4 COPY NUMBER VARIANCE

The consensus has always been that a normal person receives one copy of a gene from each parent, leaving an individual with gene sets of two. Research has identified copy number fluctuations in an upward and/or downward change in different DNA regions throughout the human genome (copy number variance), which could lead to an array of diseases/disorders.

Copy number variance (CNV) may include copy number gain or loss, mosaicism, or loss of heterozygosity (LOH) and can be classified into different copy numbers ranges: normal (2), Gain (<4), Loss (>0), High Copy Gain (>4) and Homozygous Copy Loss (<0). It is important to be able to distinguish between benign and pathogenic/high risk aberrations. This is done by analysing and comparing CNVs from apparently healthy/normal patients from different ethnic groups (Lee, Iafrate et al. 2007). Over the years several comprehensive CNV maps of the human genome have been constructed which contains both benign and pathogenic variants which are also used for comparison (Redon, Ishikawa et al. 2006, Conrad, Pinto et al. 2010, Park, Kim et al. 2010, Vlachopoulou 2011, Zarrei, MacDonald et al. 2015).

The term, copy number variance (CNV), has in the past and present been wrongly used as a synonym for other close-fitting/similar terms such as copy number alterations/aberrations or structural variants, many articles use these terms interchangeably. CNVs refer to copy number changes that have taken place in the germline cells and can be found throughout the body (Li, Lee et al. 2009). Copy number alterations and copy number aberrations are in fact synonyms and refer to copy number changes that have arisen in the somatic tissues and are localized to certain cells in the body (Li, Lee et al. 2009).

Previously, other techniques were used to test for chromosome aberrations, including: karyotyping (Caspersson, Zech et al. 1970), fluorescence in situ hybridization (FISH) (Yoon, Xuan et al. 2009) and comparative genomic hybridization (Kallioniemi, Kallioniemi et al. 1992) but CNV sequencing has shown to produce higher resolutions and to be better in detecting and quantifying CNVs (Liang, Peng et al. 2014). CNV sequencing has the ability to detect CNVs in segments < 1 kbp which is far more accurate than other techniques (Tuzun, Sharp et al. 2005).

CNVs have been linked to disease phenotypes as early as 1991 (Lupski, de Oca-Luna et al. 1991). Evidence also supports the contribution of CNVs to the development of some psychiatric disorders, such as: autism (Sebat, Lakshmi et al. 2007), bipolar disorder (Wilson, Flibotte et al. 2006) and schizophrenia (Wilson, Flibotte et al. 2006, Walsh, McClellan et al. 2008). CNVs may play an important role in the development of these disorders.

Loss of heterozygosity:

A normal diploid organism receives one copy of a gene from each parent so in other words you also receive one copy of an allele from each parent which carries these genes, loss of heterozygosity entails the loss of a part of the one allele containing genes and regions around them (Ryland, Doyle et al. 2015). The term, LOH is derived from the apparent change from heterozygosity to homozygosity when comparing germline cells to somatic cells respectively (Cavenee, Dryja et al. 1983).

LOH is a collection of copy number losses (CNL-LOH) and copy number neutral LOH (CNN-LOH), where the latter entails a LOH event brought on by a homologous recombination event ("gene conversion") or because of the presence of a duplicated normal chromosome (Ryland, Doyle et al. 2015). A well-known hypothesis has been proposed in the past to explain the complete inactivation of tumour suppressor genes: the Knudson two-hit hypothesis.

It claims that the inactivation occurs in two steps, where the first is a somatic mutation in one of the tumour suppressor genes. The patient usually does not develop cancer at this point because of a still-functional copy being present. The second hit is a deletion/mutation that inactivates the remaining functional copy which then leads to cancer development (Knudson 1971). LOH has some application in identifying novel tumour suppressor genes and is more frequently being used as biomarkers.

Mosaicism:

Mosaicism occurs when two or more genetically distinct populations of cells are present in the same human body (De Marchi, Carbonara et al. 1976). Genetic mosaicism can be brought on in various ways: A mutation can occur during cell development and the mutation is carried on by the daughter cells (De Marchi, Carbonara et al. 1976), chromosome non-disjunction, anaphase lag and endoreplication (Fitzgerald, Donald et al. 1979).

## 1.16 AIMS AND OBJECTIVES

The aim of the project was to evaluate the presence of different known and novel breast cancer variants related to breast cancer susceptibility in germline samples in women diagnosed previously with breast cancer from a southern African black population. This study was undertaken to increase our knowledge on breast cancer in southern Africa where little research has been done so far.

- Objective 1: To identify previously reported pathogenic / likely pathogenic variants in breast cancer susceptibility genes in a female black South African population (Chapter 3).
- Objective 2: To analyse pathogenic variants in hereditary cancer predisposition genes exclusively investigated for truncating variants, and variants of unknown significance (Chapter 4).
- Objective 3: To analyse non-coding variants that may play some role in cancer susceptibility (Chapter 5).

## 1.17 REFERENCES

ACS. (2008, 01/26/2016). "Breast cancer in men." from http://www.cancer.org/cancer/breastcancerinmen/detailedguide/breast-cancer-in-men-what-is-breast-cancer-in-men.

ACS. (2009, 6/12/2014 ). "The history of cancer." from www.cancer.org/acs/groups/cid/documents/webcontent/002048-pdf.pdf.

ACS. (2010, 08/18/2016). "Types of breast cancer." from http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-breast-cancer-types.

ACS. (2022, 08/18/2016). "Risk factors for breast cancer." from http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors.

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science **252**(5013): 1651-1656.

Ademuyiwa, F. O., A. Groman, T. O'Connor, C. Ambrosone, N. Watroba and S. B. Edge (2011). "Impact of body mass index on clinical outcomes in triple-negative breast cancer." (1097-0142 (Electronic)).

Akavia, U. D., O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway and D. Pe'er (2010). "An integrated approach to uncover drivers of cancer." Cell **143**(6): 1005-1017.

Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." Proc Natl Acad Sci U S A **96**(12): 6745-6750.

Amir, H., M. R. Aziz, C. K. Makwaya and S. Jessani (1997). "TNM classification and breast cancer in an African population: a descriptive study." Cent Afr J Med **43**(1): 357-359.

Amir, H., J. N. Kitinya and D. M. Parkin (1994). "A comparative study of carcinoma of the breast in an African population." East Afr Med J **71**(1): 215-218.

Anderson, W. F., P. S. Rosenberg, I. Menashe, A. Mitani and R. M. Pfeiffer (2008). "Age-related crossover in breast cancer incidence rates between black and white ethnic groups." J Natl Cancer Inst **100**(24): 1804-1814.
Angeli, D., S. Salvi and G. Tedaldi (2020). "Genetic Predisposition to Breast and Ovarian Cancers: How Many and Which Genes to Test?" Int J Mol Sci **21**(3).
Apostolou, P. and F. Fostira (2013). "Hereditary breast cancer: the era of new susceptibility genes." BioMed research international **2013**: 747318-747318.
Balinsky, D., C. E. Platz and J. W. Lewis (1983). "Isozyme patterns of normal, benign, and malignant human breast tissues." Cancer Res **43**(12 Pt 1): 5895-5901.
Baslan, T., J. Kendall, L. Rodgers, H. Cox, M. Riggs, A. Stepansky, J. Troge, K. Ravi, D. Esposito, B. Lakshmi, M. Wigler, N. Navin and J. Hicks (2012). "Genome-wide copy number analysis of single cells." Nature protocols **7**(6): 1024-1041.
Baylis, F. (2017). "Human germline genome editing and broad societal consensus." Nature Human Behaviour **1**(6): 0103.
Benes, P., V. Vetvicka and M. Fusek (2008). "Cathepsin D--many functions of one aspartic protease." Crit Rev Oncol Hematol **68**(1): 12-28.
Boder, J. M., F. B. Elmabrouk Abdalla, M. A. Elfageih, A. Abusaa, A. Buhmeida and Y. Collan (2011). "Breast cancer patients in Libya: Comparison with European and central African patients." Oncol Lett **2**(2): 323-330.
Bojesen, S. E., K. A. Pooley, S. E. Johnatty, J. Beesley, K. Michailidou, J. P. Tyrer, S. L. Edwards, H. A. Pickett, H. C. Shen, C. E. Smart, K. M. Hillman, P. L. Mai, K. Lawrenson, M. D. Stutz, Y. Lu, R. Karevan, N. Woods, R. L. Johnston, J. D. French, X. Chen, M. Weischer, S. F. Nielsen, M. J. Maranian, M. Ghoussaini, S. Ahmed, C. Baynes, M. K. Bolla, Q. Wang, J. Dennis, L. McGuffog, D. Barrowdale, A. Lee, S. Healey, M. Lush, D. C. Tessier, D. Vincent, F. Bacot, S. Australian Cancer, S. Australian Ovarian Cancer, C. Kathleen Cuningham Foundation Consortium for Research into Familial Breast, I. Gene Environment, C. Breast, S. Swedish Breast Cancer, B. Hereditary, N. Ovarian Cancer Research Group, B. Epidemiological study of, B. M. Carriers, B. M. C. Genetic Modifiers of Cancer Risk in, I. Vergote, S. Lambrechts, E. Despierre, H. A. Risch, A. Gonzalez-Neira, M. A. Rossing, G. Pita, J. A. Doherty, N. Alvarez, M. C. Larson, B. L. Fridley, N. Schoof, J. Chang-Claude, M. S. Cicek, J. Peto, K. R. Kalli, A. Broeks, S. M. Armasu, M. K. Schmidt, L. M. Braaf, B. Winterhoff, H. Nevanlinna, G. E. Konecny, D. Lambrechts, L. Rogmann, P. Guenel, A. Teoman, R. L. Milne, J. J. Garcia, A. Cox, V. Shridhar, B. Burwinkel, F. Marme, R. Hein, E. J. Sawyer, C. A. Haiman, S. Wang-Gohrke, I. L. Andrulis, K. B. Moysich, J. L. Hopper, K. Odunsi, A. Lindblom, G. G. Giles, H. Brenner, J. Simard, G. Lurie, P. A. Fasching, M. E. Carney, P. Radice, L. R. Wilkens, A. Swerdlow, M. T. Goodman, H. Brauch, M. Garcia-Closas, P. Hillemanns, R. Winqvist, M. Durst, P. Devilee, I. Runnebaum, A. Jakubowska, J. Lubinski, A. Mannermaa, R. Butzow, N. V. Bogdanova, T. Dork, L. M. Pelttari, W. Zheng, A. Leminen, H. Anton-Culver, C. H. Bunker, V. Kristensen, R. B. Ness, K. Muir, R. Edwards, A. Meindl, F. Heitz, K. Matsuo, A. du Bois, A. H. Wu, P. Harter, S. H. Teo, I. Schwaab, X. O. Shu, W. Blot, S. Hosono, D. Kang, T. Nakanishi, M. Hartman, Y. Yatabe, U. Hamann, B. Y. Karlan, S. Sangrajrang, S. K. Kjaer, V. Gaborieau, A. Jensen, D. Eccles, E. Hogdall, C. Y. Shen, J. Brown, Y. L. Woo, M. Shah, M. A. Azmi, R. Luben, S. Z. Omar, K. Czene, R. A. Vierkant, B. G. Nordestgaard, H. Flyger, C. Vachon, J. E. Olson, X. Wang, D. A. Levine, A. Rudolph, R. P. Weber, D. Flesch-Janys, E. Iversen, S. Nickels, J. M. Schildkraut, S. Silva Idos, D. W. Cramer, L. Gibson, K. L. Terry, O. Fletcher, A. F. Vitonis, C. E. van der Schoot, E. M. Poole, F. B. Hogervorst, S. S. Tworoger, J. Liu, E. V. Bandera, J. Li, S. H. Olson, K. Humphreys, I. Orlow, C. Blomqvist, L. Rodriguez-Rodriguez, K. Aittomaki, H. B. Salvesen, T. A. Muranen, E. Wik, B. Brouwers, C. Krakstad, E. Wauters, M. K. Halle, H. Wildiers, L. A. Kiemeney, C. Mulot, K. K. Aben, P. Laurent-Puig, A. M. Altena, T. Truong, L. F. Massuger, J. Benitez, T. Pejovic, J. I. Perez, M. Hoatlin, M. P. Zamora, L. S. Cook, S. P. Balasubramanian, L. E. Kelemen, A. Schneeweiss, N. D. Le, C. Sohn, A. Brooks-Wilson, I. Tomlinson, M. J. Kerin, N. Miller, C. Cybulski, B. E. Henderson, J. Menkiszak, F. Schumacher, N. Wentzensen, L. Le Marchand, H. P. Yang, A. M. Mulligan, G. Glendon, S. A. Engelholm, J. A. Knight, C. K. Hogdall, C. Apicella, M. Gore, H. Tsimiklis, H. Song, M. C. Southey, A.

Jager, A. M. den Ouweland, R. Brown, J. W. Martens, J. M. Flanagan, M. Kriege, J. Paul, S. Margolin, N. Siddiqui, G. Severi, A. S. Whittemore, L. Baglietto, V. McGuire, C. Stegmaier, W. Sieh, H. Muller, V. Arndt, F. Labreche, Y. T. Gao, M. S. Goldberg, G. Yang, M. Dumont, J. R. McLaughlin, A. Hartmann, A. B. Ekici, M. W. Beckmann, C. M. Phelan, M. P. Lux, J. Permuth-Wey, B. Peissel, T. A. Sellers, F. Ficarazzi, M. Barile, A. Ziogas, A. Ashworth, A. Gentry-Maharaj, M. Jones, S. J. Ramus, N. Orr, U. Menon, C. L. Pearce, T. Bruning, M. C. Pike, Y. D. Ko, J. Lissowska, J. Figueroa, J. Kupryjanczyk, S. J. Chanock, A. Dansonka-Mieszkowska, A. Jukkola-Vuorinen, I. K. Rzepecka, K. Pylkas, M. Bidzinski, S. Kauppila, A. Hollestelle, C. Seynaeve, R. A. Tollenaar, K. Durda, K. Jaworska, J. M. Hartikainen, V. M. Kosma, V. Kataja, N. N. Antonenkova, J. Long, M. Shrubsole, S. Deming-Halverson, A. Lophatananon, P. Siriwanarangsan, S. Stewart-Brown, N. Ditsch, P. Lichtner, R. K. Schmutzler, H. Ito, H. Iwata, K. Tajima, C. C. Tseng, D. O. Stram, D. van den Berg, C. H. Yip, M. K. Ikram, Y. C. Teh, H. Cai, W. Lu, L. B. Signorello, Q. Cai, D. Y. Noh, K. Y. Yoo, H. Miao, P. T. Iau, Y. Y. Teo, J. McKay, C. Shapiro, F. Ademuyiwa, G. Fountzilas, C. N. Hsiung, J. C. Yu, M. F. Hou, C. S. Healey, C. Luccarini, S. Peock, D. Stoppa-Lyonnet, P. Peterlongo, T. R. Rebbeck, M. Piedmonte, C. F. Singer, E. Friedman, M. Thomassen, K. Offit, T. V. Hansen, S. L. Neuhausen, C. I. Szabo, I. Blanco, J. Garber, S. A. Narod, J. N. Weitzel, M. Montagna, E. Olah, A. K. Godwin, D. Yannoukakos, D. E. Goldgar, T. Caldes, E. N. Imyanitov, L. Tihomirova, B. K. Arun, I. Campbell, A. R. Mensenkamp, C. J. van Asperen, K. E. van Roozendaal, H. Meijers-Heijboer, J. M. Collee, J. C. Oosterwijk, M. J. Hooning, M. A. Rookus, R. B. van der Luijt, T. A. Os, D. G. Evans, D. Frost, E. Fineberg, J. Barwell, L. Walker, M. J. Kennedy, R. Platte, R. Davidson, S. D. Ellis, T. Cole, B. Bressac-de Paillerets, B. Buecher, F. Damiola, L. Faivre, M. Frenay, O. M. Sinilnikova, O. Caron, S. Giraud, S. Mazoyer, V. Bonadona, V. Caux-Moncoutier, A. Toloczko-Grabarek, J. Gronwald, T. Byrski, A. B. Spurdle, B. Bonanni, D. Zaffaroni, G. Giannini, L. Bernard, R. Dolcetti, S. Manoukian, N. Arnold, C. Engel, H. Deissler, K. Rhiem, D. Niederacher, H. Plendl, C. Sutter, B. Wappenschmidt, A. Borg, B. Melin, J. Rantala, M. Soller, K. L. Nathanson, S. M. Domchek, G. C. Rodriguez, R. Salani, D. G. Kaulich, M. K. Tea, S. S. Paluch, Y. Laitman, A. B. Skytte, T. A. Kruse, U. B. Jensen, M. Robson, A. M. Gerdes, B. Ejlertsen, L. Foretova, S. A. Savage, J. Lester, P. Soucy, K. B. Kuchenbaecker, C. Olswold, J. M. Cunningham, S. Slager, V. S. Pankratz, E. Dicks, S. R. Lakhani, F. J. Couch, P. Hall, A. N. Monteiro, S. A. Gayther, P. D. Pharoah, R. R. Reddel, E. L. Goode, M. H. Greene, D. F. Easton, A. Berchuck, A. C. Antoniou, G. Chenevix-Trench and A. M. Dunning (2013). "Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer." Nat Genet **45**(4): 371-384, 384e371-372.

Bon, G. G., S. von Mensdorff-Pouilly, P. Kenemans, G. J. van Kamp, R. A. Verstraeten, J. Hilgers, S. Meijer and J. B. Vermorken (1997). "Clinical and technical evaluation of ACSYBR serum assay of MUC1 gene-derived glycoprotein in breast cancer, and comparison with CA 15-3 assays." Clin Chem **43**(4): 585-593.

BreastCancer.org. (2016, 16/05/2016). "HER2 Status." Retrieved 30/08/2016, from http://www.breastcancer.org/symptoms/diagnosis/her2.

BreastCancer.org. (2020). "Lower your risk." from http://www.breastcancer.org/risk.

BreastCancer.org. (2022). "Breast Cancer." from http://www.cancer.gov/types/breast.

Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics **83**(3): 349-360.

Bullwinkel, J., B. Baron-Lühr, A. Lüdemann, C. Wohlenberg, J. Gerdes and T. Scholzen (2006). "Ki-67 protein is associated with ribosomal RNA transcription in quiescent and proliferating cells." J Cell Physiol **206**(3): 624-635.

Bumgarner, R. (2013). "DNA microarrays: Types, Applications and their future." Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] **0 22**: Unit-22.21.

Buys, S. S., J. F. Sandbach, A. Gammon, G. Patel, J. Kidd, K. L. Brown, L. Sharma, J. Saam, J. Lancaster and M. B. Daly (2017). "A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes." Cancer.

Cancer, I. A. f. R. o. (2003). "Cancer in Africa: epidemiology and prevention." IARC Scientific Publications **153**(1): 1-414.

Carro, M. S., W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano and A. Iavarone (2010). "The transcriptional network for mesenchymal transformation of brain tumours." Nature **463**(7279): 318-325.
Carter, N. P. (2007). "Methods and strategies for analyzing copy number variation using DNA microarrays." Nat Genet.
Caspersson, T., L. Zech and C. Johansson (1970). "Differential binding of alkylating fluorochromes in human chromosomes." Exp Cell Res **60**(3): 315-319.
Cavenee, W. K., T. P. Dryja, R. A. Phillips, W. F. Benedict, R. Godbout, B. L. Gallie, A. L. Murphree, L. C. Strong and R. L. White (1983). "Expression of recessive alleles by chromosomal mechanisms in retinoblastoma." Nature **305**(5937): 779-784.
Chaouki, N. and B. Nel Gueddari (1991). "Epidemiological descriptive approach of cancer in Morocco through the activity of the National Institute of Oncology. 1986-7." Bulletin du cancer **78**(7): 603-609.
Cheang, M. C. U., S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard and J. S. Parker (2009). "Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer." Journal of the National Cancer Institute **101**(10): 736-750.
Chi, Y. Y., M. J. Gribbin, J. L. Johnson and K. E. Muller (2014). "Power calculation for overall hypothesis testing with high-dimensional commensurate outcomes." Stat Med **33**(5): 812-827.
Chin, L., J. N. Andersen and P. A. Futreal (2011). "Cancer genomics: from discovery science to personalized medicine." Nat Med **17**: 297-302.
Clement, A. A., A. Famooto, T. O. Ogundiran, T. Aniagwu, C. Nkwodimmah and E. E. Akang (2008). "Immunohistochemical and molecular subtypes of breast cancer in Nigeria." Breast Cancer Res Treat **110**(1): 183-188.
Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Wellcome Trust Case Control, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer and M. E. Hurles (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-712.
Creixell, P., J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. F. Ouellette, A. Valencia, G. D. Bader, P. C. Boutros, J. M. Stuart, R. Linding, N. Lopez-Bigas, L. D. Stein and M. C. P. A. w. g. o. t. I. C. G. Consortium (2015). "Pathway and network analysis of cancer genomes." Nature Methods **12**: 615.
Dai, X., A. Chen and Z. Bai (2014). "Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling." Sci Rep **4**: 6566.
Dai, X., T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan and B. Shi (2015). "Breast cancer intrinsic subtype classification, clinical use and future trends." Am J Cancer Res **5**(10): 2929-2943.
Danussi, C., U. D. Akavia, F. Niola, A. Jovic, A. Lasorella, D. Pe'er and A. Iavarone (2013). "RHPN2 drives mesenchymal transformation in malignant glioma by triggering RhoA activation." Cancer Res **73**(16): 5140-5150.
De Marchi, M., A. O. Carbonara, F. Carozzi, F. Massara, L. Belforte, G. M. Molinatti, D. Bisbocci, M. P. Passarino and G. Palestro (1976). "True hermaphroditism with XX/XY sex chromosome mosaicism: report of a case." Clin Genet **10**(5): 265-272.
DeRisi, J., L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su and J. M. Trent (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-460.
DeSantis, C., J. Ma, L. Bryan and A. Jemal (2014). "Breast cancer statistics, 2013." CA Cancer J Clin **64**(1): 52-62.
Dorak, M. T. and E. Karpuzoglu (2012). "Gender differences in cancer susceptibility: an inadequately addressed issue." Front Genet **3**: 268.
Easton, D. F., K. A. Pooley, A. M. Dunning, P. D. P. Pharoah, D. Thompson, D. G. Ballinger, J. P. Struewing, J. Morrison, H. Field, R. Luben, N. Wareham, S. Ahmed, C. S. Healey, R. Bowman, t. S.

collaborators, K. B. Meyer, C. A. Haiman, L. K. Kolonel, B. E. Henderson, L. Le Marchand, P. S. Brennan, S., V. Gaborieau, F. Odefrey, C. Shen, P. Wu, W. Wang, D. Eccles, D. G. Evans, J. Peto, O. Fletcher, N. Johnson, S. Seal, M. R. Stratton, N. Rahman, G. Chenevix-Trench, S. E. Bojesen, B. G. Nordestgaard, C. K. Axelsson, M. Garcia-Closas, L. Brinton, S. Chanock, J. Lissowska, B. Peplonska, H. Nevanlinna, R. Fagerholm, H. Eerola, D. Kang, K. Yoo, D. Noh, S. Ahn, D. J. Hunter, S. E. Hankinson, D. G. Cox, P. Hall, S. Wedren, J. Liu, Y. Low, N. Bogdanova, P. Schürmann, T. Dörk, R. A. E. M. Tollenaar, C. E. Jacobi, P. Devilee, J. G. M. Klijn, A. J. Sigurdson, M. M. Doody, B. H. Alexander, J. Zhang, A. Cox, I. W. Brock, G. MacPherson, M. W. R. Reed, F. J. Couch, E. L. Goode, J. E. Olson, H. Meijers-Heijboer, A. van den Ouweland, A. Uitterlinden, F. Rivadeneira, R. L. Milne, G. Ribas, A. Gonzalez-Neira, J. Benitez, J. L. Hopper, M. McCredie, M. Southey, G. G. Giles, C. Schroen, C. Justenhoven, H. Brauch, U. Hamann, Y. Ko, A. B. Spurdle, J. Beesley, X. Chen, kConFab, A. M. Group, A. Mannermaa, V. Kosma, V. Kataja, J. Hartikainen, N. E. Day, D. R. Cox and B. A. J. Ponder (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." Nature **447**(1): 1087-1094.

Eroles, P., A. Bosch, J. A. Pérez-Fidalgo and A. Lluch (2012). "Molecular biology in breast cancer: intrinsic subtypes and signaling pathways." Cancer treatment reviews **38**(6): 698-707.

Evrony, G. D., X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, P. J. Park and C. A. Walsh (2012). "Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain." Cell **151**(3): 483-496.

Fabbri, A., M. L. Carcangiu and A. Carbone (2008). Histological classification of breast cancer. Breast Cancer, Springer**:** 3-14.

Farmer, P., J. Frenk, F. M. Knaul, L. N. Shulman, G. Alleyne, L. Armstrong, R. Atun, D. Blayney, L. Chen, R. Feachem, M. Gospodarowicz, J. Gralow, S. Gupta, A. Langer, J. Lob-Levyt, C. Neal, A. Mbewu, D. Mired, P. Piot, K. S. Reddy, J. D. Sachs, M. Sarhan and J. R. Seffrin (2010). "Expansion of cancer care and control in countries of low and middle income: a call to action." Lancet **376**(9747): 1186-1193.

Ferguson, J. A., F. J. Steemers and D. R. Walt (2000). "High-density fiber-optic DNA random microsphere array." Anal Chem **72**(22): 5618-5624.

Ferlay, J., M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram and F. Bray. (2018). "Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer." 2020, from https://gco.iarc.fr/today.

Ferlay, J., I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman and F. Bray (2015). "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012." Int J Cancer **136**(5): E359-386.

Fitzgerald, P. H., R. A. Donald and R. L. Kirk (1979). "A true hermaphrodite dispermic chimera with 46,XX and 46,XY karyotypes." Clin Genet **15**(1): 89-96.

Fodor, S. P., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu and D. Solas (1991). "Light-directed, spatially addressable parallel chemical synthesis." Science **251**(4995): 767-773.

Freed, D., E. L. Stevens and J. Pevsner (2014). "Somatic mosaicism in the human genome." Genes **5**(4): 1064-1094.

Fregene, A. and L. A. Newman (2005). "Breast cancer in sub-Saharan Africa: how does it relate to breast cancer in African-American women?" Cancer **103**(8): 1540-1550.

Gao, Y., X. Shu, Q. Dai, J. D. Potter, L. A. Brinton, W. Wen, T. A. Sellers, L. H. Kushi, Z. Ruan, R. M. Bostick, F. Jin and W. Zheng (2000). "Association of menstrual and reproductive factors with breast cancer risk: results from the Shanghai breast cancer study." Int J Cancer **87**(1): 295-300.

Gerdes, J., H. Lemke, H. Baisch, H. H. Wacker, U. Schwab and H. Stein (1984). "Cell cycle analysis of a cell proliferation associated human nuclear antigen defined by the monoclonal antibody Ki-67." J Immunol **133**(1): 1710-1715.

Gerstung, M., E. Papaemmanuil and P. J. Campbell (2014). "Subclonal variant calling with multiple samples and prior knowledge." Bioinformatics **30**(9): 1198-1204.

Glas, A. M., A. Floore, L. J. Delahaye, A. T. Witteveen, R. C. Pover, N. Bakx, J. S. Lahti-Domenici, T. J. Bruinsma, M. O. Warmoes, R. Bernards, L. F. Wessels and L. J. Van't Veer (2006). "Converting a breast cancer microarray signature into a high-throughput diagnostic test." (1471-2164 (Electronic)).

Greenough, R. B. (1925). "Varying Degrees of Malignancy in Cancer of the Breast." The Journal of Cancer Research **9**(4): 453-463.
Grunstein, M. and D. S. Hogness (1975). "Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene." Proc Natl Acad Sci U S A **72**(10): 3961-3965.
Gunderson, K. L., S. Kruglyak, M. S. Graige, F. Garcia, B. G. Kermani, C. Zhao, D. Che, T. Dickinson, E. Wickham, J. Bierle, D. Doucet, M. Milewski, R. Yang, C. Siegmund, J. Haas, L. Zhou, A. Oliphant, J. B. Fan, S. Barnard and M. S. Chee (2004). "Decoding randomly ordered DNA arrays." Genome Res **14**(5): 870-877.
Hammarstrom, S. (1999). "The carcinoembryonic antigen CEA family: structures, suggested functions and expression in normal and malignant tissues." Cancer Biol **9**(1): 67-81.
Han, M. R., J. Long, J. Y. Choi, S. K. Low, S. S. Kweon, Y. Zheng, Q. Cai, J. Shi, X. Guo, K. Matsuo, M. Iwasaki, C. Y. Shen, M. K. Kim, W. Wen, B. Li, A. Takahashi, M. H. Shin, Y. B. Xiang, H. Ito, Y. Kasuga, D. Y. Noh, K. Matsuda, M. H. Park, Y. T. Gao, H. Iwata, S. Tsugane, S. K. Park, M. Kubo, X. O. Shu, D. Kang and W. Zheng (2016). "Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci." Hum Mol Genet.
Harris, L., H. Fritsche, R. Mennel, L. Norton, P. Ravdin, S. Taube, M. R. Somerfield, D. F. Hayes, R. C. Bast, Jr. and O. American Society of Clinical (2007). "American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer." J Clin Oncol **25**(33): 5287-5212.
Herschkowitz, J. I., K. Simin, V. J. Weigman, I. Mikaelian, J. Usary, Z. Hu, K. E. Rasmussen, L. P. Jones, S. Assefnia, S. Chandrasekharan, M. G. Backlund, Y. Yin, A. I. Khramtsov, R. Bastein, J. Quackenbush, R. I. Glazer, P. H. Brown, J. E. Green, L. Kopelovich, P. A. Furth, J. P. Palazzo, O. I. Olopade, P. S. Bernard, G. A. Churchill, T. Van Dyke and C. M. Peroun (2007). "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumours." Genome Biol **8**(5): R76.
HHS (2012). Help me Understand Genetics: Genetic Testing. U.S.A., CreateSpace Independent Publishing Platform.
Hinds, P. W., S. Mittnacht, V. Dulic, A. Arnold, S. I. Reed and R. A. Weinberg (1992). "Regulation of retinoblastoma protein functions by ectopic expression of human cyclins." Cell **70**(6): 993-1006.
Ho-Yen, C., R. L. Bowen and J. L. Jones (2012). "Characterization of basal-like breast cancer: an update." Diagnostic Histopathology **18**(3): 104-111.
Hoadley, K. A., C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, M. D. M. Leiserson, B. Niu, M. D. McLellan, V. Uzunangelov, J. Zhang, C. Kandoth, R. Akbani, H. Shen, L. Omberg, A. Chu, A. A. Margolin, L. J. Van't Veer, N. Lopez-Bigas, P. W. Laird, B. J. Raphael, L. Ding, A. G. Robertson, L. A. Byers, G. B. Mills, J. N. Weinstein, C. Van Waes, Z. Chen, E. A. Collisson, C. C. Benz, C. M. Perou and J. M. Stuart (2014). "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." Cell **158**(4): 929-944.
Holley, R. W., G. A. Everett, J. T. Madison and A. D. A. Zamir (1965). "NucleotideTransfer Sequences in the Yeast Alanine Ribonucleic Acid." J Biol Chem **240**(5): 2122-2128.
Holliday, R. (1979). "A new theory of carcinogenesis." Br J Cancer **40**(4): 513-522.
Howard, J. (1987). "Using Mammography for Cancer Control: An Unrealized Potential." CA CANCER J Clin **37**(1): 33-48.
Howlader, N., A. M. Noone, M. Krapcho, J. Garshell, D. Miller, S. F. Altekruse, C. l. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer and K. A. Cronin (2014). SEER cancer statistics review, 1975-2012. Bethesda, MD.
Huo, D., Y. Zheng, T. O. Ogundiran, C. Adebamowo, K. L. Nathanson, S. M. Domchek, T. R. Rebbeck, M. S. Simon, E. M. John, A. Hennis, B. Nemesure, S. Y. Wu, M. C. Leske, S. Ambs, Q. Niu, J. Zhang, N. J. Cox and O. I. Olopade (2012). "Evaluation of 19 susceptibility loci of breast cancer in women of African ancestry." Carcinogenesis **33**(4): 835-840.

Ignatov, A., T. Ignatov, A. Roessner, S. D. Costa and T. Kalinski (2010). "Role of GPR30 in the mechanisms of tamoxifen resistance in breast cancer MCF-7 cells." Breast Cancer Res Treat **123**(1): 87-96.
Imai, K., L. J. Kricka and P. Fortina (2011). "Concordance study of 3 direct-to-consumer genetic-testing services." Clin Chem **57**(3): 518-521.
Isobe, M., B. S. Emanuel, D. Givol, M. Oren and C. M. Croce (1986). "Localization of gene for human p53 tumour antigen to band 17p13." Nature **340**(1): 84-85.
Jia, P., L. Wang, H. Y. Meltzer and Z. Zhao (2010). "Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data." Schizophrenia research **122**(1-3): 38-42.
Joseph, C. G., E. Darrah, A. A. Shah, A. D. Skora, L. A. Casciola-Rosen, F. M. Wigley, F. Boin, A. Fava, C. Thoburn, I. Kinde, Y. Jiao, N. Papadopoulos, K. W. Kinzler, B. Vogelstein and A. Rosen (2014). "Association of the autoimmune disease scleroderma with an immunologic response to cancer." Science **343**(6167): 152-157.
Kallioniemi, A., O. P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman and D. Pinkel (1992). "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors." Science **258**(5083): 818-821.
Kastner, P., A. Krust, B. Turcotte, U. Stropp, L. Tora, G. H. and C. P. (1990). "Two distinct estrogen-regulated promoters generate transcripts encoding the two functionally different human progesterone receptor forms A and B." EMBO J **9**(5): 1603-1613.
Keating, B. J., S. Tischfield, S. S. Murray, T. Bhangale, T. S. Price, J. T. Glessner, L. Galver, J. C. Barrett, S. F. A. Grant, D. N. Farlow, H. R. Chandrupatla, M. Hansen, S. Ajmal, G. J. Papanicolaou, Y. Guo, M. Li, S. DerOhannessian, P. I. W. de Bakker, S. D. Bailey, A. Montpetit, A. C. Edmondson, K. Taylor, X. Gai, S. S. Wang, M. Fornage, T. Shaikh, L. Groop, M. Boehnke, A. S. Hall, A. T. Hattersley, E. Frackelton, N. Patterson, C. W. K. Chiang, C. E. Kim, R. R. Fabsitz, W. Ouwehand, A. L. Price, P. Munroe, M. Caulfield, T. Drake, E. Boerwinkle, D. Reich, A. S. Whitehead, T. P. Cappola, N. J. Samani, A. J. Lusis, E. Schadt, J. G. Wilson, W. Koenig, M. I. McCarthy, S. Kathiresan, S. B. Gabriel, H. Hakonarson, S. S. Anand, M. Reilly, J. C. Engert, D. A. Nickerson, D. J. Rader, J. N. Hirschhorn and G. A. FitzGerald (2008). "Concept, Design and Implementation of a Cardiovascular Gene-Centric 50 K SNP Array for Large-Scale Genomic Association Studies." PLOS ONE **3**(10): e3583.
Kiguli-Malwadde, E., M. A. Gonzaga, B. Francis, K. G. Michael, N. Rebecca, B. K. Rosemary and M. Zeridah (2010). "Current knowledge, attitudes and practices of women on breast cancer and
mammography at Mulago Hospital " Pan Afr Med J **5**(9): 1-13.
Knudson, A. G. (1971). "Mutation and Cancer: Statistical Study of Retinoblastoma." Proceedings of the National Academy of Sciences of the United States of America **68**(4): 820-823.
Krijgsman, O., P. Roepman, Z. W., J. S. Carroll, S. Tian, F. A. de Snoo, R. A. Bender, R. Bernards and A. M. Glas (2012). "A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response." (1573-7217 (Electronic)).
Kris A. Wetterstrand, M. S. (2016, October 30, 2019). "National Human Genome Research Institute. The cost of sequencing a human genome." Retrieved August 12, 2020, 2020, from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
Kukurba, K. R. and S. B. Montgomery (2015). "RNA Sequencing and Analysis." Cold Spring Harbor protocols **2015**(11): 951-969.
Lee, C., A. J. Iafrate and A. R. Brothman (2007). "Copy number variations and clinical cytogenetic diagnosis of constitutional disorders." Nat Genet **39**(7 Suppl): S48-54.
Levin, E. R. (2005). "Integration of the extranuclear and nuclear actions of estrogen." Mol Endocrinol **19**(8): 1951-1959.
Li, H., J. He, S. Li and A. P. Turner (2013). "Electrochemical immunosensor with N-doped graphene-modified electrode for label-free detection of the breast cancer biomarker CA 15-3." Biosens Bioelectron **43**: 25-29.

Li, J., K. Humphreys, H. Darabi, G. Rosin, U. Hannelius, T. Heikkinen, K. Aittomaki, C. Blomqvist, P. D. Pharoah, A. M. Dunning, S. Ahmed, M. J. Hooning, A. Hollestelle, R. A. Oldenburg, L. Alfredsson, A. Palotie, L. Peltonen-Palotie, A. Irwanto, H. Q. Low, G. H. Teoh, A. Thalamuthu, J. Kere, M. D'Amato, D. F. Easton, H. Nevanlinna, J. Liu, K. Czene and P. Hall (2010). "A genome-wide association scan on estrogen receptor-negative breast cancer." Breast Cancer Res **12**(6): R93.
Li, W., A. Lee and P. K. Gregersen (2009). "Copy-number-variation and copy-number-alteration region detection by cumulative plots." BMC Bioinformatics **10**(Suppl 1): S67-S67.
Liang, D., Y. Peng, W. Lv, L. Deng, Y. Zhang, H. Li, P. Yang, J. Zhang, Z. Song, G. Xu, D. S. Cram and L. Wu (2014). "Copy number variation sequencing for comprehensive diagnosis of chromosome disease syndromes." J Mol Diagn **16**(5): 519-526.
Liang, L., N. Morar, A. L. Dixon, G. M. Lathrop, G. R. Abecasis, M. F. Moffatt and W. O. Cookson (2013). "A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines." Genome Res **23**(4): 716-726.
Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub (2005). "MicroRNA expression profiles classify human cancers." Nature **435**(7043): 834-838.
Lupski, J. R., R. M. de Oca-Luna, S. Slaugenhaupt, L. Pentao, V. Guzzetta, B. J. Trask, O. Saucedo-Cardenas, D. F. Barker, J. M. Killian, C. A. Garcia, A. Chakravarti and P. I. Patel (1991). "DNA duplication associated with Charcot-Marie-Tooth disease type 1A." Cell **66**(2): 219-232.
Marshall, C. R., A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari and S. W. Scherer (2008). "Structural variation of chromosomes in autism spectrum disorder." Am J Hum Genet **82**(2): 477-488.
Mavaddat, N., T. R. Rebbeck, S. R. Lakhani, D. F. Easton and A. C. Antoniou (2010). "Incorporating tumour pathology information into breast cancer risk prediction algorithms." Breast Cancer Res **12**(R28): 1-12.
McCann, J. (2000). "Gender differences in cancer that don't make sense--or do they?" J Natl Cancer Inst **92**(19): 1560-1562.
McCann, J. and B. N. Ames (1976). "Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion." Proc Natl Acad Sci U S A **73**(3): 950-954.
Mehta, J., J. M. Kling and J. E. Manson (2021). "Risks, Benefits, and Treatment Modalities of Menopausal Hormone Therapy: Current Concepts." (1664-2392 (Print)).
Michael, K. L., L. C. Taylor, S. L. Schultz and D. R. Walt (1998). "Randomly ordered addressable high-density optical sensor arrays." Anal Chem **70**(7): 1242-1248.
Michailidou, K., P. Hall, A. Gonzalez-Neira, M. Ghoussaini, J. Dennis, R. L. Milne, M. K. Schmidt, J. Chang-Claude, S. E. Bojesen, M. K. Bolla, Q. Wang, E. Dicks, A. Lee, C. Turnbull, N. Rahman, Breast, C. Ovarian Cancer Susceptibility, O. Fletcher, J. Peto, L. Gibson, I. Dos Santos Silva, H. Nevanlinna, T. A. Muranen, K. Aittomaki, C. Blomqvist, K. Czene, A. Irwanto, J. Liu, Q. Waisfisz, H. Meijers-Heijboer, M. Adank, B. Hereditary, N. Ovarian Cancer Research Group, R. B. van der Luijt, R. Hein, N. Dahmen, L. Beckman, A. Meindl, R. K. Schmutzler, B. Muller-Myhsok, P. Lichtner, J. L. Hopper, M. C. Southey, E. Makalic, D. F. Schmidt, A. G. Uitterlinden, A. Hofman, D. J. Hunter, S. J. Chanock, D. Vincent, F. Bacot, D. C. Tessier, S. Canisius, L. F. Wessels, C. A. Haiman, M. Shah, R. Luben, J. Brown, C. Luccarini, N. Schoof, K. Humphreys, J. Li, B. G. Nordestgaard, S. F. Nielsen, H. Flyger, F. J. Couch, X. Wang, C. Vachon, K. N. Stevens, D. Lambrechts, M. Moisse, R. Paridaens, M. R. Christiaens, A. Rudolph, S. Nickels, D. Flesch-Janys, N. Johnson, Z. Aitken, K. Aaltonen, T. Heikkinen, A. Broeks, L. J. Veer, C. E. van der Schoot, P. Guenel, T. Truong, P. Laurent-Puig, F. Menegaux, F. Marme, A. Schneeweiss, C. Sohn, B. Burwinkel, M. P. Zamora, J. I. Perez, G. Pita, M. R. Alonso, A. Cox, I. W. Brock, S. S. Cross, M. W. Reed, E. J. Sawyer, I. Tomlinson, M. J. Kerin, N. Miller, B. E. Henderson, F. Schumacher, L. Le

Marchand, I. L. Andrulis, J. A. Knight, G. Glendon, A. M. Mulligan, I. kConFab, G. Australian Ovarian Cancer Study, A. Lindblom, S. Margolin, M. J. Hooning, A. Hollestelle, A. M. van den Ouweland, A. Jager, Q. M. Bui, J. Stone, G. S. Dite, C. Apicella, H. Tsimiklis, G. G. Giles, G. Severi, L. Baglietto, P. A. Fasching, L. Haeberle, A. B. Ekici, M. W. Beckmann, H. Brenner, H. Muller, V. Arndt, C. Stegmaier, A. Swerdlow, A. Ashworth, N. Orr, M. Jones, J. Figueroa, J. Lissowska, L. Brinton, M. S. Goldberg, F. Labreche, M. Dumont, R. Winqvist, K. Pylkas, A. Jukkola-Vuorinen, M. Grip, H. Brauch, U. Hamann, T. Bruning, G. Network, P. Radice, P. Peterlongo, S. Manoukian, B. Bonanni, P. Devilee, R. A. Tollenaar, C. Seynaeve, C. J. van Asperen, A. Jakubowska, J. Lubinski, K. Jaworska, K. Durda, A. Mannermaa, V. Kataja, V. M. Kosma, J. M. Hartikainen, N. V. Bogdanova, N. N. Antonenkova, T. Dork, V. N. Kristensen, H. Anton-Culver, S. Slager, A. E. Toland, S. Edge, F. Fostira, D. Kang, K. Y. Yoo, D. Y. Noh, K. Matsuo, H. Ito, H. Iwata, A. Sueta, A. H. Wu, C. C. Tseng, D. Van Den Berg, D. O. Stram, X. O. Shu, W. Lu, Y. T. Gao, H. Cai, S. H. Teo, C. H. Yip, S. Y. Phuah, B. K. Cornes, M. Hartman, H. Miao, W. Y. Lim, J. H. Sng, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, C. Y. Shen, C. N. Hsiung, P. E. Wu, S. L. Ding, S. Sangrajrang, V. Gaborieau, P. Brennan, J. McKay, W. J. Blot, L. B. Signorello, Q. Cai, W. Zheng, S. Deming-Halverson, M. Shrubsole, J. Long, J. Simard, M. Garcia-Closas, P. D. Pharoah, G. Chenevix-Trench, A. M. Dunning, J. Benitez and D. F. Easton (2013). "Large-scale genotyping identifies 41 new loci associated with breast cancer risk." Nat Genet **45**(4): 353-361.

Miller, B. A., L. N. Kolonel, L. Bernstein, J. L. Young, G. M. Swanson, D. West, C. R. Key, J. M. Liff, C. S. Glover and G. A. Alexander (1996). Racial/ethnic patterns of cancer in the United States, 1988-1992.

Mitri, Z., T. Constantine and R. O'Regan (2012). "The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy." Chemother Res Pract **2012**(1): 743193.

Mraz, M., K. Malinova, J. Kotaskova, S. Pavlova, B. Tichy, J. Malcikova, K. Stano Kozubik, J. Smardova, Y. Brychtova, M. Doubek, M. Trbusek, J. Mayer and S. Pospisilova (2009). "miR-34a, miR-29c and miR-17-5p are downregulated in CLL patients with TP53 abnormalities." Leukemia **23**(6): 1159-1163.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein and M. Snyder (2008). "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing." Science (New York, N.Y.) **320**(5881): 1344-1349.

Nawy, T. (2014). "Single-cell sequencing." Nat Methods **11**(1): 18.

NCI. (2021, 09/02/2015). "What is cancer?", from http://www.cancer.gov/about-cancer/understanding/what-is-cancer.

Ng, P. C. and E. F. Kirkness (2010). "Whole genome sequencing." Methods Mol Biol **628**: 215-226.

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. Jabs, D. A. Nickerson, J. Shendure and M. J. Bamshad (2010). "Exome sequencing identifies the cause of a mendelian disorder." Nature genetics **42**(1): 30-35.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson and J. Shendure (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

O'Brien, K. M., S. R. Cole, C. Poole, J. T. Bensen, A. H. Herring, L. S. Engel and R. C. Millikan (2014). "Replication of breast cancer susceptibility loci in whites and African Americans using a Bayesian approach." Am J Epidemiol **179**(3): 382-394.

O'Brien, K. M., S. R. Cole, C. K. Tse, C. M. Perou, L. A. Carey, W. D. Foulkes, L. G. Dressler, J. Geradts and R. C. Millikan (2010). "Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study." Clin Cancer Res **16**(24): 6100-6110.

Ormerod, M. G. (2008). DNA Analysis, Ploidy. Flow Cytometry - A Basic Introduction.

Park, H., J. Kim, Y. Ju, O. Gokcumen, R. E. Mills, S. Kim, S. Lee, D. Suh, D. Hong, H. Kang, Y. Yoo, J. Shin, H. Kim, M. Yavartanoo, Y. Chang, J. Ha, W. Chong, G. Hwang, K. Darvishi, H. Kim, S. Yang, K. Yang, H. Kim, M. E. Hurles, S. W. Scherer, N. P. Carter, C. Tyler-Smith, C. Lee and J. Seo (2010). "Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing." Nature genetics **42**(5): 400-405.

Parkin, D. M., F. Bray, J. Ferlay and P. Pisani (2005). "Global Cancer Statistics, 2002." CA A Cancer Journal for Clinicians **55**(1): 74-110.
Pease, A. C., D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes and S. P. Fodor (1994). "Light-generated oligonucleotide arrays for rapid DNA sequence analysis." Proc Natl Acad Sci U S A **91**(11): 5022-5026.
Perou, C. M., T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen and L. A. Akslen (2000). "Molecular portraits of human breast tumours." Nature **406**(6797): 747-752.
Prat, A., J. S. Parker, O. Karginova, C. Fan, C. Livasy, J. I. Herschkowitz, X. He and C. M. Perou (2010). "Phenotypic and molecular characterization of the
claudin-low intrinsic subtype of breast cancer." Breast Cancer Res **12**(5): R68-R85.
Rebbeck, T. R. (2020). "Cancer in sub-Saharan Africa." (1095-9203 (Electronic)).
Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-454.
Reeves, M. D., T. M. Yawitch, N. C. van der Merwe, H. J. van den Berg, G. Dreyer and E. J. van Rensburg (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**(5): 677-682.
Rehm, H. L. (2013). "Disease-targeted sequencing: a cornerstone in the clinic." Nature reviews. Genetics **14**(4): 295-300.
Reuning, U., S. Sperl, C. Kopitz, H. Kessler, A. Kruger, M. Schmitt and V. Magdolen (2003). "Urokinase-type Plasminogen Activator (uPA) and its Receptor (uPAR): Development of Antagonists of uPA / uPAR Interaction and their Effects In Vitro and In Vivo." Curr Pharm Des **9**(19): 1529-1543.
Ronnov-Jessen, L., O. W. Petersen and M. J. Bissell (1996). "Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction." Physiological reviews **76**(1): 69-125.
Ruiz-Narvaez, E. A., L. Rosenberg, S. Yao, C. N. Rotimi, A. L. Cupples, E. V. Bandera, C. B. Ambrosone, L. L. Adams-Campbell and J. R. Palmer (2013). "Fine-mapping of the 6q25 locus identifies a novel SNP associated with breast cancer risk in African-American women." Carcinogenesis **34**(2): 287-291.
Russo, G., C. Zegar and A. Giordano (2003). "Advantages and limitations of microarray technology in human cancer." Oncogene **22**(42): 6497-6507.
Ryland, G. L., M. A. Doyle, D. Goode, S. E. Boyle, D. Y. Choong, S. M. Rowley, J. Li, D. D. Bowtell, R. W. Tothill, I. G. Campbell and K. L. Gorringe (2015). "Loss of heterozygosity: what is it good for?" BMC Med Genomics **8**: 45.
Ryland, G. L., M. A. Doyle, D. Goode, S. E. Boyle, D. Y. H. Choong, S. M. Rowley, J. Li, D. L. Bowtell, R. W. Tothill, I. G. Campbell and K. L. Gorringe (2015). "Loss of heterozygosity: what is it good for?" BMC Medical Genomics **8**(1): 45.
Saeed, I. E., H. Y. Weng, K. H. Mohamed and S. I. Mohammed (2014). "Cancer incidence in Khartoum, Sudan: first results from the Cancer Registry, 2009-2010." Cancer Med **3**(4): 1075-1084.
Sager, R. (1979). "Transposable elements and chromosomal rearrangements in cancer--a possible link." Nature **282**(5738): 447-449.
Saibu, M., A. James, O. Adu, F. Faduyile, O. F. Adewale, O. Iyapo, S. Soyemi and A. Benjamin (2017). "Epidemiology and Incidence of Common Cancers in Nigeria." Journal of Cancer Biology & Research **5**: 1105.
Sankaranarayanan, R. (2006). Strategies for implementation of screening programs in low- and medium resource settings. UICC World Cancer Congress 2006, Washington, DC, USA.

Saudi Mendeliome, G. (2015). "Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases." Genome Biology **16**(1): 134.
Schneeberger, K. (2014). "Using next-generation sequencing to isolate mutant genes from forward genetic screens." Nat Rev Genet **15**(10): 662-676.
Schon, K. and M. Tischkowitz (2018). "Clinical implications of germline mutations in breast cancer: TP53." Breast Cancer Research and Treatment **167**(2): 417-423.
Schwartz, A. M., D. E. Henson, D. Chen and S. Rajamarthandan (2014). "Histologic Grade Remains a Prognostic Factor for Breast Cancer Regardless of the Number of Positive Lymph Nodes and Tumor Size: A Study of 161 708 Cases of Breast Cancer From the SEER Program." Archives of Pathology & Laboratory Medicine **138**(8): 1048-1052.
Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye and M. Wigler (2007). "Strong association of de novo copy number mutations with autism." Science **316**(5823): 445-449.
Service, R. F. (2006). "Gene sequencing. The race for the $1000 genome." Science **311**(5767): 1544-1546.
Shlien, A. and D. Malkin (2009). "Copy number variations and cancer." Genome Med **1**(6): 62.
Sonabend, A. M., M. Bansal, P. Guarnieri, L. Lei, B. Amendolara, C. Soderquist, R. Leung, J. Yun, B. Kennedy, J. Sisti, S. Bruce, R. Bruce, R. Shakya, T. Ludwig, S. Rosenfeld, P. A. Sims, J. N. Bruce, A. Califano and P. Canoll (2014). "The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression." Cancer Res **74**(5): 1440-1451.
Sorlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning and A. L. Borresen-Dale (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proc Natl Acad Sci U S A **98**(19): 10869-10874.
Srour, N., M. A. Reymond and R. Steinert (2008). "Lost in translation? A systematic database of gene expression in breast cancer." Pathobiology **75**(2): 112-118.
Stabel, S. and P. J. Parker (1991). "Protein kinase C." Pharmac Ther **51**(1): 71-95.
Steemers, F. J., J. A. Ferguson and D. R. Walt (2000). "Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays." Nat Biotechnol **18**(1): 91-94.
Stephens, P. J., C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal and P. J. Campbell (2011). "Massive genomic rearrangement acquired in a single catastrophic event during cancer development." Cell **144**(1): 27-40.
Steyn, K., J. Fourie and N. Temple (2006). Chronic Diseases of Lifestyle in South Africa: 1995 - 2005
. Technical Report. C
ape Town. M. Krisela Steyn MSc, NED, M. Jean Fourie BA (Nursing) and N. T. PhD.
Stoffel, E. M., E. Koeppe, J. Everett, P. Ulintz, M. Kiel, J. Osborne, L. Williams, K. Hanson, S. B. Gruber and L. S. Rozek (2018). "Germline Genetic Features of Young Individuals With Colorectal Cancer." Gastroenterology **154**(4): 897-905.e891.
Strausberg, R. L. and A. J. Simpson (2010). "Whole-genome cancer analysis as an approach to deeper understanding of tumour biology." Br J Cancer **102**(2): 243-248.
Taub, F. E., J. M. DeLeo and E. B. Thompson (1983). "Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs." Dna **2**(4): 309-327.
Thompson, E. R., S. M. Rowley, N. Li, S. McInerny, L. Devereux, M. W. Wong-Brown, A. H. Trainer, G. Mitchell, R. J. Scott, P. A. James and I. G. Campbell (2016). "Panel Testing for Familial Breast Cancer: Calibrating the Tension Between Research and Clinical Care." J Clin Oncol **34**(13): 1455-1459.

Tung, N., C. Battelli, B. Allen, R. Kaldate, S. Bhatnagar, K. Bowles, K. Timms, J. E. Garber, C. Herold, L. Ellisen, J. Krejdovsky, K. DeLeonardis, K. Sedgwick, K. Soltis, B. Roa, R. J. Wenstrup and A. R. Hartman (2015). "Frequency of mutations in individuals with breast cancer referred for BRCA1 and BRCA2 testing using next-generation sequencing with a 25-gene panel." Cancer **121**(1): 25-33.
Tung, N., N. U. Lin, J. Kidd, B. A. Allen, N. Singh, R. J. Wenstrup, A. R. Hartman, E. P. Winer and J. E. Garber (2016). "Frequency of Germline Mutations in 25 Cancer Susceptibility Genes in a Sequential Series of Patients With Breast Cancer." J Clin Oncol **34**(13): 1460-1468.
Turnbull, C., S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, S. Seal, M. Ghoussaini, S. Hines, C. S. Healey, D. Hughes, M. Warren-Perry, W. Tapper, D. Eccles, D. G. Evans, C. Breast Cancer Susceptibility, M. Hooning, M. Schutte, A. van den Ouweland, R. Houlston, G. Ross, C. Langford, P. D. Pharoah, M. R. Stratton, A. M. Dunning, N. Rahman and D. F. Easton (2010). "Genome-wide association study identifies five new breast cancer susceptibility loci." Nat Genet **42**(6): 504-507.
Tuzun, E., A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M. V. Olson and E. E. Eichler (2005). "Fine-scale structural variation of the human genome." Nat Genet **37**(7): 727-732.
Vallejos, C. S., H. L. Gomez, W. R. Cruz, J. A. Pinto, R. R. Dyer, R. Velarde, J. F. Suazo, S. P. Neciosup, M. Leon, M. A. de la Cruz and C. E. Vigil (2010). "Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database." Clin Breast Cancer **10**(4): 294-300.
van 't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.
van Rensburg, E. J., N. C. van der Merwe, M. D. Sluiter and C. M. Schlebusch (2007). Impact of the BRCA-genes on the burden of familial breast/ovarian cancer in South Africa. American Society of Human Genetics. San Diego, California.
Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression." Science **270**(5235): 484.
Vlachopoulou, E. (2011). "Genome-Wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls."
Volinia, S., G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris and C. M. Croce (2006). "A microRNA expression signature of human solid tumors defines cancer gene targets." Proc Natl Acad Sci U S A **103**(7): 2257-2261.
Vora, G. J., C. E. Meador, D. A. Stenger and J. D. Andreadis (2004). "Nucleic Acid Amplification Strategies for DNA Microarray-Based Pathogen Detection." Applied and Environmental Microbiology **70**(5): 3047-3054.
Walsh, T., M. K. Lee, S. Casadei, A. M. Thornton, S. M. Stray, C. Pennil, A. S. Nord, J. B. Mandell, E. M. Swisher and M. King (2010). "Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing." Proceedings of the National Academy of Sciences of the United States of America **107**(28): 12629-12633.
Walsh, T., J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M. C. King and J. Sebat (2008). "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia." Science **320**(5875): 539-543.
Walt, D. R. (2000). "Techview: molecular biology. Bead-based fiber-optic arrays." Science **287**(5452): 451-452.

Wang, Y., J. Waters, M. L. Leung, A. Unruh, W. Roh, X. Shi, K. Chen, P. Scheet, S. Vattathil, H. Liang, A. Multani, H. Zhang, R. Zhao, F. Michor, F. Meric-Bernstam and N. E. Navin (2014). "Clonal evolution in breast cancer revealed by single nucleus genome sequencing." Nature **512**(7513): 155-160.
Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.
Whitworth, P., L. Stork-Sloots, F. A. de Snoo, P. Richards, M. Rotkis, J. Beatty, A. Mislowsky, J. V. Pellicane, B. Nguyen, L. Lee, C. Nash, M. Gittleman, S. Akbari and P. D. Beitsch (2014). "Chemosensitivity predicted by BluePrint 80-gene functional subtype and MammaPrint in the Prospective Neoadjuvant Breast Registry Symphony Trial (NBRST)." (1534-4681 (Electronic)).
Wilson, G. M., S. Flibotte, V. Chopra, B. L. Melnyk, W. G. Honer and R. A. Holt (2006). "DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling." Hum Mol Genet **15**(5): 743-749.
Wingate, H., A. Puskas, M. Duong, T. Bui, D. Richardson, Y. Liu, S. L. Tucker, C. van Pelt, L. Meijer, H. K. and K. K. (2009). "Low molecular weight cyclin E is specific in breast cancer and is associated with mechanisms of tumor progression." Cell Cycle **8**(7): 1062-1068.
Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard and P. M. Visscher (2015). "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index." Nat Genet **47**(10): 1114-1120.
Yoon, S., Z. Xuan, V. Makarov, K. Ye and J. Sebat (2009). "Sensitive and accurate detection of copy number variants using read depth of coverage." Genome Research **19**(9): 1586-1592.
Zarrei, M., J. R. MacDonald, D. Merico and S. W. Scherer (2015). "A copy number variation map of the human genome." Nat Rev Genet **16**(3): 172-183.
Zhang, X. L., D. Zhang, F. J. Michel, J. L. Blum, F. A. Simmen and R. C. Simmen (2003). "Selective interactions of Kruppel-like factor 9/basic transcription element-binding protein with progesterone receptor isoforms A and B determine transcriptional activity of progesterone-responsive genes in endometrial epithelial cells." J Biol Chem **278**(24): 21474-21482.
Zhao, S., W. P. Fung-Leung, A. Bittner, K. Ngo and X. Liu (2014). "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells." PLoS One **9**(1): e78644.
Zheng, W., J. Long, Y. T. Gao, C. Li, Y. Zheng, Y. B. Xiang, W. Wen, S. Levy, S. L. Deming, J. L. Haines, K. Gu, A. M. Fair, Q. Cai, W. Lu and X. O. Shu (2009). "Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1." Nat Genet **41**(3): 324-328.
Zheng, W., B. Zhang, Q. Cai, H. Sung, K. Michailidou, J. Shi, J. Y. Choi, J. Long, J. Dennis, M. K. Humphreys, Q. Wang, W. Lu, Y. T. Gao, C. Li, H. Cai, S. K. Park, K. Y. Yoo, D. Y. Noh, W. Han, A. M. Dunning, J. Benitez, D. Vincent, F. Bacot, D. Tessier, S. W. Kim, M. H. Lee, J. W. Lee, J. Y. Lee, Y. B. Xiang, Y. Zheng, W. Wang, B. T. Ji, K. Matsuo, H. Ito, H. Iwata, H. Tanaka, A. H. Wu, C. C. Tseng, D. Van Den Berg, D. O. Stram, S. H. Teo, C. H. Yip, I. N. Kang, T. Y. Wong, C. Y. Shen, J. C. Yu, C. S. Huang, M. F. Hou, M. Hartman, H. Miao, S. C. Lee, T. C. Putti, K. Muir, A. Lophatananon, S. Stewart-Brown, P. Siriwanarangsan, S. Sangrajrang, H. Shen, K. Chen, P. E. Wu, Z. Ren, C. A. Haiman, A. Sueta, M. K. Kim, U. S. Khoo, M. Iwasaki, P. D. Pharoah, W. Wen, P. Hall, X. O. Shu, D. F. Easton and D. Kang (2013). "Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls." Hum Mol Genet **22**(12): 2539-2550.

# Chapter 2

# Materials and Methods

## 2.1 SAMPLING, PATIENT INFORMATION AND CONSENT

Peripheral blood samples were previously collected from South African women with breast cancer, who attended the Oncology Clinic at Steve Biko Hospital, Pretoria, between 1993 and 2001. The study population were of self-reported African ancestry, at least 18 years old and were included regardless of age at diagnosis or family history. In total we received blood samples from 286 patients with age at diagnosis ranging from 21 to 85 years (mean 49.52 years ± 12.93 years). DNA was extracted from the blood samples using the method described by Johns and Paulus-Thomas (Johns and Paulus-Thomas 1989). For the current study we selected 165 of these patients (Table 2.1) beginning with the youngest patients. With the exception of four cases (BRB130, BRB290, BRC134 and BRC210) all the samples were previously screened for BRCA1/2 deleterious variants using SSCP/Heteroduplex analyses and multiplex ligation-dependent probe amplification (MLPA) and were thought to be negative for pathogenic or likely pathogenic variants. These samples were collected with the patients consent and stored as frozen blood samples. Most of the samples were tested using conventional methods [ SSCP/Heteroduplex analyses and MLPA] for the presence of BRCA variants.

Table 2.1: Patient information.

| BRB lab # | Age @ Dx (yrs,months) | Ethnicity/ Language | Family History | Histology* | Grade |
|---|---|---|---|---|---|
| **2** | 35, 6 months | Sepedi | None | inf duct | IV |
| **3** | 48, 10 months | Tswana | None | inf duct | IV |
| **5** | 49, 10 months | N. Sotho | Mother breast ca | inf duct | IV |
| **6** | 43, 8 months | N. Sotho | None | inf duct | III a |
| **8** | 28, 0 months | N. Sotho | None | inf duct | IV |
| **9** | 36, 5 months | N. Sotho | None | inf duct | IV |
| **10** | 42, 8 months | N. Sotho | None | medullary ductal | III |
| **11** | 44, 0 months | Tswana | None | inf duct | III |
| **14** | 47, 8 months | Zulu | None | Unknown | IV |
| **17** | 44, 11 months | Ndebele | Cousin breast ca | inf duct | II b |
| **18** | 31, 9 months | Tswana | None | inf duct | III |

| | | | | | |
|---|---|---|---|---|---|
| **19** | 36, 0 months | Swazi | None | inf duct | IV |
| **20** | 34, 1 months | N. Sotho | None | inf duct | IV |
| **21** | 31, 0 months | Unknown | None | medullary ductal | III b |
| **28** | 29, 4 months | Zulu | None | inf duct | III a |
| **34** | 43, 1 month | Tswana | None | inf duct | II |
| **37** | 37, 5 months | Ndebele | Mother breast ca | inf duct | IV |
| **38** | 46, 10 months | Tswana | None | Unknown | II a |
| **39** | 54, 0 months | Zulu | None | inf duct | IV |
| **42** | 49, 11 months | Tswana | None | Unknown | II a |
| **44** | 30, 0 months | N. Sotho | None | Unknown | ? |
| **46** | 46, 6 months | Tswana | None | inf duct | IV |
| **47** | 52, 8 months | Zulu | None | nos | IV |
| **48** | 42, 3 months | Tswana | None | inf duct | II b |
| **49** | 43, 4 months | Zulu | Aunt breast ca | inf duct | II b |
| **50** | 40, 4 months | Tswana | Mother breast ca | inf duct | IV |
| **51** | 52, 6 months | Unknown | None | inf duct | IV |
| **52** | 46, 11 months | Unknown | None | inf duct | II b |
| **53** | 43, 5 months | Zulu | None | inf duct | IV |
| **55** | 32, 9 months | N. Sotho | None | Unknown | II a |
| **57** | 26, 1 month | S. Sotho | None | inf duct | III b |
| **58** | 33, 7 months | N. Sotho | None | inf duct | II a |
| **59** | 39, 0 months | Swazi | None | inf duct | IV |
| **62** | 43, 7 months | N. Sotho | None | inf duct, with tubular differentiation | IV |
| **68** | 43, 0 months | Ndebele | Mother breast ca | inf duct | III b |
| **70** | 42, 7 months | Tswana | Maternal aunt breast ca | inf duct | IV |
| **72** | 53, 8 months | Unknown | None | inf duct | IV |
| **73** | 29, 11 months | Tswana | None | inf duct | II b |
| **74** | 47, 4 months | N. Sotho | None | inf duct | III b |
| **75** | 46, 11 months | Zulu | None | inf duct | II b |
| **77** | 47, 7 months | Zulu | None | inf duct | IV |
| **78** | 36, 4 months | N. Sotho | None | inf duct | III b |
| **81** | 52, 6 months | Sotho | None | invasive lobular | IV |
| **83** | 53, 0 months | Sepedi | None | Unknown | III b |
| **84** | 43, 7 months | Tswana | None | Unknown | II a |
| **87** | 50, 0 months | Swazi | None | inf duct | III |
| **88** | 39, 3 months | N. Sotho | None | inf duct | IV |
| **89** | 48, 8 months | N. Sotho | Uncle prostate ca | inf duct | II a |
| **91** | 48, 0 months | Unknown | None | inf duct | III a |
| **94** | 35, 3 months | Zulu | None | inf duct | III b |

| | | | | | |
|---|---|---|---|---|---|
| **96** | 30, 5 months | N. Sotho | None | inf duct | IV |
| **98** | 43, 3 months | Tswana | None | inf duct | II b |
| **99** | 45, 3 months | Zulu | None | inf duct | II b |
| **101** | 50, 8 months | Tswana | None | inf duct | IV |
| **102** | 40, 9 months | Zulu | None | inf duct | IV |
| **104** | 47, 0 months | Zulu | None | inf duct with areas of DCIS | II b |
| **106** | 35, 0 months | Unknown | None | inf duct | IV |
| **108** | 39, 6 months | Zulu | None | inf duct | III b |
| **111** | 42, 7 months | Tswana | None | inf duct + DCIS | IV |
| **113** | 41, 7 months | Tswana | None | inf duct | III b |
| **114** | 47, 1 month | Zulu | None | inf duct | II |
| **118** | 33, 10 months | N. Sotho | Father oesophagus ca | inf duct | IV |
| **120** | 52, 6 months | Zulu | None | inf duct | IV |
| **121** | 54, 0 months | Zulu | None | Unknown | ? |
| **122** | 43, 10 months | Tswana | None | inf duct | IV |
| **123** | 44, 4 months | Unknown | None | inf duct | II |
| **124** | 25, 9 months | Tswana | None | inf duct | IV |
| **125** | 38, 2 months | Tswana | None | inf duct | IV |
| **129** | 43, 2 months | Shangaan | None | inf duct | IV |
| **130** | 45, 8 months | Tswana | None | inf duct | III b |
| **131** | 45, 3 months | Zulu | None | inf duct | IV |
| **132** | 42, 4 months | Unknown | None | inf duct | IV |
| **137** | 48, 0 months | Sotho | None | inf duct | III b |
| **138** | 48, 6 months | Tswana | None | inf duct | III a |
| **139** | 48, 3 months | Tswana | None | inf lobular | IV |
| **142** | 52, 11 months | Unknown | None | inf duct | III a |
| **143** | 42, 0 months | Ndebele | None | inf duct | IV |
| **146** | 52, 2 months | Tswana | None | inf duct | IV |
| **147** | 52, 10 months | Zulu | Mother breast ca | Unknown | IV |
| **148** | 39, 8 months | Ndebele | None | papillary with DCIS | II a |
| **150** | 45, 6 months | Sotho | None | inf duct | II b |
| **152** | 39, 4 months | Sepedi | None | inf duct | II b |
| **153** | 22, 10 months | Tswana | None | inf duct | IV |
| **154** | 51, 2 months | Tswana | None | inf duct | II |
| **156** | 47, 0 months | Unknown | None | inf duct | II b |
| **158** | 53, 7 months | Ndebele | None | inf duct | IV |
| **160** | 40, 3 months | Xhosa | None | inf duct | II a |
| **161** | 29, 6 months | Sotho | None | Unknown | IV |
| **162** | 42, 10 months | Sepedi | Sister breast ca | Unknown | III |
| **166** | 28, 10 months | N. Sotho | Sister & aunt breast ca | inf duct | III b |
| **167** | 44, 6 months | Unknown | None | inf duct | IV |

| | | | | | |
|---|---|---|---|---|---|
| **169** | 38, 1 month | N. Sotho | None | inf duct | IV |
| **170** | 40, 2 months | Sepedi | None | inf duct | II |
| **171** | 37, 4 months | Tsonga | None | inf duct | III |
| **172** | 31, 3 months | N. Sotho | None | inf duct | III a |
| **173** | 34, 9 months | Tswana | None | inf duct | III a |
| **174** | 37, 8 months | N. Sotho | None | inf duct | III |
| **175** | 37, 9 months | Swazi | None | inf mucinous | IV |
| **177** | 26, 5 months | N. Sotho | None | inf duct | II b |
| **182** | 41, 2 months | Tswana | None | inf duct | III |
| **185** | 30, 4 months | Unknown | None | inf duct | IV |
| **186** | 35, 9 months | Sotho | None | inf duct | II b |
| **187** | 41, 2 months | S. Sotho | None | inf duct | IV R |
| **188** | 43, 10 months | Zulu | None | inf duct | II a |
| **189** | 40, 3 months | Tsonga | None | inf duct | III |
| **190** | 38, 9 months | Unknown | None | inf duct | IV |
| **191** | 28, 8 months | Tswana | None | inf duct | II b |
| **193** | 43, 5 months | Ndebele | None | inf duct | II b |
| **194** | 44, 1 month | Sotho | None | inf duct | II b |
| **197** | 51, 0 months | Zulu | None | inf duct | IV |
| **199** | 45, 2 months | Zulu | None | inf duct | III b |
| **200** | 41, 6 months | Sepedi | None | inf duct | III a |
| **201** | 42, 4 months | Ndebele | Mother's aunt breast ca | inf duct | III b |
| **203** | 39, 7 months | Zulu | None | inf duct | IV |
| **205** | 43, 6 months | Tswana | None | medullary ductal | III b |
| **207** | 49, 9 months | Tswana | None | inf duct | IV |
| **208** | 44, 8 months | Zulu | None | inf duct | III a |
| **215** | 40, 5 months | Ndebele | None | inf duct | IV |
| **220** | 38, 11 months | Sotho | None | inf duct | II |
| **224** | 26, 7 months | Swazi | None | inf duct | II a |
| **225** | 34, 4 months | Zulu | None | inf duct | IV |
| **226** | 26, 2 months | Zulu | None | inf duct | III b |
| **229** | 38, 11 months | Zulu | None | inf duct | III |
| **233** | 41, 1 month | Ndebele | None | inf duct | III |
| **234** | 39, 10 months | Tswana | None | inf duct | III a |
| **236** | 47, 8 months | Swazi | None | inf duct | III |
| **237** | 28, 4 months | Zulu | None | inf duct | IV |
| **238** | 44, 10 months | Zulu | None | inf duct | III a |
| **239** | 51, 11 months | Zulu | None | inf duct | III |
| **240** | 48, 10 months | Sotho | None | inf duct | IV |
| **241** | 40, 1 month | Xhosa | None | Unknown | IV |
| **242** | 46, 10 months | Zulu | None | inf duct | II b |
| **245** | 35, 10 months | N. Sotho | Two sisters breast ca | medullary ductal | ? |
| **246** | 42, 1 month | Tswana | Mother breast ca | inf duct | I |

| | | | | | |
|---|---|---|---|---|---|
| **249** | 32, 4 months | N. Sotho | None | inf duct | II b |
| **252** | 40, 9 months | Sepedi | None | inf duct | III b |
| **253** | 46, 4 months | N. Sotho | Father stomach ca | inf duct | IV |
| **254** | 49, 0 months | Ndebele | None | inf duct | IV |
| **255** | 51, 4 months | Sotho | None | inf duct | IV |
| **257** | 46, 9 months | N. Sotho | None | inf duct | IV |
| **258** | 28, 1 month | Ndebele | None | inf duct | II a |
| **259** | 26, 11 months | Sepedi | None | inf duct | IV |
| **260** | 39, 0 months | Tswana | None | Unknown | IV |
| **261** | 38, 1 month | Xhosa | None | Unknown | II a |
| **264** | 42, 3 months | Shangaan | None | inf duct | IV |
| **265** | 43, 7 months | Zulu | None | Unknown | IV |
| **267** | 45, 11 months | Tswana - Venda | None | Unknown | II b |
| **268** | 46, 2 months | N. Sotho | None | inf duct | IV |
| **270** | 41, 2 months | Tswana | None | inf duct | IV |
| **271** | 41, 3 months | Zulu | None | inf duct | III |
| **272** | 40, 2 months | Zulu | None | inf duct | II b |
| **273** | 48, 9 months | N. Sotho | Daughter breast ca | inf duct | ? |
| **275** | 47, 4 months | S. Sotho | None | inf duct | IV |
| **276** | 46, 6 months | N. Sotho | None | inf duct | ? |
| **279** | 42, 4 months | Zulu | None | Unknown | ? |
| **281** | 51, 0 months | Zulu | Sister Ovarian ca | inf duct | IV |
| **282** | 49, 11 months | Swazi | None | inf duct | III b |
| **283** | 47, 0 months | Tswana | Mother breast ca | inf duct | IV |
| **284** | 30, 10 months | Zulu | None | inf duct | II b |
| **286** | 48, 11 months | Unknown | None | inf duct | III b |
| **287** | 50, 7 months | Ndebele | None | inf duct | III a |
| **288** | 52, 11 months | Swazi | None | inf duct | IV |
| **290** | 26, 4 months | N. Sotho | None | Unknown | III b |
| | | | | | |
| **BRC134** | 38, 0 months | Sotho | Mother & sister breast ca | Unknown | ? |
| **BRC210** | 36, 0 months | Sotho | Mother & 2 aunts breast ca | Unknown | ? |
| **2** | 35, 6 months | Sepedi | None | inf duct | IV |

* Inf duct = infiltrating carcinoma of no special type; inf lobular = infiltrating lobular carcinoma; nos = carcinoma not otherwise specified.

## 2.2 DNA EXTRACTIONS

Peripheral blood samples were previously collected from South African women with breast cancer, who attended the Oncology Clinic at Steve Biko Hospital, Pretoria, between 1993 and 2001. Freezing of these samples may introduce DNA strand breaks and/or artifacts (Peng, S. et al. 2008), which may introduce new variants unrelated to the cancer diagnosed in the patient. But the DNA extracted from these samples was of high quality and in large amounts. DNA was previously extracted from peripheral blood samples using the method described by Johns and Paulus-Thomas (Johns and Paulus-Thomas 1989).

## 2.3 ETHICS APPROVAL

This study was approved by the Research Ethics Committee of the Faculty of Health Sciences, University of Pretoria (Protocol no. 260/2018). All experiments were performed in accordance with guidelines and regulations. The patients gave written informed consent for participation in the study.

## 2.4 ANALYSED CANCER GENES

The Illumina TruSight Cancer sequencing panel at The Institute of Cancer Research (ICR), London, which targets 94 cancer related genes was used (Table 2.2). This panel requires low sample DNA input and covers a wide range of cancer genes and SNPs across the human genome. The targeted sequences are 350–650 bases centered symmetrically around the midpoint of an 80-mer probe. The targets covers exonic DNA with 50 bp flanking non-coding regions.

All 94 genes were assessed for nonsense, frameshift, or splice-site variants affecting the invariant splice sites. A subset of nineteen established and candidate breast or ovarian cancer genes (ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, MLH1, MSH2, MSH6, NBN, NF1, PALB2, PMS2, PTEN, RAD51C, RAD51D, STK11 and TP53) were further investigated for all sequence variants. The results from the previous BRCA1/BRCA2 screening were also verified.

Table 2.2: Alphabetic list of genes analysed with the TruSight cancer panel.

*AIP, ALK, APC, ATM, BAP1, BLM, BMPR1A, BRCA1, BRCA2, BRIP1, BUB1B, CDC73, CDH1, CDK4, CDKN1C, CDKN2A, CEBPA, CEP57, CHEK2, CYLD, DDB2, DICER1, DIS3L2, EGFR, EPCAM, ERCC2, ERCC3, ERCC4, ERCC5, EXT1, EXT2, EZH2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, FH, FLCN, GATA2, GPC3, HNF1A, HRAS, KIT, MAX, MEN1, MET, MLH1, MSH2, MSH6, MUTYH, NBN, NF1, NF2, NSD1, PALB2, PHOX2B, PMS1, PMS2, PRF1, PRKAR1A, PTCH1, PTEN, RAD51C, RAD51D, RB1, RECQL4, RET, RHBDF2, RUNX1,*

*SBDS, SDHAF2, SDHB, SDHC, SDHD, SLX4, SMAD4, SMARCB1, STK11, SUFU, TMEM127, TP53, TSC1, TSC2, VHL, WRN, WT1, XPA,* and *XPC*

## 2.5 LIBRARY PREPARATION AND SEQUENCING

Patient DNA samples were sent to Omega Biotech in Georgia, USA, where DNA libraries were produced with the TruSight Rapid Capture kit (Illumina) and sequenced using the Illumina TruSight Cancer sequencing panel.

## 2.6 SEQUENCING DATA ANALYSIS

### 2.6.1 EXECUTION OF ANALYSIS

The sample set was collectively run through a BCBIO pipeline (June 2021 release, detailed tool versions provided in supplementary information) (Chapman, Kirchner et al. 2020) that includes the Genome Analysis Toolkit (GATK) best practices, Base Quality Score Recalibration (BQSR), realignment and HaplotypeCaller variant calling. The genome build that was used was GRCh37. *Bgzip* was used to compress the sample fastq's, indexing was done using by *grabix*. *FastQC* (version 0.11.7) was incorporated to evaluate the quality of fastq files. Reads were subsequently pre-processed with the FastX toolkit (version 0.0.14) to trim five nucleotides from the 5'- and 3'-ends of the 100 bp paired-end reads (G. 2018). *BWA MEM* (Li and Durbin 2009) was used to aligned sample sequences to the reference genome (UCSC hg19) and then sorting and creation of a bam file was handled by *samtools* while *Picard* handled the duplicate marking. The base values from the bam file were recalibrated using the *GATK* workflow (*Base Recalibrator*), the next step in the flow was variant calling (*HaplotypeCaller*). Afterwards, *GenotypeGVCFs* was ran for joint genotyping on the gVCF files. *SelectVariants* was used to split SNPs and indels and *VarianFitration with the* suggested manual filtration parameters was used to score variant quality for filtering purposes on both SNPs and indels. All steps after alignment forms part of the GATK best practices (Van der Auwera, Carneiro et al. 2013) approach.

GATK uses an integrated post-filtering system to validate the significance of all the variants tested. Different tests have been worked into the pipeline to ensure statistical significance of variants tested including, Pre-filtering: Base quality score recalibration (quality scores made based on how confident the sequencing machine was on calls it made). Post-filtering: Fisher's exact test (Calculates the confidence of a specific nucleotide being identified in a specific position), Inbreeding Coefficient (Measures excess heterozygosity at a position), HaplotypeCaller (Estimate a confidence level for the

presence of SNP's and indels at specific positions), Rank sum test (validate that the variant is not an artefact), Variant quality score recalibration (Assign quality scores to variants which is used to filter out non-relevant variants). Other calculations that affect the confidence levels will include: QualByDepth, FisherStrand, SrandOddsRatio, RMSMappingQuality, MappingQualityRankSumTest and ReadPosRankSumTest. The GATK pipeline has been specifically set-up to handle and evaluate the type of cancer samples present in this study. For significantly mutated pathways the program uses the PathScan algorithm and the Mutation relation test. The Clinical correlation test is incorporated to determine the relationship between observed mutations and clinical phenotypes. The Proximity analysis is another module part of the suite which has the function of identifying other overlooked mutations that cluster closely to more prominent mutations. Lastly, the Pfam annotation module groups significant genes together based on the frequency of mutations in specific protein domains. This helps to identify putative function of grouped genes.

All processes were executed on a Linux cluster with 10× nodes, each having 28× cores and 128 GB of RAM, running CentoS 7.4.

### 2.6.2 VARIANT CALLING

Variant calling was carried out using the HaplotypeCaller in gVCF mode. Variants were classified in accordance with the American College of Medical Genetics and Genomics guidelines (Richards, Aziz et al. 2015), as pathogenic, likely pathogenic, likely benign, benign or as variants of uncertain significance (VUS). For clarification, pathogenic/likely pathogenic variants were defined as "deleterious variants" linked to the condition "hereditary cancer-predisposition syndrome". Variants were described according to the Human Genome Variation Society (HGVS) recommendations (den Dunnen, Dalgleish et al. 2016). For BRCA1 the most common human transcript (NM_007294.3) was used with custom numbering of the exons (missing exon 4).

### 2.6.3 VARIANT ANNOTATION

Functional variant annotation was done using Variant Effect Predictor (VEP), the default parameters were used in concordance with documentation (McLaren, Gil et al. 2016). Quality-filtered variants were uploaded to the VEP web interface, and additional output fields were activated in the dbNSFP section for LRT_pred (Chun and Fay 2009), MutationTaster (Schwarz, Cooper et al. 2014), PROVEAN (Choi and Chan 2015), CADD (Kircher, Witten et al. 2014) and FATHMM (Shihab, Rogers et al. 2015). Filtering of common variants was not performed in VEP.

### 2.6.4 VARIANT FILTRATION AND IN SILICO EVALUATION OF VARIANTS

In-house Python code (available on request) was developed for the selection of variants for inclusion in this study. Variants with a minor allele frequency (MAF) of ≥ 1% in the 1000 Genomes African database were removed. For non-synonymous variants in the breast cancer susceptibility genes, the results of five in silico functional effect predictors were considered, being LRT_pred (Chun and Fay 2009), MutationTaster (Schwarz, Cooper et al. 2014), PROVEAN (Choi and Chan 2015), CADD (Kircher, Witten et al. 2014) and FATHMM (Shihab, Rogers et al. 2015) with variants being selected if at least 3/5 methods predicted a variant to be deleterious. A threshold of 2.0 for GERP_RS and 10.0 for CADD was used. For the other methods, a prediction of 'D' was selected. As VEP provides results for multiple transcripts per gene, only canonical transcripts are reported on, as determined by mapping of REFSEQ identifiers to Ensembl canonical transcripts via UCSC tables (Karolchik, Hinrichs et al. 2004), accessed July 2018.

### 2.6.5 ENRICHED PATHWAYS

The project also set out to test for any enriched pathways in the set of breast cancer patients. *PathScore* is a web-based executable program which quantifies the level of enrichment of somatic mutations in known pathways according to the Molecular Signatures Database (MsigDB) (Gaffney and Townsend 2016). The program uses a hypergeometric test to estimate pathway alteration probabilities and sets itself apart from other over-representation analysis tools because of three major points: 1) Estimates are split into individual patients rather than calculating over-representation by group, which give raise to patient specific pathway alteration probabilities, 2) the program takes gene transcript length into account, which means a larger gene has a higher probability of carrying a mutation than a shorter one, 3) Lastly, it incorporates empirically-derived background mutation rates to account for varied mutation probabilities across the whole genome, which is novel in over-representation analysis tools according to the authors (Gaffney and Townsend 2016).

The program takes a set of patient-gene pairs that represent all the genes that contained deleterious mutations as inputs. Other selectable options when running the program included: 'BMR-scaled gene length' and 'gene count', which can be modified. Outputs include matrix plots of patient-gene pairs, pie charts for corresponding pathways, volcano plots of altered pathways, overlapping relationship between pathways in the form of tree plots and comparison plots for the different affected pathways.

## 2.7 NATURE AND FORM OF RESULTS

- The research done here was written up in the form of a thesis for the partial fulfilment of the requirements for the degree PhD Bioinformatics.
- A research papers was accepted by Scientific Reports, which corresponds to Chapters 3 and 4.
- This project aims to increase the research community's knowledge on the mutational spread of breast cancer in South Africa, which up until now has not been focused on. These mutations may be novel, or they may share characteristics from other ethnic groups around the world.

## 2.8 DATA MANAGEMENT

Data is stored on a high performance and high capacity Lustre filesystem at the Centre of Bioinformatics and Computational Biology, University of Pretoria. Data is backed up to a ZFS JBOD. Data is kept confidential as it contains personal medical information. Variants identified in this study, together with the relevant sequences, will be made available in the European Variant Archives.

## 2.9 AUTHORS' LIST OF AGREEMENT FOR PUBLICATIONS AND PRESENTATIONS

Results obtained in this research study will be submitted for possible publication in international and local accredited journals. The agreement among the authors, regarding the order of appearance if the findings are published or presented, can be viewed in Table 2.3.

Table 2.3: The contribution and order of author's appearance, if results are published, in this research study.

| | Name | Department | Contribution | Author or Acknowledgement |
|---|---|---|---|---|
| 1. | Dewald Eygelaar | Bioinformatics and Computational Biology | Analysis, Author, Student | Author |
| 2. | Lizette Jansen van Rensburg | Genetics | Sample handling (Gathering, Extraction), | Author |

| | | | | |
|---|---|---|---|---|
| | | | Author, Co-ordinator | |
| 3. | Fourie Joubert | Bioinformatics and Computational Biology | Analysis, Author, Co-ordinator | Author |

## 2.10 REFERENCES

Chapman, B., R. Kirchner, L. Pantano, M. De Smet, L. Beltrame, T. Khotiainsteva, S. Naumenko, V. Saveliev, R. V. Guimera, I. Sytchev, J. Kern, C. Brueffer, G. Carrasco, M. Giovacchini, P. Tang, M. Ahdesmaki, S. Kanwal, J. J. Porter, S. Möller, V. Le, A. Coman, V. Svensson, bogdang989, M. Mistry, M. Edwards, J. Hammerbacher, B. Pedersen, P. Cock, apastore and S. Turner. (2020). "bcbio/bcbio-nextgen: v1.2.3." from https://doi.org/10.5281/zenodo.3743344.

Choi, Y. and A. P. Chan (2015). "PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels." Bioinformatics **31**(16): 2745-2747.

Chun, S. and J. C. Fay (2009). "Identification of deleterious mutations within three human genomes." Genome Res **19**(9): 1553-1561.

den Dunnen, J. T., R. Dalgleish, D. R. Maglott, R. K. Hart, M. S. Greenblatt, J. McGowan-Jordan, A. F. Roux, T. Smith, S. E. Antonarakis and P. E. Taschner (2016). "HGVS Recommendations for the Description of Sequence Variants: 2016 Update." Hum Mutat **37**(6): 564-569.

G., H. (2018). "The FastX Toolkit." Retrieved 2020/01/19, 2020, from http://hannonlab.cshl.edu/fastx_toolkit/.

Gaffney, S. G. and J. P. Townsend (2016). "PathScore: a web tool for identifying altered pathways in cancer data." Bioinformatics **32**(23): 3688-3690.

Johns, M. B., Jr. and J. E. Paulus-Thomas (1989). "Purification of human genomic DNA from whole blood using sodium perchlorate in place of phenol." Anal Biochem **180**(2): 276-278.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler and W. J. Kent (2004). "The UCSC Table Browser data retrieval tool." Nucleic Acids Res **32**(Database issue): D493-496.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure (2014). "A general framework for estimating the relative pathogenicity of human genetic variants." Nat Genet **46**(3): 310-315.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek and F. Cunningham (2016). "The Ensembl Variant Effect Predictor." Genome Biol **17**(1): 122.

Peng, L., W. S., Y. S., L. C., L. Z., W. S. and Q. Liu (2008). "Autophosphorylation of H2AX in a cell-specific frozen dependent way." (1090-2392 (Electronic)).

Richards, S., N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding and H. L. Rehm (2015). "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." Genet Med **17**(5): 405-424.

Schwarz, J. M., D. N. Cooper, M. Schuelke and D. Seelow (2014). "MutationTaster2: mutation prediction for the deep-sequencing age." Nat Methods **11**(4): 361-362.

Shihab, H. A., M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. Day, T. R. Gaunt and C. Campbell (2015). "An integrative approach to predicting the functional effects of non-coding and coding sequence variation." Bioinformatics **31**(10): 1536-1543.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel and M. A.

DePristo (2013). "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline." Curr Protoc Bioinformatics **43**(1110): 11.10.11-11.10.33.

# Chapter 3

# General results and pathogenic / likely pathogenic variants in known breast cancer susceptibility genes

## 3.1 INTRODUCTION

Breast cancer has been reported numerous times as the most prominent cancer in women (Torre, Bray et al. 2015, Siegel, Miller et al. 2016) and is 100 times more likely to be identified in females compared to males (Borgen, Wong et al. 1992, Fentiman, Fourquet et al. 2006). Globally, breast cancer has been ranked as having the second highest incidence level when compared to other cancers (Torre, Bray et al. 2015). Until recently, breast cancer was also rated second in America (DeSantis, Ma et al. 2014) but new estimates suggest that breast cancer may have overtaken lung cancer as the most prominent cancer for 2016 in America (Society 2016).

The GLOBOCAN 2020 database of the International Agency for Research on Cancer (IARC), estimated the current age standardised breast cancer incidence per 100,000 women in Southern (50.4), Western (41.5), Eastern (33), and Central Africa (32.7) with associated mortality rates estimated at 15.7, 22.3, 17.9, and 18, respectively (Ferlay, Laversanne et al. 2020). Newly diagnosed breast cancer cases in South Africa accounts for 27.1% of female cancers in 2020, with age-standardized (World) incidence and mortality rates of 52.6 and 16 (per 100,000 women) respectively (Ferlay, Laversanne et al. 2020).

Anatomically, the breast consists of soft tissue, connective and fatty tissues, blood- and lymph vessels, milk ducts and lobules, the areola and nipple. Thus far, cancer has been identified to affect the ducts, lobules, nipple and soft tissues (BreastCancer.org 2022). Cancers do affect these areas at varying levels, where the more common and less aggressive cancers are found in the ducts and lobules. More rare aggressive cancers have been identified in the nipple and soft tissues of the breast (BreastCancer.org 2022).

Cancer results from a process of genetic changes, some inherited, some induced by environmental exposures and some occurring by chance. (ACS 2022). Genetically, mutations in known oncogenes increases the risk of developing breast cancer (ACS 2022). The most well defined and understood oncogene to date is BRCA2 (Wooster, Bignell et al. 1995). In South Africa, studies have shown that

up to 47% of familial breast cancer are caused by mutations in the BRCA oncogenes (Reeves, Yawitch et al. 2004, van Rensburg, van der Merwe et al. 2007). Up to now, most breast cancer research in South Africa focused on the evaluation of the BRCA genes (Reeves, Yawitch et al. 2004, van Rensburg, van der Merwe et al. 2007, van der Merwe, Hamel et al. 2012) and very little work has been done on other genes. Except for BRCA1 and BRCA2, an association with breast cancer susceptibility has internationally also been reported for a further eleven high- to moderate-penetrance genes: TP53, PALB2, PTEN, STK11, CDH1, ATM, BRIP1, CHEK2, RAD51B, RAD51C, and RAD51D (Couch, Shimelis et al. 2017, Samadder, Giridhar et al. 2019). In addition, pathogenic variants in genes from the mismatch repair pathway (MLH1, MSH2, MSH6 and PMS2) have been identified in breast cancer and ovarian cancer patients (Couch, Shimelis et al. 2017).

One of the biggest challenges faced when researching breast cancer is the variety of different mutations that may cause the same type of cancer separately. Within the BRCA1 gene alone, 1 800 mutations have been identified (GHR 2020), and large amounts of mutations have been identified in the other prominent genes as well. Another point of consideration is that ethnic groups may be affected differently by gene mutations (Neuhausen 2000).

Therefore, different databases have been compiled to document all the various oncogenic mutations from studies around the world. The *1000 Genomes Project* is one example, it set out to sequence whole genomes of a large group of healthy individuals around the world to try and account for variants present in as many different ethnic groups as possible (Figure 3.1) (Auton, Brooks et al. 2015). The project is supported by the *Genome Reference Consortium* (GRC), a collaboration between top research and sequencing facilities to present researchers with a ‘normal’ reference genome to which they can compare their own subjects’ sequences for analysing possible variants. In this study, the hg19 (GRCh37) human reference genome was used, but recently GRCh38 has become the more common reference (Raney, Cline et al. 2011).


Figure 3.1: Graphical representation of the countries that have been included in the 1000 Genomes Project (Chen, Ghandikota et al. 2020).

Another well-known database, *ClinVar* (Landrum, Lee et al. 2016), is a compilation of the relationships of variants with the different clinically relevant phenotypes they cause. *ClinVar* mostly focuses on germline variants and they all are supported by research work done across the world. The database, if the information is available, can estimate the outcome of a variant, whether its non-pathogenic or pathogenic (Landrum, Lee et al. 2016). The *ExAC* (Exome Aggregation Consortium) aggregates sequencing data from large-scale exome studies (Lek, Karczewski et al. 2016). This data is publicly available and contains germline (blood) samples to all researchers, that use it to estimate variant frequencies. This database incorporates six well defined populations around the world: African/African-American, American, Finnish, non-Finnish European, South Asian and East Asians (Lek, Karczewski et al. 2016).

In 1998, the National Cancer for Biotechnology Information (NCBI) envisioned a general catalog of genome variation, the Single Nucleotide Polymorphism Database (dbSNP) (Smigielski, Sirotkin et al. 2000). The database was only established in 1999 through a collaboration between the NCBI and the National Human Genome Research Institute (NHGRI). It is populated with human single nucleotide variations, microsatellites, and small-scale insertions and deletions. These variants in most cases are accompanied by publications as well as molecular consequence, and genomic and RefSeq mapping information. Also, it freely offers users/researchers half a billion non-redundant and uniquely accessioned Reference SNP (RefSNP) records.

The next well known database, The Cancer Genome Atlas (TCGA), was established in 2006 through a joint effort by the National Cancer Institute and the National Human Genome Research Institute(NCI

2021). TGCA originally started out to only characterize three cancers: glioblastoma multiforme, lung, and ovarian cancer (Nosrati Nahook and Sh 2021), but over the years has been expanded to include over 20 000 sets of cancer versus normal tissues over 33 different cancer types(NCI 2021). Data from this database is available to researchers, where some samples are open access, and others require the permission of a data access committee (DAC).

The International Cancer Genome Consortium (ICGC) on the other hand is a global initiative to build a database containing somatic mutations from major tumour types. This collaboration has yielded mutational abnormalities from 50 different cancer types and is ever growing (Zhang, Baran et al. 2011). The database already contained over 77 million somatic mutations in 2019 (Zhang, Bajari et al. 2019).

The Genome Aggregation Database (gnomAD) database made an effort to compile and compare exome and genome sequencing datasets from a wide variety of large-scale sequencing projects. Data from this database is freely available to further scientific knowledge. In its infancy the database only contained exome data and was called the Exome Aggregation Consortium (ExAC) but was later merged into the gnomAD database and ceases to exist.

Arguably the most well-defined database that frequently gets referred to or compared against, is the Genome-wide association studies (GWAS) catalogue (Welter, MacArthur et al. 2014). Currently, the database contains a collection of over a 100 000 different literature-derived SNP's. The GWAS catalogue was founded by the National Human Genome Research Institute (NHGRI) in 2008 where after the European Bioinformatics Institute (EMBL-EBI) joined them in a collaboration from 2010 to continue building it (Welter, MacArthur et al. 2014).

Germline mutations are present in egg/sperm cells of an individual, which are then inherited by the progeny, where they are present in all the cells. These types of mutations may ultimately give rise to different cancer family syndromes. Somatic mutations are accumulated throughout one's life and are not inherited by the progeny. Research has shown that the frequency of somatic mutations is far higher than that of germline mutations (Milholland, Dong et al. 2017).

Germline breast cancer data can be analysed in different ways including NGS and microarrays. Microarrays and other chip-related assays are a cost-effective alternative to the more expensive NGS analysis with the ability to compare a wide range of known variants across a large amount of samples (Liu, So et al. 2015). On the other hand, NGS can be used for whole genome sequencing which leads to the identification of novel variants in cancer related genes. NGS can be downscaled in terms of the parts of the genome that are sequenced to reduce computational time as well as costs

by using whole exome sequencing or targeted panel sequencing which would only include specific genes for the analysis (Liu, So et al. 2015).

Several studies utilizing NGS gene panels have been carried out, mainly on breast cancer cases from west European and Asian populations (Easton, Pharoah et al. 2015, Castéra, Harter et al. 2018). Some studies have included African-Americans, but this data can be difficult to interpret in an African context due to the fact that they are an admixed population. The estimated proportion of African, European and Native American ancestry in African-American groups vary from 76 to 85% African, 14% to 21% European and 1% to 3% Native American ancestry (Baharian, Barakatt et al. 2016). Most South African breast cancer research has focussed on the analysis of the well-defined BRCA genes, but studies have shown females developing breast cancer without BRCA mutations present (Bayraktar, Gutierrez-Barrera et al. 2011, Noori, Gangi et al. 2014). It is evident that more in depth research is required in other cancer-related genes (eg. ATM, BARD1, CHEK2, NBN, PALB2 and RAD51, just to name a few). In some cases, the effect that mutations in the other genes might have in breast cancer is still unclear (Robson 2016) and warrants further research.

Here, we discuss genes that have been reported in the literature to have the highest likelihood of being related to breast cancer susceptibility, specifically BRCA1, BRCA2, TP53 (high penetrance), ATM, CHEK2, BRIP1, PALB2, RAD50, NBN (medium penetrance) and PTEN, RAD51C, BARD1, STK11, CDH1 (low penetrance) (Mahdavi, Nassiri et al. 2019). Thereafter, we provide general results from the study. This is followed by a report of known or highly-likely pathogenic variants associated with breast cancer susceptibility. Additional types of variants are described in subsequent chapters.

The BRCA1 gene was localized to the long arm of chromosome 17 in 1990 through the use of genetic linkage (Ford, Easton et al. 1998) and contains 23 exons (Kwong, Chen et al. 2015), but cloning of the gene only occurred four years later (Miki, Swensen et al. 1994). The BRCA2 gene was identified on the long arm of chromosome 13 in 1994 (Ford, Easton et al. 1998) and consists of 27 exons (Kwong, Chen et al. 2015). The BRCA1 protein carries 1 863 amino acids and consists out of an N-terminal ring domain and two tandem BRCA1 C Terminus (BRCT) domains (Wu, Paul et al. 2016) and BRCA2 contains 3 418 amino acids (Shamoo 2003).

Table 3.1: A selection of some well-known variants in BRCA1 and BRCA2 founder mutations across different populations.

| **Population or subgroup** | **BRCA1 mutation(s)(Chen, Morrical et al. 2015)** | **Reference(s)** |
|---|---|---|

| | | |
|---|---|---|
| African-Americans | BRCA1 943ins10, M1775R | (Gao, Neuhausen et al. 1997, C., L. et al. 1999) |
| Afrikaners | BRCA1 1374delC, 2641G>T<br>BRCA2 7934delG | (Reeves, Yawitch et al. 2004, van der Merwe and Jansen van Rensburg 2009) |
| Ashkenazi Jewish | BRCA1 185delAG, 188del11, 5382insC,<br>BRCA2 6174delT | (Struewing, Abeliovich et al. 1995, Tonin, Serova et al. 1995, Neuhausen, Gilewski et al. 1996) |
| Austrians | BRCA1 2795delA, C61G, 5382insC, Q1806stop | (Wagner, Möslinger et al. 1998) |
| Belgians | BRCA1 2804delAA, IVS5+3A>G | (Peelen, van Vliet et al. 1997, Claes, Machackova et al. 1999) |
| Dutch | BRCA1 Exon 22 deletion, exon 13 deletion, 2804delAA | (Peelen, van Vliet et al. 1997, Petrij- |

| | | |
|---|---|---|
| | | Bosch, Peelen et al. 1997, Verhoog, van den Ouweland et al. 2001) |
| Finns | BRCA1 3745delT, IVS11-2A>G,<br>BRCA2 999del5, IVS23-2A>G | (Pääkkönen, Sauramo et al. 2001) (Huusko, Pääkkönen et al. 1998) |
| French | BRCA1 3600del11, G1710X | (Muller, Bonaiti-Pellié et al. 2004) |
| French Canadians | BRCA1 R1443X<br>BRCA2 8765delAG | (Simard, Tonin et al. 1994) (Tonin, Mes-Masson et al. 1999) |
| Germans | BRCA1 5382insC, C61G | (Backe, Hofferbert et al. 1999) |
| Greeks | BRCA1 5382insC | (Ladopoulou, Kroupis et al. 2002) |
| Hungarians | BRCA1 300T>G, 5382insC, 185delAG<br>BRCA2 9326insA | (Van Der Looij, Szabo et al. 2000) |
| Icelanders | BRCA2 999del5 | (Thorlacius, Olafsdottir |

| | | |
|---|---|---|
| | | et al. 1996) |
| Italians | BRCA1 5083del19<br>BRCA2 8765delAG | (Baudi, Quaresima et al. 2001) (Pisano, Cossu et al. 2000) |
| Japanese | BRCA1 L63X, Q934X | (Sekine, Nagata et al. 2001) |
| Latvians | BRCA1 C61G, 5382insC, 4153delA | (Csokay, Tihomirova et al. 1999) |
| Native North Americans | BRCA1 1510insG, 1506A>G | (Liede, Jack et al. 2002) |
| Northern Irish | BRCA1 2800delAA<br>BRCA2 6503delTT | (Consortium " 2003) |
| Norwegians | BRCA1 816delGT, 1135insA, 1675delA, 3347delAG | (Borg, Dorum et al. 1999, Heimdal, Maehle et al. 2003) |
| Pakistanis | BRCA1 2080insA, 3889delAG, 4184del4, 4284delAG, IVS14-1A>G<br>BRCA2 3337C>T | (Liede, Malik et al. 2002) |
| Polish | BRCA1 300T>G, 5382insC, C61G, 4153delA | (Gorski, Byrski et al. 2000, Perkowska, BroZek et al. 2003) |
| Russians | BRCA1 5382insC, 4153delA | (Gayther, |

| | | |
|---|---|---|
| | | Harrington et al. 1997) |
| Scottish | BRCA1 2800delAA<br>BRCA2 6503delTT | (Liede, Cohen et al. 2000) (Consortium " 2003) |
| Slovenians | BRCA2 IVS16-2A>G | (Krajc, De Greve et al. 2002) |
| Spanish | BRCA1 R71G<br>BRCA2 3034delAAAC(codon936), 9254del5 | (Vega, Campos et al. 2001) (Osorio, Robledo et al. 1998, Campos, Diez et al. 2003) |
| Swedish | BRCA1 Q563X, 1201del11, 2594delC, 3166ins5, 3171ins5<br>BRCA2 4486delG | (Johannsson, Ostermeyer et al. 1996, Bergman, Einbeigi et al. 2001) (Hakansson, Johannsson et al. 1997) |

Adapted from (Neuhausen 2000).

These genes and their proteins play important roles in repairing cell damage and maintaining normal growing breast cells but at different stages. Notably not all variants identified in BRCA are associated with increased risk for breast cancer. Some of these variants have been coupled with other cancers while others have no effect on cancer risk at all.

Female carriers of pathogenically mutated BRCA1 and BRCA2 genes have a 55-65% and 45% increased chance of developing breast cancer respectively before the age of 70 (Kwong, Chen et al. 2015). Early detection of these mutations along with primary and secondary cancer prevention strategies will decrease this risk.

Table 3.1 contains a list of some well-known BRCA founder mutations previously identified around the world. Differences in disease prevalence have been observed in many ethnic groups, a good example is sickle-cell anaemia which is identified more in people from African descent (Kaback, Lim-Steele et al. 1993). A founder mutation can occur in a small population or a bottle-necked population when a specific mutation expands and the mutation gets fixed in the larger population. In some cases, the detrimental mutation may also have a positive effect and lead to an increase in fitness (Neuhausen 2000).

Work has shown that the BRCA proteins may be associated or may have interactions with the following variety of cancer related genes: ATM (Gatei, Zhou et al. 2001), BARD1 (Nishikawa, Ooka et al. 2004), BRCC3 (Dong, Hakimi et al. 2003), BRIP1 (Yu, Chini et al. 2003), CHEK2 (Chabalier-Taste, Racca et al. 2008), LMO4 (Sutherland, Visvader et al. 2003), MRE11A (Yuan, Hou et al. 2012), MSH2 (Guerrette, Wilson et al. 1998), MSH3 (Guerrette, Wilson et al. 1998), MSH6 (Guerrette, Wilson et al. 1998), Myc (Mac Partlin, Homer et al. 2003), NBN (Berlin, Lalonde et al. 2014), NPM1 (Falini, Nicoletti et al. 2007), P53 (Abramovitch S. 2003), PALB2 (Xia, Sheng et al. 2006), RAD51 (Pellegrini, Yu et al. 2002), RELA (Kim, Gazourian et al. 2000), RB1 (Ali, Parsam et al. 2010), SMARCA4 (Medina, Romero et al. 2008) and CDC48 (p97) (Ye 2006).

The ataxia–telangiectasia-mutated (ATM) protein kinase has been identified to control cell cycle initiation/arrest to some degree. A gene aptly named after the condition it leads to when mutations are present, it plays a pivotal role in the phosphorylation of specific proteins that are involved in cell cycle checkpoint control, apoptotic responses, and DNA repair (Shiloh 2006). Mutations in ATM have been linked to an increase in risk of cancer development (Shiloh 1997, Petrini 2000, Shiloh 2006).The gene is located on the q arm of chromosome 11 and gives rise to 66 exons (Ahmed and Rahman 2006). Previously, the mutated gene has been shown to be associated with different cancers including lymphoid and epithelial cancers. People diagnosed with ataxia-telangiectasia are estimated to be 100-fold more prone to cancer (Ahmed and Rahman 2006).

The checkpoint kinase 2 (CHEK2) gene was identified to be a mammalian homolog of *S. cerevisiae* RAD53 and *S. pombe* Cds1 in 1998 (Matsuoka, Huang et al. 1998). This gene is localized to the human chromosome 22q12.1 and forms part of a cascade of activation/phosphorylation of proteins

in the occurrence of DNA damage and activates BRCA1 and p53 through phosphorylation (Vahteristo, Bartkova et al. 2002). With this knowledge and further research this gene has emerged as a good candidate tumour suppressor gene (Bartek, Falck et al. 2001, Bartek and Lukas 2003).

BRCA1 Interacting Protein C- terminal helicase 1 (BRIP1), previously known as Brca1-Associated C-terminal Helicase (BACH1), a tumour suppressor gene has previously been reported to have a possible role in breast cancer susceptibility . It was discovered in 2001 and was identified as a gene of interest because of its interaction with BRCA1 (Cantor, Bell et al. 2001). More recent work has shown a decrease in BRIP1's value as a BC susceptibility gene with possibly a more active role in ovarian cancer (Seal, Thompson et al. 2006, Rafnar, Gudbjartsson et al. 2011, Easton, Lesueur et al. 2016). Research has classified BRIP1 mutations as important inducers of germline breast cancer but this may only be in more rare mutations (Moyer, Ivanovich et al. 2020).

During further research into BRCA2 complexes in 2006, the Partner and localizer of BRCA2 (PALB2) was described for the first time (Xia, Sheng et al. 2006). It has been classified as the third most prevalent breast cancer gene after BRCA1 and BRCA2 (Evans and Longo 2014, Snyder, Metcalfe et al. 2015). The translated protein has a three-fold function, an interaction with BRCA2 to localize the formed complex into the nucleus of the cell where it prevents accumulation of DNA damage, serves as a molecular scaffold in the BRCA1-PALB2-BRCA2 complex and also with the help of BRCA2 and RAD51 replace replication protein A's on processed single-stranded DNA ends (Stecklein and Jensen 2012).

The RAD50 gene was originally isolated from *Saccharomyces cerevisiae* mutants which had an abnormally high sensitivity to DNA damage (Cox and Parry 1968). RAD50 is a highly conserved gene which has been implicated in double stranded DNA repair (Bhaskara, Dupré et al. 2007, Lamarche, Orazio et al. 2010, Williams, Lees-Miller et al. 2010).

RecA-like DNA strand transferase (RAD51) mutations have been associated with increased chances of developing breast cancer (Chen, Morrical et al. 2015). Also, it plays a role in repair of DNA double strand breaks and homologous recombination (Chen, Morrical et al. 2015). Research has shown that RAD51 interacts directly or indirectly with the following cancer associated genes/proteins: ATM, BRCA1, BRCA2, p53 and PALB2 (Richardson 2005).

The nibrin (NBN) gene has previously been associated with breast cancer (Lu, Wei et al. 2006). NBN interacts with MRE11A and RAD50 to repair DNA, NBN's main function is to transport MRE11A and RAD50 to the cell nucleus and guide them to the damaged DNA (Lu, Wei et al. 2006). The gene is located on chromosome 8q21 and mutations in this gene is characterised by a disorder called

Nijmegen breakage syndrome, which may lead to microcephaly, growth retardation, immunodeficiency, and cancer susceptibility (Lu, Wei et al. 2006, di Masi and Antoccia 2008).

The Phosphatase and Tensin homolog (PTEN) gene codes for the production of a phosphatase enzyme that regulates cell division by acting as a tumour suppressor (Shaw and Cantley 2006). The functionality of the gene was discovered in 1997 by three independent research groups (Li and Sun 1997, Li, Yen et al. 1997, Steck, Pershouse et al. 1997). Different from its phosphatase function, PTEN has been linked to various other biological mechanisms including binding of TP53 and anaphase-promoting complex/cyclosome (APC/C), increasing both their transcriptional and tumour suppressive activity (Freeman, Li et al. 2003, Song, Carracedo et al. 2011) and also a role in DNA repair, where PTEN binds to damaged chromatin and recruits RAD51 to initiated DNA repair (Ma, Benitez et al. 2019). Mutations have been linked to a variety of cancers including thyroid cancer, melanomas, lung cancer and low grade/secondary glioblastoma multiforme (GBM) tumours (Tamguney and Stokoe 2007).

Until recently the importance of RAD51C mutations in breast cancer susceptibility was relatively questionable while research has clearly made a connection to ovarian cancer susceptibility (Clague, Wilhoite et al. 2011, Pelttari, Heikkinen et al. 2011, Loveday, Turnbull et al. 2012, Thompson, Boyle et al. 2012, Blanco, Gutiérrez-Enríquez et al. 2014, Song, Dicks et al. 2015, Jønson, Ahlborn et al. 2016). As with a previously discussed gene, BRIP1, RAD51C drew attention because of its interactions with other known breast cancer susceptibility genes. Mutations in RAD51C have been associated with tumour formation.

BRCA1-associated RING domain protein-1 (BARD1 was first described as a BRCA1-interacting protein in 1996 (Wu, Wang et al. 1996). The formed protein has been shown to have a structural similarity with BRCA1, including shared N-terminal RING finger domains as well as BRCA1 C-terminal (BRCT) domains (Brzovic, Meza et al. 2001). With the use of their N-terminal RING finger domains, both have the ability to form homodimers (Brzovic, Rajagopal et al. 2001). This gene encodes a protein that binds to the N-terminal region of BRCA1 and together these proteins play a role in DNA repair/tumour suppression specifically double-stranded DNA breaks (Woditschka, Evans et al. 2014). Interactions between BARD1 and BRCA1 are hampered by tumorigenic amino acid substitutions in BRCA1 (Woditschka, Evans et al. 2014). A mutation in BARD1 does not necessarily mean BRCA1 is also mutated. A recent study has shed light on the breast cancer association of BARD1 and promising results were established regarding germline breast cancer mutations. The study concluded with a significant role of BARD1 in germline mutations related to breast cancer and suggested that a more

intensive focus should be placed on germline pathogenic variants (Weber-Lassalle, Borde et al. 2019).

*Serine/threonine kinase 11* (STK11) also known as liver kinase B1 (LKB1) or renal carcinoma antigen NY-REN-19 was identified as a tumour suppressor in 1997 (Hemminki, Tomlinson et al. 1997). The transcribed protein has been shown to be involved in various processes, including apoptosis (Esteve-Puig, Gil et al. 2014), cell control (Tiainen, Vaahtomeri et al. 2002, Scott, Nath-Sain et al. 2007) , metabolism (Spicer and Ashworth 2004, Esteve-Puig, Canals et al. 2009), polarity (Baas, Kuipers et al. 2004) and DNA damage response (Esteve-Puig, Gil et al. 2014). STK11 has also been identified to control/activate other important cancer-related genes, PTEN (Mehenni, Lin-Marq et al. 2005) and TP53 (Zeng and Berger 2006, Hou, Liu et al. 2011), through phosphorylation.

CDH1 is a gene that can be found on chromosome 16 and translates into the *E-cadherin* protein (Berx, Cleton-Jansen et al. 1995). The origin of this protein dates to 1981, where it was labelled, *Uvomorulin* (Hyafil, Babinet et al. 1981). Over the years this protein carried several different names: *Arc-1* (Imhof, Vollmers et al. 1983), *cell-CAM 120/80* (Damsky, Richa et al. 1983), *E-cadherin* (Shirayoshi, Okada et al. 1983), *L-CAM* (Gallin, Edelman et al. 1983). The gene identifies as a tumour suppressor and its mutated counterpart has previously been linked to lobular breast cancer (Berx, Cleton-Jansen et al. 1995, Sarrió, Moreno-Bueno et al. 2003, Benusiglio, Malka et al. 2013, Corso, Intra et al. 2016).

Tumour protein p53 (TP53/p53): This protein serves as a multi-functional transcription factor which influences cell cycle progression, cell survival and DNA integrity in cells where DNA-damaging agents are present (Pharoah, Day et al. 1999). Research shows that in breast tumours, frequently mutations are found in the p53 tumour suppressor gene (Ingvarsson 2001).

To our knowledge, only two studies in Africa, one on Nigerian women (Zheng, Walsh et al. 2018) and one on women from Uganda and Cameroon (Adedokun, Zheng et al. 2020) have used multigene panel sequencing to test for germline variants in patients, unselected for family history or age at diagnosis. In the present study, we included South African women of African ancestry (self-identified) diagnosed with breast cancer, who were unselected for age at diagnosis or family history of cancer. With the exception of four cases, all others were previously investigated for BRCA1 and BRCA2 pathogenic variants using alternate methods and were deemed negative for pathogenic/likely pathogenic BRCA1/BRCA2 variants. We used targeted next-generation sequencing of a multigene panel, comprised of 94 cancer susceptibility genes (Illumina TruSight cancer panel) in order to assess the frequency of deleterious germline variants in this cohort.

## 3.3 RESULTS

A total of 165 breast cancer patients of African ancestry (self-reported), patients were not specifically selected because of a family history of breast cancer or their age at diagnosis, were included in this study (Table 2.1). Their mean age (SD) at diagnosis was 41.28 (7.35) years (age range 22 to 54 years). Figure 3.2 depicts the patients' age at diagnosis displayed in 5-year intervals. Furthermore, 9% (15/165) of the patients reported either a 1st and/or 2nd degree relative with breast and/or ovarian cancer (Table 2.1). Sequencing was performed with the Illumina Trusight Cancer kit for 94 cancer-related genes and 284 SNPs that had previously been identified in GWAS studies to be associated with cancer.

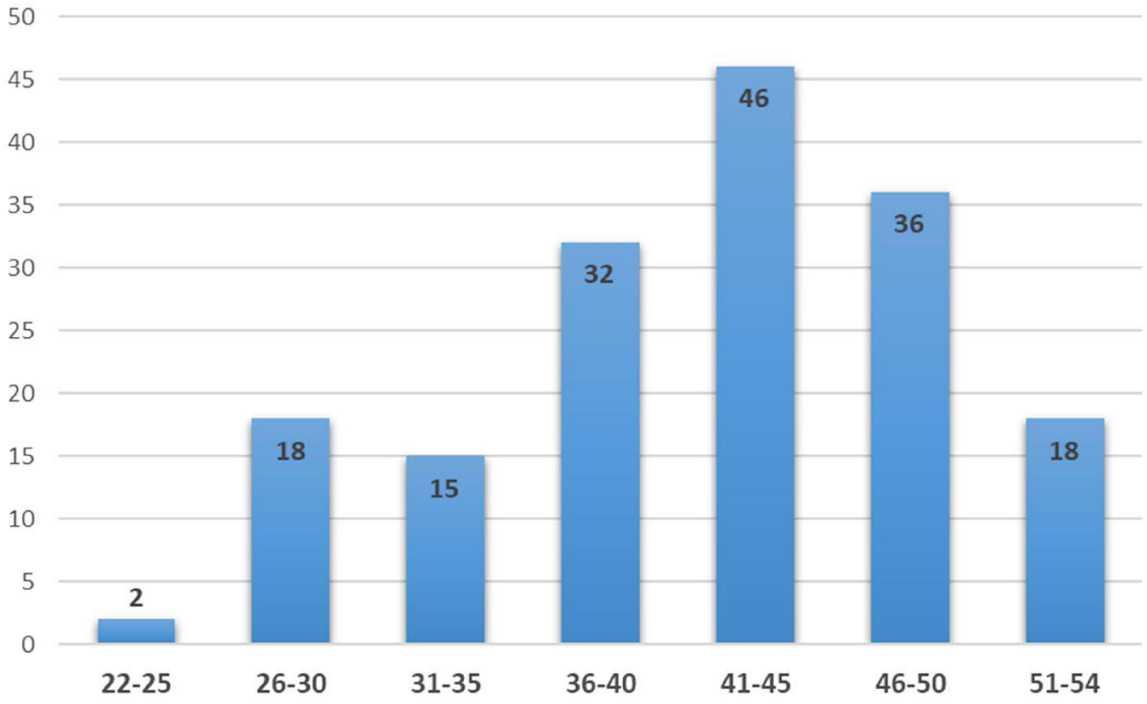


Figure 3.2: Distribution of patient age at first breast cancer diagnosis displayed in five year intervals (Figure generated using Microsoft Excel).

Nine percent (15/165) of the patients reported either a 1st and/or 2nd degree relative with breast and/or ovarian cancer (Supplementary Table 3.1). Information on the histology type was available for 145 of the 165 patients. The most common type was infiltrating ductal carcinoma (81.8%), followed by medullary ductal carcinoma (2.4%), invasive lobular carcinoma (1.2%), and at 0.6% each, tubular ductal carcinoma, papillary carcinoma, infiltrating mucinous carcinoma and carcinoma not otherwise specified. Cancer grade information was unavailable for eight of the 165 patients. Only

one of the patients was diagnosed with grade I (0.6%), 40 (24.2%) with grade II, 46 (27.9%) with grade III and 70 (42.4%) with grade IV breast cancer. High-grade tumours (grade III and IV) were by far the most common, accounting for 70.3% of all carcinomas.

### 3.3.1 GENERAL RESULTS

Initially, a total of 1 616 variants were identified for the various patients. According to VEP, initially 135 (8.4%) of these were novel variants. Coding variants included: synonymous (50%), missense (48%), stop gained (1%), inframe deletion (1%). Of these, 22 (2.1%) were identified for BRCA1 and 29 (2.8%) for BRCA2. Other genes that contained high numbers of mutations included: ALK (3.1%), APC (2.0%), ATM (2.8%), FANCA (3.0%), NSD1 (2.0%), RECQL4 (2.8%), and SLX4 (4.2%). The type of mutation that was most prominent overall in the mutation list was downstream gene variants (18%), followed by synonymous variants (17%), missense variants (16%) and intron variants (12%). The total number of variants was further reduced to 1 153, to only include non-synonymous and synonymous variants that were present at less than 1% in the 1000 Genomes African database to discard more common polymorphisms. The different variants identified in the coding regions included: missense (49%), splice region variant (40%), synonymous (30%), splice donor variant (14%), inframe deletions (5%), frameshift variants (0.6%), inframe insertions (0.4%), stop gained (0.3%), 5 prime UTR variant (0.3%) and splice acceptor variant (0.3%).

### 3.3.2 VARIANTS OF INTEREST

In the final data set, variants classified as inframe insertions and deletions, truncating, nonsense, frameshift, or splice-site variants affecting the invariant splice sites were retained. Additionally, non-synonymous variants were filtered by concordant deleterious effect prediction for missense in the breast cancer susceptibility genes (selected if predicted by at least 3/5 methods), population allele frequencies (< 1% in African populations of 1000 genomes phase 1 and 3), read depth (≥ 20). This resulted in the identification of 52 unique variants in 20 genes. Of these 52 variants, eleven were classified as Pathogenic variant/Likely Pathogenic variant (PV/LPV) and are presented in detail in Table 3.2 and are primarily discussed in this chapter. Fourteen were classified as benign/likely benign (Supplementary Table 3.1). These variants were present in 76 of the patients in Figure 3.6. Missense variants dominated (39), followed by frameshift variants (3), nonsense variants (3), variants affecting canonical splice sites (3), and in-frame deletions (3) (Figure 3.3).

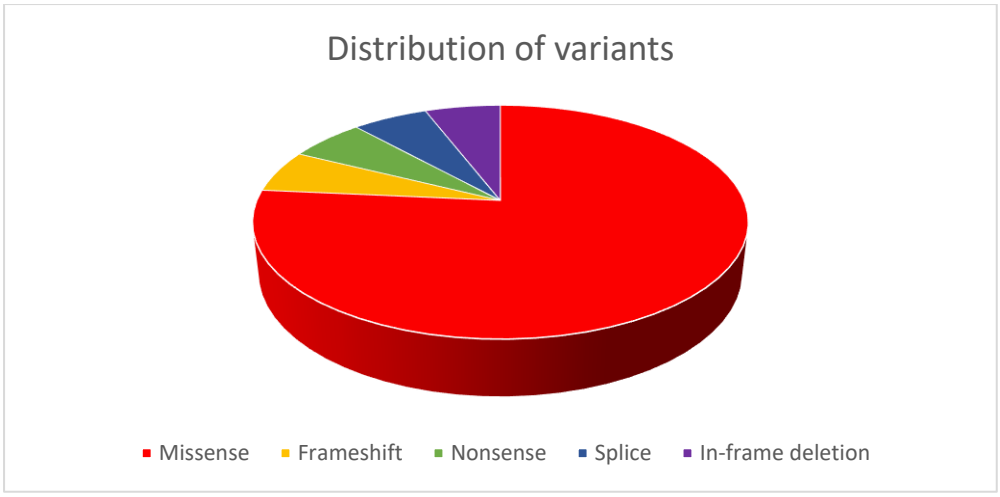


Figure 3.3: Distribution of pathogenic relevant variants (Figure generated using Microsoft Excel).

The concordance of the five in silico functional effect predictors is shown in Figure 3.4 (full concordance for 61 variants), and the number of distinct deleterious variants per predictor is shown in Figure 3.5.

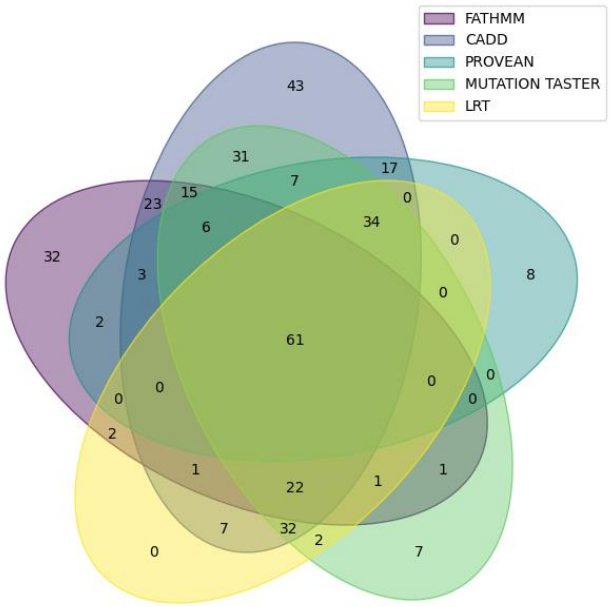


Figure 3.4: Concordance of variant effect predictors for distinct deleterious variants.

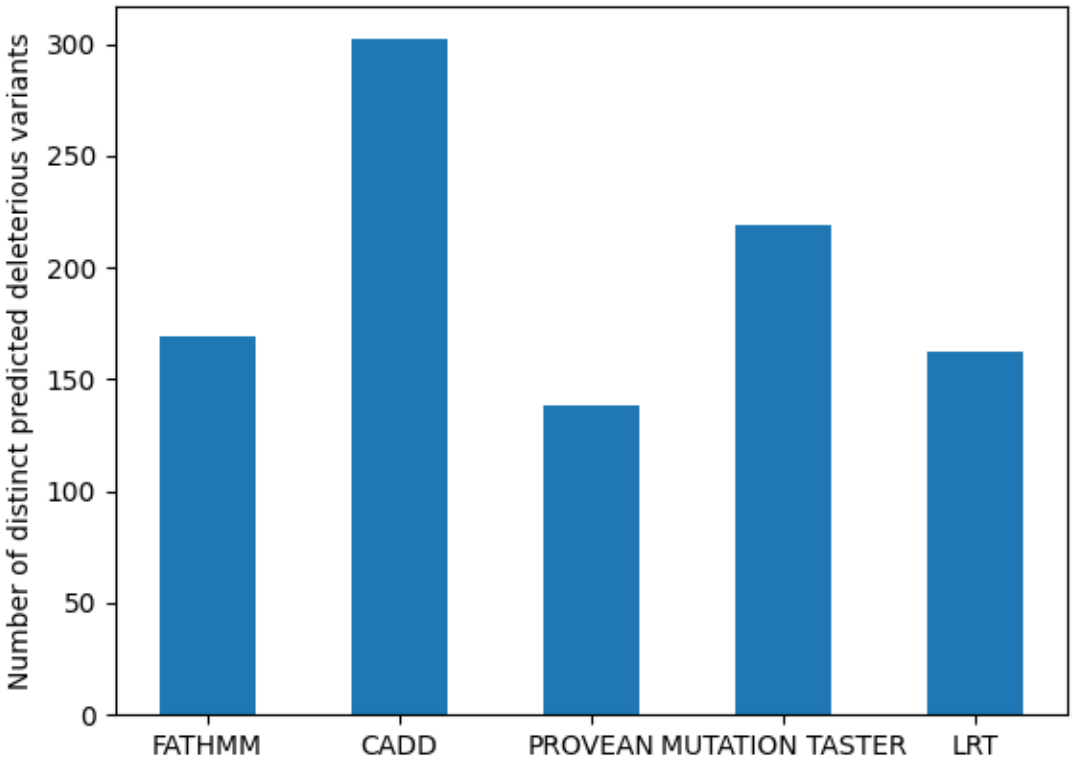


Figure 3.5: Number of distinct predicted deleterious variants by predictor.

Variants classified as in-frame insertions or deletions, truncating, nonsense, frameshift, or splice-site variants affecting the invariant splice sites as well as variants indicated as deleterious by 3/5 functional effect predictors have been summarized in Figure 3.6.

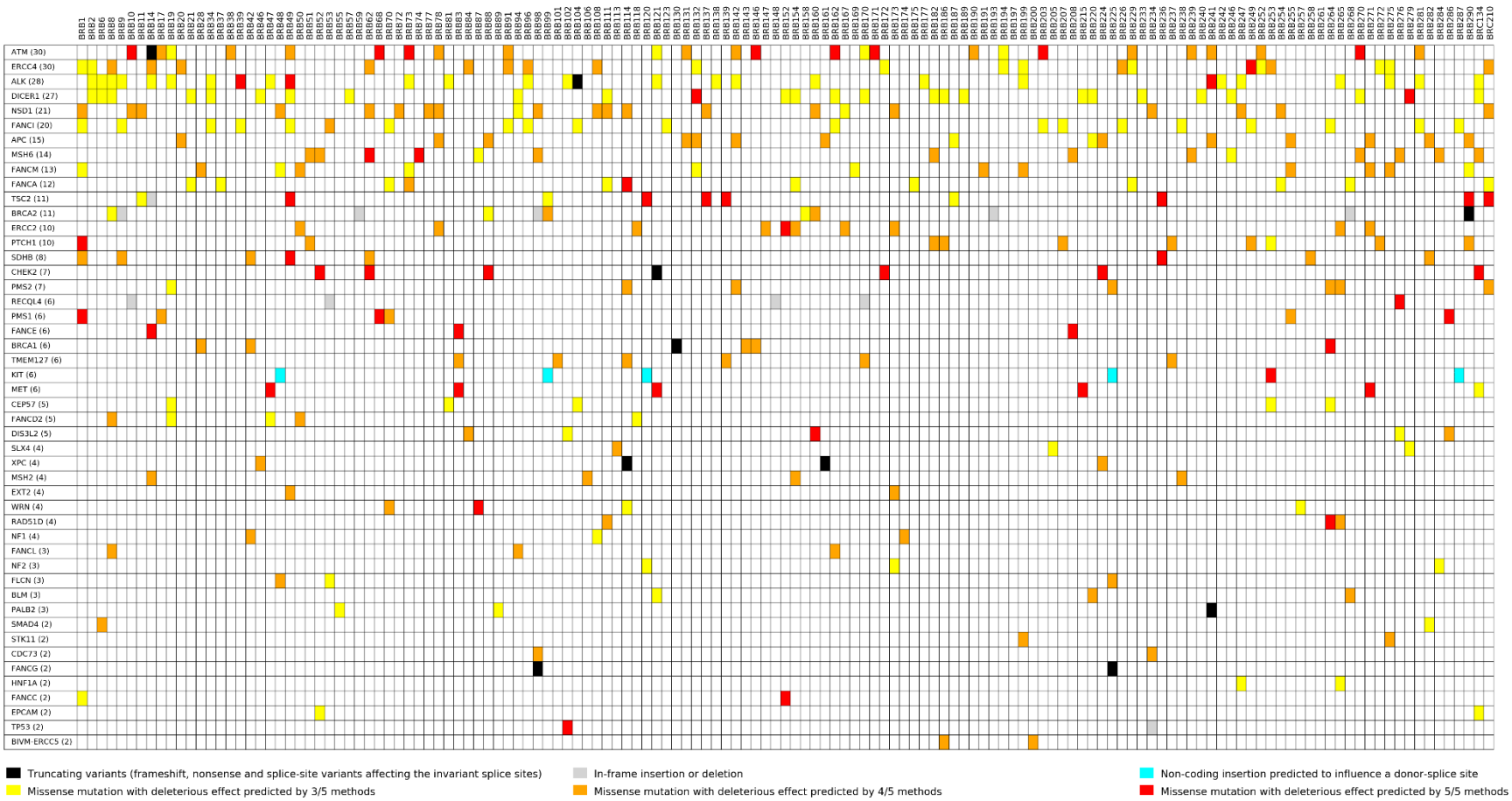


Figure 3.6: Matrix of patients vs. genes with sequence variants in breast cancer susceptibility genes and genes exclusively investigated for truncating variants (multiple variants per gene may be present). The genes are sorted from the most to least number of variants per gene as indicated in brackets. Black indicates truncating variants (frameshift, nonsense and splice-site variants affecting the invariant splice sites); Grey indicates an in-frame insertion or deletion. Missense variants are indicated according to the five in silico functional effect predictors, where yellow indicates a deleterious effect predicted by 3/5 methods, orange 4/5 methods and red 5/5 methods (Figure generated using Matplotlib 3.4.2: https:// matplotlib.org).

Genes in the figure were sorted by the amount of variants linked to them. ATM (12), BRCA2 (7), MSH6 (6) and BRCA1 (3) were the most highly represented genes. Variants marked with black (10) indicated carriers of truncating variants (frameshift, nonsense and splice-site variants affecting the invariant splice sites). For the remaining variants, red (53) represented patients with a variant that was classified as deleterious by 5/5 effect predictors, orange (148) was 4/5 effect predictors and yellow (131) was 3/5 and grey (11) indicated in-frame insertions and deletions. Patients marked with a blue colour (5) were carriers of a non-coding insertion that disrupted a splice site.

### 3.3.3 PATHOGENIC / LIKELY PATHOGENIC VARIANTS IN KNOWN BREAST CANCER SUSCEPTIBILITY GENES

This section focuses on the 11 genes that were identified as pathogenic / likely pathogenic (see section 3.3.2). Six patients (3.6%) were found to carry a pathogenic or likely pathogenic (P/LP) variant in one of five **known breast cancer susceptibility genes**: 1.2% in BRCA1, 0.6% in each of BRCA2, ATM, CHEK2 and PALB. A further seven patients carried deleterious variants in one of five **hereditary cancer predisposition genes** exclusively investigated for truncating variants, specifically ALK, BUB1B, FANCG, RB1 and XPC (Table 3.2), which will be discussed detail in the next chapter. None of these patients reported any family history of cancer.

Table 3.2: Pathogenic/likely pathogenic variants detected in a South African breast cancer cohort of African ancestry. *Reference sequences obtained from the NCBI database. For BRCA1 the most common human transcript (NM_007294.3) was used with custom numbering of the exons (missing exon 4). Variant nomenclature is according to the Human Genome Variation Society (HGVS) where complimentary DNA (cDNA) numbering + 1 corresponds to the A of the ATG translation initiation codon. # Not reported in dbSNP (http:// www. ncbi.nlm.nih.gov/ SNP), EVS (http://evs gs. washington. edu/ EVS), gnomAD (https:// gnomad.broadinstitute. org) or ClinVar (https://www. ncbi.nlm.nih. gov/ clinvar).

| Gene (RefSeq)* | Nucleotide change | Location | Predicted protein consequence | dbSNP | Patient | Age at diagnosis (yrs:mnths) |
|---|---|---|---|---|---|---|
| **Pathogenic variants in known breast cancer susceptibility genes** | | | | | | |
| *ATM (NM_000051.3)* | c.162T>A | Exon 3 | p.Tyr54Ter | - | BRB14 | 47:8 |
| *BRCA1* | c.4524G>A | Exon 15 | p.Trp1508Ter | rs80356885 | BRB130 | 45:8 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *(NM_007294.3)* | c.5096G>A | Exon 18 | p.Arg1699Gln | rs41293459 | BRB264 | 42:3 |
| *BRCA2 (NM_000059.3)* | c.5771_5774del | Exon 11 | p.Ile1924ArgfsTer38 | rs80359535 | BRB290 | 26:6 |
| *CHEK2 (NM_001005735.1)* | c.283C>T | Exon 2 | p.Arg95Ter | rs587781269 | BRB121 | 54:0 |
| *PALB2 (NM_024675.3)* | c.2835-1G>C | Intron 8 | p.(?) | rs515726099 | BRB241 | 40:1 |
| **Pathogenic variants in hereditary cancer predisposition genes exclusively investigated for truncating variants** | | | | | | |
| *ALK (NM_004304.4)* | c.2782dup | Exon 16 | p.Cys928LeufsTer20 | - | BRB104 | 47:0 |
| *BUB1B (NM_001211.5)* | c.2848C>T | Exon 1 | p.Gln950Ter | - | BRB261 | 38:1 |
| *FANCG (NM_004629.1)* | c.637_643del | Exon 5 | p.Tyr213LysfsTer6 | rs587776640 | BRB225<br>BRB98 | 34:4<br>43:3 |
| *RB1 (NM_000321.2)* | c.1127+1G>A | Intron | p.(?) | - | BRB73 | 29:11 |
| *XPC (NM_004628.4)* | c.2251-1G>C | Intron 13 | p.(?) | rs754673606 | BRB114<br>BRB161 | 47:1<br>29:6 |

To follow will be an in-depth look at the genes linked to the variants described in the top section of Table 3.2. The remaining variants in the second part of the table will be discussed in Chapter 4.

**ATM** (NM_000051.3)**:** c.162T>A

The ATM protein plays a role in cell cycle and regulates important proteins: NBS1, CHEK2, TP53 and BRCA1 downstream (Shiloh 2006). The nucleotide change that was identified for this variant in patient BRB14, c.162T>A changes the commonly found tyrosine amino acid to a stop codon and causes the formation of a truncated protein. The gene that is located on the long arm of chromosome 11, encodes for a protein 3056 nucleotides long, which will be transformed into a 162 bp nucleotide sequence which will not be able to function normally. The variant exists in exon 1 of patient BRB14, which leads to the loss of most of the protein.

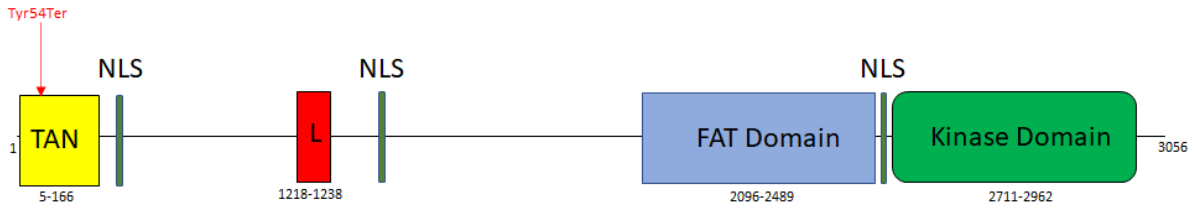


Figure 3.7: Protein structure of ATM. The variant from this study was located within the Chromatin-association domain (TAN) where chromatin or partner proteins would bind. Nuclear Localization Signal (NLS) regions have an important role in nuclear translocation and the leucine-zipper (L) has been reported to have a role in dimerization and interaction with other partners/substrates. FRAP, ATM and TRRAP (FAT) proteins also play a role in substrate binding while the kinase domain is used for phosphorylation.

This exact variant has not been published previously, but the c.162T>C variant has been identified multiple times as being a synonymous change that carries no significance (NCBI 2020). The inefficiency of the truncated protein will lead to a disrupted cell cycle which in turn will lead to decreased apoptotic responses to damaged/old cells, and DNA repair (Shiloh 2006). The identified variant has the potential to increase the risk of developing breast cancer.

**BRCA1** (NM_007294.3)**:**

Two pathogenic variants in BRCA1 were identified in two different patients. Firstly, in patient BRB130 the variant c.4524G>A gives rise to a stop codon which usually should have been a tryptophan amino acid. This variant has previously been seen by other research groups and in all cases it has been classified as being pathogenic (NCBI 2016). Patients carrying PV's in BRCA1 have an increased risk of developing breast cancer of up to 72% by the age of 80 (Kuchenbaecker, Hopper et al. 2017). This mutation was identified previously in patients from two different independent studies in the middle east (Bu, Siraj et al. 2016, Abulkhair, Al Balwi et al. 2018). All patients with this variant were diagnosed with breast cancer previously.

The second variant was c.5096G>A, transforming an arginine to a glutamate amino acid. An intensive study into a large cohort of patients carrying the variant reveal that it conveys an intermediate risk of developing breast cancer (Moghadasi, Meeks et al. 2018). Initially, this variant was deemed to have no breast cancer significance at all (Plon, Eccles et al. 2008), a subsequent study on the variant was deemed inconclusive (Lovelock, Spurdle et al. 2007). Five years later, researchers investigated the cumulative risk of developing breast cancer by age.

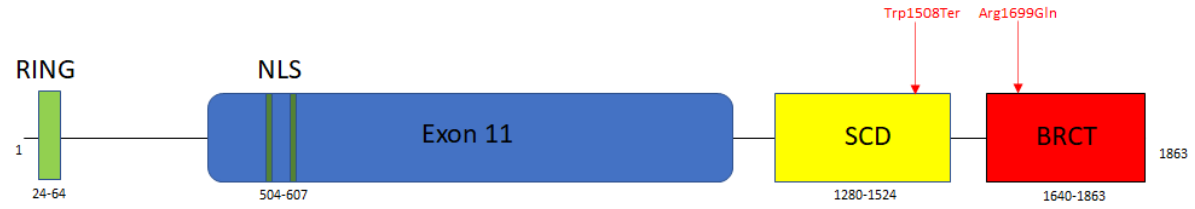

Figure 3.8: Protein structure of BRCA1. The protein contains a RING finger domain used as zinc-binding motif. Already, BARD1 and BAP1 have been identified to bind to this domain. Two nuclear localization signals (NLS) are also present within exon 11 which also harbours DNA-binding sites, a serine containing domain (SCD) and two BRCA1 C-terminus, T tower (BRCT) domains next to each other.

This was done by comparing breast cancer development through truncating BRCA1 variants, through c.5096G>A and people not carrying any specific breast cancer-related variants. This work made it evident that the c.5096G>A variant is not as dangerous as high-risk truncating variants but shows a definite increased risk compared to the general population (Spurdle, Whiley et al. 2012). In 2017, the findings of the previous paper were reiterated and they also concluded that this variant only has an intermediate effect on risk for breast cancer (Moghadasi, Meeks et al. 2018).

**BRCA2** (NM_000059.3)**:**

Another pathogenic BRCA variant was identified in this study. This variant was originally identified in a Dutch population but more recently in the Coloured and Xhosa population of the Western Cape in South Africa, no common ancestry could be established (van der Merwe, Hamel et al. 2012). Previously, classified as 'BRCA2 5999del4', it is known as the most common pathogenic variant in the Black and Coloured female populations of South Africa (Oosthuizen, Kotze et al. 2020). It is believed that this deletion leads to a truncated protein or to a degradation of a non-sense mRNA (Spugnesi, Balia et al. 2013, Oosthuizen, Kotze et al. 2020). One patient, BRB290, was found to carry this variant.

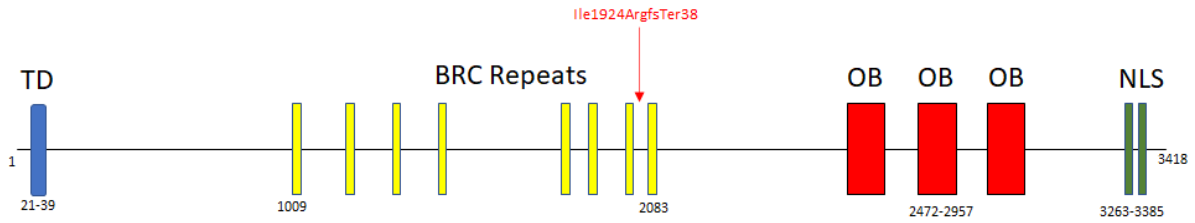


Figure 3.9: Graphical representation of the BRCA2 protein. The protein has a transactivation domain (TD) which binds transcription co-activator, like P/CAF. Eight repetitive sequence motifs (BRC), approx. 30 bases long and have some interactions with the protein, RAD51. Three oligonucleotide/oligosaccharide-binding (OB) sites and two nuclear localization signals (NLS).

The deletion occurs in the BRC domain which facilitates the binding of RAD51 and may disrupt this interaction (Spugnesi, Balia et al. 2013).

**CHEK2** (NM_001005735.1)**:**

Most breast cancer studies focus on high-risk genes like BRCA1 and BRCA2 and less on other moderate risk genes: ATM, CHEK2 and PALB2 (Kleibl and Kristensen 2016). This leads to a large gap in our understanding of moderate risk genes and the effect they may have on carriers. Studies have illuminated the fact that CHEK2 has one of the highest mutation rates after BRCA1 and BRCA in patients with Ashkenazi Jewish and European descent (Leedom, LaDuca et al. 2016, Couch, Shimelis et al. 2017, Fan, Ouyang et al. 2018, Hauke, Horvath et al. 2018).

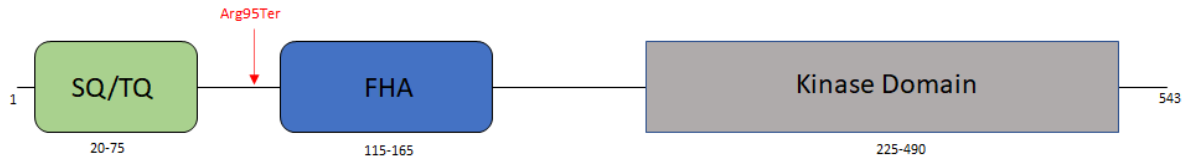


Figure 3.10: Protein structure of CHEK2. The protein contains a Ser-Gln/Thr-Gln (SQ/TQ) cluster domain which are characteristic of ATM phosphorylation sites, a forkhead-associated (FHA) domain , known for protein-protein interactions and a serine/threonine kinase domain.

Carrying this variant on the CHEK2 gene may increase the risk of breast cancer as much as three times (Walsh, Mandell et al. 2017). The variant was first identified in a study on lymphoid malignancies and (Tavor, Takeuchi et al. 2001) was later identified as a candidate breast cancer risk gene (Shaag, Walsh et al. 2005). In this study only one patient, BRB121, was identified with this variant.

**PALB2** (NM_024675.3):

This splice acceptor variant was first discovered in a 2012 (Tischkowitz, Capanu et al. 2012), where they found no clear evidence that this variant affects breast cancer risk. Studies up to this point suggested pathogenic PALB2 variants may increase risk of developing breast cancer up 2.3 times, but little is known and understood of intronic variants (Rahman, Seal et al. 2007). The variant is found in a region of PALB2 which could interact with another protein KEAP1 (Antoniou, Casadei et al. 2014), which play an important role in the sensing cellular reactive oxygen species (ROS) levels. One patient, BRB241 , was affected by this variant.

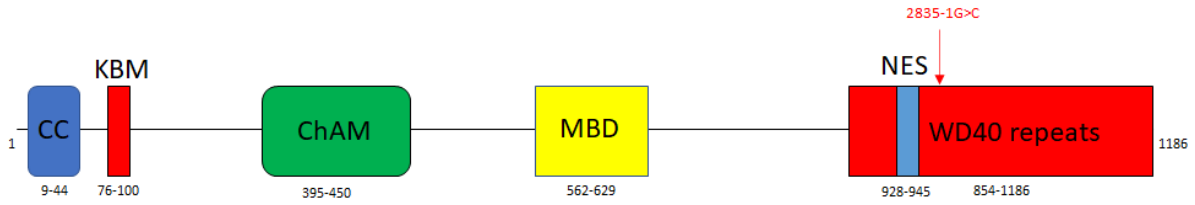


Figure 3.11: Protein structure of PALB2. The protein carries a coiled-coil domain (CC) for BRCA1 interactions, a KEAP1-binding motif (KBM). Both chromatin association motif (ChAM) and MRG15-binding domain (MBD) play roles in tethering to nucleosomes and other proteins. Seven WD40 repeats, known to promote BRCA1 interactions, on the C-terminal and located within the WD40 repeats, a nuclear export signal (NES). The variant is located in the intron in between the WD40 repeats it should be absent in the protein, the illustration is only to show where it should be located in regard to exons.

In normal circumstances, KEAP1 would bind to NRF2 to activate its degradation but PALB2 completes for this linkage and causes a disequilibrium in cellular redox homeostasis (Ma, Cai et al. 2012).

The remaining variants from the second part of Table 3.2 will be discussed in Chapter 4.

### 3.3.4 PATHWAY ENRICHMENT

*PathScore* was used to estimate pathways that were overrepresented specifically in germline mutational patients. A variety of pathways were highlighted by the program but only the top ten most enriched pathways were summarized in Figure 3.12. The top enriched pathways were dominated by repair mechanisms, eight of these were linked to some form of repair pathway.

The pathway that was mostly overrepresented was DNA repair (158 patients), which corresponds with the presence of germline mutations. Damage to DNA has long been shown to lead to cancer development. Defective DNA repair would only promote cancer.

The Fanconi anemia pathway (156 patients) and the regulation there of (156 patients) also plays an important role in DNA repair and has been linked to three classic DNA repair pathways: homologous recombination, nucleotide excision repair, and mutagenic translesion synthesis.

Other than repair pathways, meiotic cell division was the most affected pathway, meiosis (70 patients) and meiotic recombination (70 patients). The overexpression of meiosis genes gives rise to genomic instability which ultimately leads to cancer development.

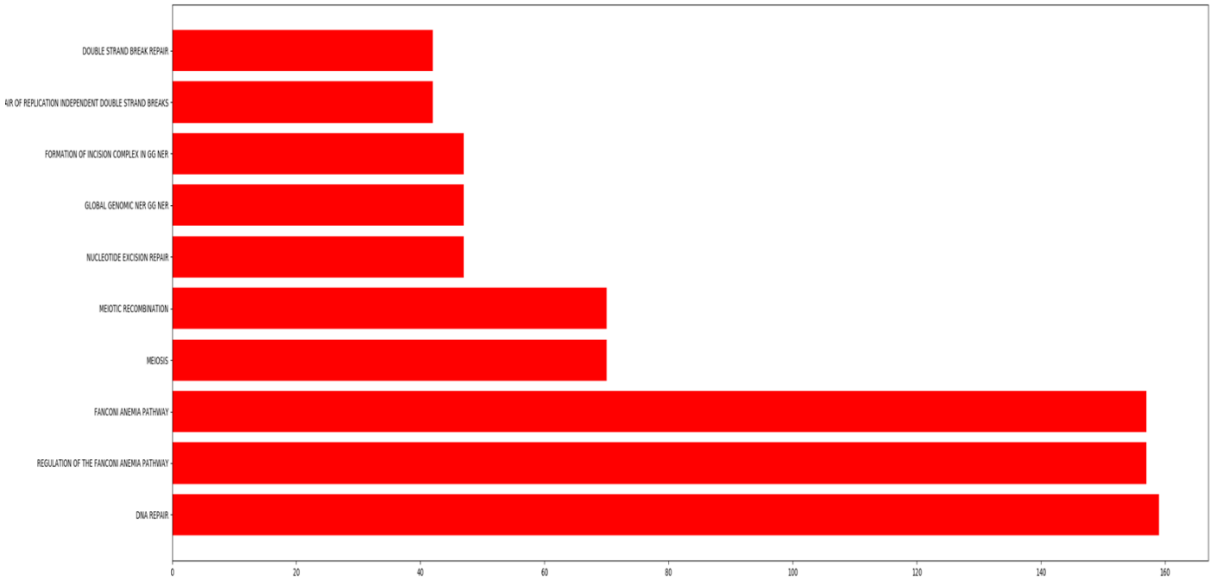


Figure 3.12: Top ten most affected pathways as described by *Pathscore*. Pathways which were most active in the patients from this study were listed. The figure represents the pathway with the total number of patients where it was overrepresented. GG-NER – Global genome nucleotide excision repair

### 3.3.5 COMPARISON BETWEEN GATK 3.8 AND GATK 4.0

Lastly, the analysis initially done in GATK 3.8 was redone in GATK 4.0 platform to draw a comparison between their agreement, as the variant detection was initially done with GATK 3.8, but was subsequently redone with GATK 4.0 (Figure 3.13).

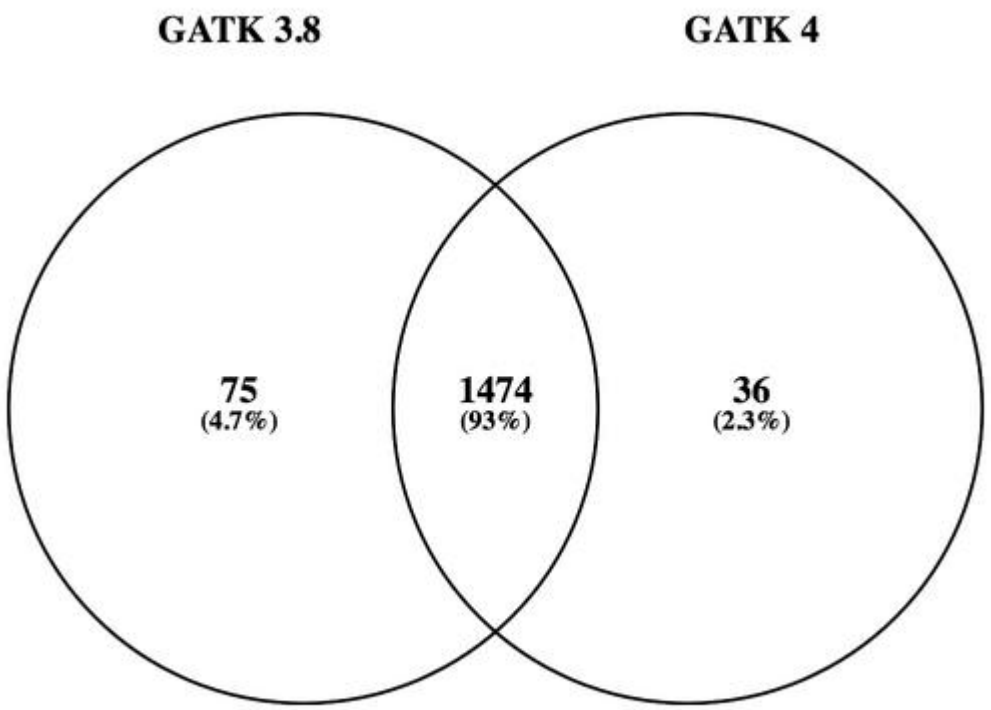


Figure 3.13: Comparison between GATK 3.8 and GATK 4.0.

The different analysis resulted in an agreement of 93% of the calls made. GATK 4.0 at the end of the analysis removed more samples that it considered not to be deleterious (Figure 3.13). Updated repositories of variants may have led to better analysis of these calls resulting in a smaller more accurate set. Our analysis was finally done using GATK 4.0.

## 3.4 DISCUSSION

This study screened 165 South African breast cancer patients of African ancestry (self-identified) for the presence of deleterious germline sequence variants in 94 genes associated with hereditary cancer. The patients were unselected for age at diagnosis or family history of cancer. With the exception of four cases (BRB130, BRB290, BRC134 and BRC210) all others were previously screened for BRCA1/BRCA2 variants using non-NGS methods and found to be negative for pathogenic/likely pathogenic variants.

Although the patients were unselected for family history of breast or ovarian cancer, 9% did report some family history of breast or ovarian cancer. This is higher than that reported for similar studies in breast cancer patients from Cameroon/Uganda (6.6%) and Nigeria (6%) (Zheng, Walsh et al. 2018, Adedokun, Zheng et al. 2020). With regards to tumour stage, 70.3% of patients were diagnosed with stage III/IV at diagnosis. It is thought that low survival rates in sub-Saharan Africa is mostly attributable to late-stage presentation. The stage at presentation of our cohort is similar to that reported in 83 studies across 17 sub-Saharan African countries, with 77% of cases presenting at stage III/IV (Jedy-Agba, McCormack et al. 2016).

We identified pathogenic/likely pathogenic variants (P/LP) in 13 patients, in ten different genes (Table 3.2), which represents 7.9% of the cohort. Six of these patients (3.6%) have PV/LPVs in genes that are confirmed to confer an increased risk for breast cancer. The mean age of patients who carried deleterious variant in BRCA1/BRCA2 was 39 years and 8 months compared to 47 years and 3 months among women who carried a deleterious variant in other breast cancer susceptibility genes. Pathogenic variants in non-BRCA1/BRCA2 breast cancer susceptibility genes accounted for 1.8% of our cohort. None of these women reported any family history of cancer. In addition, 14 benign/likely benign variants were detected in eight breast cancer genes (Supplementary Table 3.1). This includes six variants not previously described, detected in 12 established and candidate breast cancer genes. In the studied cohort, variants in the ATM gene were the most frequently identified (Table 3.2).

Pathogenic ATM variants act in a recessive manner to cause Ataxia telangiectasia (a neurodegenerative disease), whereas heterozygous carriers are at moderately increased risk for breast cancer (Renwick, Thompson et al. 2006, Marabelli, Cheng et al. 2016). Patient BRB14 (Zulu-speaking patient), diagnosed with breast cancer at age 47 years and 8 months, was a carrier of the novel ATM likely pathogenic variant, c.162T > A. It is predicted to be a nonsense variant, p.(Tyr54Ter), that may cause the transcript to be exposed to nonsense-mediated mRNA decay. If ATM is synthesized it will lack most of the protein sequence and thus be non-functional. Interestingly, a recent study that explored the clinico-pathological characteristics of breast cancers developed by ATM mutation carriers reported the median age at first diagnosis to be 46.9 years in their cohort (Toss, Tenedini et al. 2021). Unfortunately, we do not have any further histopathologic information on the breast cancer of BRB14. There has been some debate on whether mono-allelic truncating ATM variants are associated with increased breast cancer risk. Early on it was hypothesised that some missense variants in ATM might have dominant negative effects and confer a particularly high risk of breast cancer when heterozygous, compared to truncating variants (Gatti, Tward et al. 1999).

In a meta-analysis of ATM variants, a later study found strong evidence that a subset of rare evolutionary unlikely missense variants confer increased cancer risk. They found marginal evidence that protein truncating and splice-site variants contribute to breast cancer risk (Tavtigian, Oefner et al. 2009). Goldgar *et al*. further investigated the issue and reported risk estimates that women who carry either a pathogenic missense or truncating variant have a significantly increased risk of breast cancer (Goldgar, Healey et al. 2011). Obtaining accurate risk estimates require a large sample size, which a recent large study of more than 113,000 women (mostly population-based samples), addressed (Dorling, Carvalho et al. 2021). This study identified ATM protein-truncating variants to confer significant disease risks (odds ratio 2.1), compared to rare missense variants (odds ratio 1.06) (Dorling, Carvalho et al. 2021).

Two of the four patients (BRB130 and BRB290) who had not previously been screened for BRCA1/BRCA2 variants, were found to carry a BRCA1 or BRCA2 deleterious variant (Table 3.2). The BRCA1 c.4524G > A p.(Trp1508Ter) variant was identified in BRB130, a Tswana-speaking woman diagnosed with breast cancer at age 45 years and 8 months. The variant is predicted to introduce a stop codon that will produce a transcript that may be targeted for nonsense-mediated mRNA decay (NMD). This nonsense variant has been detected in multiple families with hereditary breast ovarian cancers (Loman, Johannsson et al. 2001, Laitman, Borsthein et al. 2011, Walsh, Casadei et al. 2011, Kang, Seong et al. 2015, Lynce, Smith et al. 2015, Bu, Siraj et al. 2016, Plaskocinska, Shipman et al. 2016, Briceño-Balcázar, Gómez-Gutiérrez et al. 2017). Of note, the variant is also designated as 4643G > A in published literature.

BRB264 (diagnosed at 42 years and three months, Tsonga-speaking patient) carried the BRCA1 c.5096G > A p.(Arg1699Gln), intermediate risk variant. It is in the BRCA1 carboxyl terminal region of the transcriptional transactivation domain. The cancer risks associated with this variant was first defined by the ENIGMA consortium (Evidence-based Network for the Interpretation of Germline Mutant Alleles) in 2012 and in a follow up study in 2017 the risk estimates were confirmed (Spurdle, Whiley et al. 2012, Moghadasi, Meeks et al. 2018). Functional assays showed this variant to have impaired homology-directed DNA repair activity and it was classified as being a hypomorphic allele (Petitalot, Dardillac et al. 2019). Interestingly, this pathogenic missense was also found in a Nigerian woman with breast cancer (Zheng, Walsh et al. 2018).

The BRCA2 frameshift variant, c.5771_5774del p.(Ile1924ArgfsTer38), was identified in BRB290 who was diagnosed with breast cancer at 26 years and 6 months of age. The variant is expected to result in loss of function due to an absent or disrupted protein. This alteration has been reported in multiple individuals (of European ancestry) with hereditary breast and ovarian cancer syndrome

(NHGRI 2019) and has been reported as a founder mutation in Bantu-speaking Xhosa women from the Western Cape of South Africa (van der Merwe, Hamel et al. 2012). BRB290 is however a Bantu-speaking Sotho individual, and at this time it is not possible to do any haplotype analysis to ascertain whether she carries this PV on the same haplotype as that of the Xhosa founder variant.

The pathogenic CHEK2, c.283C > T p.(Arg95Ter), variant detected in BRB121 (diagnosed at 54 years, Zulu speaking patient) was previously identified in the germline of two Norwegian patients diagnosed with locally advanced breast cancer (Chrisanthar, Knappskog et al. 2008). Of interest, both patients were resistant to anthracycline therapy. In vitro assays of the p.(Arg95Ter) variant found the CHEK2 protein to be non-functional in terms of kinase activity and dimerization. Loss of heterogeneity (LOH) analysis of the tumours found that the wild type allele of the CHEK2 gene was lost for both of the patients (Chrisanthar, Knappskog et al. 2008). The possibility that this nonsense variant together with LOH is associated with resistance to anthracyclines in cancer patients underlines its potential clinical importance. In a follow up case control study of 7 081 incident cancer cases from Norway, Knappskog et al. (2016), detected the p.(Arg95Ter) variant in 0.23% breast cancer cases and in 0.16% prostate cancer cases (Knappskog, Leirvaag et al. 2016). This variant is also reported as pathogenic by multiple laboratories in ClinVar (Variation ID: 140772). In our study 0.61% (1/165) of cases carried a pathogenic CHEK2 variant. There is substantial variation in the prevalence of germline CHEK2 pathogenic variants among different populations and ethnicities, with individuals of European ancestry that have the highest prevalence (Stolarova, Kleiblova et al. 2020). A multi-ethnic population-based study of a cohort of breast cancer and ovarian cancer patients found that for breast cancer 2.3% (95% CI 1.8% to 2.8%) of white individuals and only 0.15% (95% CI 0% to 0.82%) of black individuals carried a pathogenic CHEK2 variant (Kurian, Ward et al. 2019).

The PALB2 variant, c.2835-1G > C, located in a canonical acceptor splice-site (in Intron 8) was identified in a Xhosa-speaking patient (BRB241, diagnosed at 40 years of age). The variant has been reported in the literature in persons affected with breast or ovarian cancer (Tischkowitz, Capanu et al. 2012, Antoniou, Casadei et al. 2014, Norquist, Harrell et al. 2016, Eliade, Skrzypski et al. 2017). Several in silico bioinformatic tools predicted this variant to abolish the 3'-acceptor splice site, which would alter the natural splicing of PALB2. The expected effect is an in-frame deletion in the PALB2 mRNA by skipping exon nine (deletion of 162 bp, 54 amino acids: Ala946 to Gly999). Another possibility is that an alternative cryptic splice site could be used. The strongest alternative site is in exon nine at c.2864, and should this be used, the result would be the loss of 30 bp (10 amino acids: Ala946 to Glu955) from exon nine.

cBROCA analysis of mRNA from patients with the c.2835-1G > C variant showed that it preferentially leads to skipping of exon 9 (r.2835–2996) and is therefore expected to produce an abnormal PALB2 protein, lacking the 54 amino acids (Casadei, Gulsuner et al. 2019). The deleted section is part of the second and third blades of the WD40 domain of PALB2. This seven bladed region is essential for the interaction of BRCA2 with PALB2 (Xia, Sheng et al. 2006, Oliver, Swift et al. 2009). When BRCA2 is unable to bind to PALB2, homologous recombination repair is severely disrupted.

A limitation of this study is that no copy number variation using NGS data or MLPA was used to investigate the genes. Large deletions or duplications could be undetected. Furthermore, the relatively small sample size and unavailability of hormone receptor status precluded any investigation of the prevalence of sequence variants by breast cancer subtype. While precision medicine is currently still mostly out of reach in African countries due to economic reasons, the rapidly declining costs of genomic technologies will in future necessitate population-specific variant information, particularly in diseases such as cancer.

To our knowledge, this is the first study that has investigated South African breast cancer patients of African ancestry for germline sequence variants in a multigene panel. Although we investigated a relatively small cohort of patients, our study provides some insights towards the genetic breast cancer risk factors in South African women of African ancestry. In conclusion, our study has shown that the 3.6% of women who carry a pathogenic/likely pathogenic variant in a breast cancer susceptibility gene do not necessarily have a family history of breast cancer. In our cohort there was an equal proportion of women who carried a deleterious variant in BRCA1/BRCA2 (1.8%) and women who carried a deleterious variant in other breast cancer susceptibility genes (1.8%). These findings must however be treated with caution because of the small sample size. Further studies of a larger patient cohort is warranted to assess the distribution of variants in clinically relevant cancer susceptibility genes.

## 3.5 SUPPLEMENTARY TABLES

Supplementary Table 3.1: Benign/Likely benign variants detected.

| Gene (RefSeq) | Variant | Predicted protein change | dbSNP | Patient |
|---|---|---|---|---|
| *ATM* (NM_000051.3) | c.334G>A | NP_000042.3:p.Ala112Thr | rs146382972 | BRB171 |
| | | | | BRB68 |
| *ATM* | c.2096A>G | NP_000042.3:p.Glu699Gly | rs147934285 | BRB142 |
| | | | | BRB190 |
| | | | | BRB91 |
| *ATM* | c.7313C>T | NP_000042.3:p.Thr2438Ile | rs147604227 | BRB171 |
| *BRCA1* (NM_007294.3) | c.4682C>T | NP_009225.1:p.Thr1561Ile | rs56158747 | BRB143 |
| | | | | BRB146 |
| | | | | BRB28 |
| | | | | BRB42 |
| *BRCA2* (NM_000059.3) | c.3858_3860del | NP_000050.2:p.Lys1286del | rs80359406 | BRB59 |
| | | | | BRB9 |
| *BRCA2* | c.9875C>T | NP_000050.2:p.Pro3292Leu | rs56121817 | BRB160 |
| | | | | BRB99 |
| *CHEK2* (NM_001005735.1) | c.254C>T | NP_001005735.1:p.Pro85Leu | rs17883862 | BRB172 |
| | | | | BRB224 |
| | | | | BRB52 |
| | | | | BRB62 |
| | | | | BRB88 |
| | | | | BRC134 |
| *MSH6* (NM_000179.2) | c.3911G>A | NP_000170.1:p.Arg1304Lys | rs34625968 | BRB87 |
| *NF1* (NM_001042492.2) | c.3169G>A | NP_001035957.1:p.Ala1057Thr | rs1367746167 | BRB108 |
| *NF1* | c.7539G>C | NP_001035957.1:p.Gln2513His | rs2070170345 | BRB108 |
| *PMS2* (NM_000535.6) | c.1268C>T | NP_000526.2:p.Ala423Val | rs756883400 | BRB19 |
| *PMS2* | c.612T>A | NP_000526.2:p.Asn204Lys | - | BRC210 |
| *PMS2* | c.497T>C | NP_000526.2:p.Leu166Pro | rs116349687 | BRB114 |
| | | | | BRB142 |
| | | | | BRB225 |
| | | | | BRB264 |
| | | | | BRB265 |
| *RAD51D* (NM_002878.3) | c.146C>T | NP_002869.3:p.Ala49Val | rs140317560 | BRB264 |

## 3.6 REFERENCES

Abulkhair, O., M. Al Balwi, O. Makram, L. Alsubaie, M. Faris, H. Shehata, A. Hashim, B. Arun, A. Saadeddin and E. Ibrahim (2018). "Prevalence of BRCA1 and BRCA2 Mutations Among High-Risk Saudi Patients With Breast Cancer." J Glob Oncol **4**: 1-9.

Adedokun, B., Y. Zheng, P. Ndom, A. Gakwaya, T. Makumbi, A. Y. Zhou, T. F. Yoshimatsu, A. Rodriguez, R. K. Madduri, I. T. Foster, A. Sallam, O. I. Olopade and D. Huo (2020). "Prevalence of Inherited Mutations in Breast Cancer Predisposition Genes among Women in Uganda and Cameroon." Cancer Epidemiol Biomarkers Prev **29**(2): 359-367.

Antoniou, A. C., S. Casadei, T. Heikkinen, D. Barrowdale, K. Pylkäs, J. Roberts, A. Lee, D. Subramanian, K. De Leeneer, F. Fostira, E. Tomiak, S. L. Neuhausen, Z. L. Teo, S. Khan, K. Aittomäki, J. S. Moilanen, C. Turnbull, S. Seal, A. Mannermaa, A. Kallioniemi, G. J. Lindeman, S. S. Buys, I. L. Andrulis, P. Radice, C. Tondini, S. Manoukian, A. E. Toland, P. Miron, J. N. Weitzel, S. M. Domchek, B. Poppe, K. B. Claes, D. Yannoukakos, P. Concannon, J. L. Bernstein, P. A. James, D. F. Easton, D. E. Goldgar, J. L. Hopper, N. Rahman, P. Peterlongo, H. Nevanlinna, M. C. King, F. J. Couch, M. C. Southey, R. Winqvist, W. D. Foulkes and M. Tischkowitz (2014). "Breast-cancer risk in families with mutations in PALB2." N Engl J Med **371**(6): 497-506.

Briceño-Balcázar, I., A. Gómez-Gutiérrez, N. A. Díaz-Dussán, M. C. Noguera-Santamaría, D. Díaz-Rincón and M. C. Casas-Gómez (2017). "Mutational spectrum in breast cancer associated BRCA1 and BRCA2 genes in Colombia." Colomb Med (Cali) **48**(2): 58-63.

Bu, R., A. K. Siraj, K. A. S. Al-Obaisi, S. Beg, M. Al Hazmi, D. Ajarim, A. Tulbah, F. Al-Dayel and K. S. Al-Kuraya (2016). "Identification of novel BRCA founder mutations in Middle Eastern breast cancer patients using capture and Sanger sequencing analysis." International journal of cancer **139**(5): 1091-1097.

Casadei, S., S. Gulsuner, B. H. Shirts, J. B. Mandell, H. M. Kortbawi, B. S. Norquist, E. M. Swisher, M. K. Lee, Y. Goldberg, R. O'Connor, Z. Tan, C. C. Pritchard, M. C. King and T. Walsh (2019). "Characterization of splice-altering mutations in inherited predisposition to cancer." Proc Natl Acad Sci U S A **116**(52): 26798-26807.

Chrisanthar, R., S. Knappskog, E. Løkkevik, G. Anker, B. Østenstad, S. Lundgren, E. O. Berge, T. Risberg, I. Mjaaland, L. Maehle, L. F. Engebretsen, J. R. Lillehaug and P. E. Lønning (2008). "CHEK2 mutations affecting kinase activity together with mutations in TP53 indicate a functional pathway associated with resistance to epirubicin in primary breast cancer." PLoS One **3**(8): e3062.

Couch, F. J., H. Shimelis, C. Hu, S. N. Hart, E. C. Polley, J. Na, E. J. Hallberg, R. Moore, A. Thomas, J. Lilyquist, B. Feng, R. McFarland, T. Pesaran, R. Huether, H. LaDuca, E. C. Chao, D. E. Goldgar and J. S. Dolinsky (2017). "Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer." JAMA oncology **3**(9): 1190-1196.

Dorling, L., S. Carvalho, J. Allen, A. González-Neira, C. Luccarini, C. Wahlström, K. A. Pooley, M. T. Parsons, C. Fortuno, Q. Wang, M. K. Bolla, J. Dennis, R. Keeman, M. R. Alonso, N. Álvarez, B. Herraez, V. Fernandez, R. Núñez-Torres, A. Osorio, J. Valcich, M. Li, T. Törngren, P. A. Harrington, C. Baynes, D. M. Conroy, B. Decker, L. Fachal, N. Mavaddat, T. Ahearn, K. Aittomäki, N. N. Antonenkova, N. Arnold, P. Arveux, M. Ausems, P. Auvinen, H. Becher, M. W. Beckmann, S. Behrens, M. Bermisheva, K. Białkowska, C. Blomqvist, N. V. Bogdanova, N. Bogdanova-Markov, S. E. Bojesen, B. Bonanni, A. L. Børresen-Dale, H. Brauch, M. Bremer, I. Briceno, T. Brüning, B. Burwinkel, D. A. Cameron, N. J. Camp, A. Campbell, A. Carracedo, J. E. Castelao, M. H. Cessna, S. J. Chanock, H. Christiansen, J. M. Collée, E. Cordina-Duverger, S. Cornelissen, K. Czene, T. Dörk, A. B. Ekici, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, H. Flyger, A. Försti, M. Gabrielson, M. Gago-Dominguez, V. Georgoulias, F. Gil, G. G. Giles, G. Glendon, E. B. G. Garcia, G. I. G. Alnæs, P. Guénel, A. Hadjisavvas, L. Haeberle, E. Hahnen, P. Hall, U. Hamann, E. F. Harkness, J. M. Hartikainen, M. Hartman, W. He, B. A. M. Heemskerk-Gerritsen, P. Hillemanns, F. B. L. Hogervorst, A. Hollestelle, W. K. Ho, M. J. Hooning, A. Howell, K. Humphreys, F. Idris, A. Jakubowska, A. Jung, P. M. Kapoor, M. J. Kerin, E. Khusnutdinova, S. W. Kim, Y. D. Ko, V. M. Kosma, V. N. Kristensen, K. Kyriacou, I. M. M. Lakeman, J. W. Lee, M. H. Lee, J. Li, A. Lindblom, W. Y.

Lo, M. A. Loizidou, A. Lophatananon, J. Lubiński, R. J. MacInnis, M. J. Madsen, A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M. E. Martinez, T. Maurer, D. Mavroudis, C. McLean, A. Meindl, A. R. Mensenkamp, K. Michailidou, N. Miller, N. A. Mohd Taib, K. Muir, A. M. Mulligan, H. Nevanlinna, W. G. Newman, B. G. Nordestgaard, P. S. Ng, J. C. Oosterwijk, S. K. Park, T. W. Park-Simon, J. I. A. Perez, P. Peterlongo, D. J. Porteous, K. Prajzendanc, D. Prokofyeva, P. Radice, M. U. Rashid, V. Rhenius, M. A. Rookus, T. Rüdiger, E. Saloustros, E. J. Sawyer, R. K. Schmutzler, A. Schneeweiss, P. Schürmann, M. Shah, C. Sohn, M. C. Southey, H. Surowy, M. Suvanto, S. Thanasitthichai, I. Tomlinson, D. Torres, T. Truong, M. Tzardi, Y. Valova, C. J. van Asperen, R. M. Van Dam, A. M. W. van den Ouweland, L. E. van der Kolk, E. M. van Veen, C. Wendt, J. A. Williams, X. R. Yang, S. Y. Yoon, M. P. Zamora, D. G. Evans, M. de la Hoya, J. Simard, A. C. Antoniou, Å. Borg, I. L. Andrulis, J. Chang-Claude, M. García-Closas, G. Chenevix-Trench, R. L. Milne, P. D. P. Pharoah, M. K. Schmidt, A. B. Spurdle, M. P. G. Vreeswijk, J. Benitez, A. M. Dunning, A. Kvist, S. H. Teo, P. Devilee and D. F. Easton (2021). "Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women." N Engl J Med **384**(5): 428-439.

Eliade, M., J. Skrzypski, A. Baurand, C. Jacquot, G. Bertolone, C. Loustalot, C. Coutant, F. Guy, P. Fumoleau, Y. Duffourd, L. Arnould, A. Delignette, M. M. Padéano, C. Lepage, G. Raichon-Patru, A. Boudrant, M. C. Bône-Lépinoy, A. L. Villing, A. Charpin, K. Peignaux, S. Chevrier, F. Vegran, F. Ghiringhelli, R. Boidot, N. Sevenet, S. Lizard and L. Faivre (2017). "The transfer of multigene panel testing for hereditary breast and ovarian cancer to healthcare: What are the implications for the management of patients and families?" Oncotarget **8**(2): 1957-1971.

Fan, Z., T. Ouyang, J. Li, T. Wang, Z. Fan, T. Fan, B. Lin, Y. Xu and Y. Xie (2018). "Identification and analysis of CHEK2 germline mutations in Chinese BRCA1/2-negative breast cancer patients." Breast Cancer Res Treat **169**(1): 59-67.

Gatti, R. A., A. Tward and P. Concannon (1999). "Cancer risk in ATM heterozygotes: a model of phenotypic and mechanistic differences between missense and truncating mutations." Mol Genet Metab **68**(4): 419-423.

Goldgar, D. E., S. Healey, J. G. Dowty, L. Da Silva, X. Chen, A. B. Spurdle, M. B. Terry, M. J. Daly, S. M. Buys, M. C. Southey, I. Andrulis, E. M. John, K. K. Khanna, J. L. Hopper, P. J. Oefner, S. Lakhani and G. Chenevix-Trench (2011). "Rare variants in the ATM gene and risk of breast cancer." Breast Cancer Res **13**(4): R73.

Hauke, J., J. Horvath, E. Groß, A. Gehrig, E. Honisch, K. Hackmann, G. Schmidt, N. Arnold, U. Faust, C. Sutter, J. Hentschel, S. Wang-Gohrke, M. Smogavec, B. H. F. Weber, N. Weber-Lassalle, K. Weber-Lassalle, J. Borde, C. Ernst, J. Altmüller, A. E. Volk, H. Thiele, V. Hübbel, P. Nürnberg, K. Keupp, B. Versmold, E. Pohl, C. Kubisch, S. Grill, V. Paul, N. Herold, N. Lichey, K. Rhiem, N. Ditsch, C. Ruckert, B. Wappenschmidt, B. Auber, A. Rump, D. Niederacher, T. Haaf, J. Ramser, B. Dworniczak, C. Engel, A. Meindl, R. K. Schmutzler and E. Hahnen (2018). "Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German Consortium for Hereditary Breast and Ovarian Cancer." Cancer Med **7**(4): 1349-1358.

Jedy-Agba, E., V. McCormack, C. Adebamowo and I. Dos-Santos-Silva (2016). "Stage at diagnosis of breast cancer in sub-Saharan Africa: a systematic review and meta-analysis." Lancet Glob Health **4**(12): e923-e935.

Kang, E., M. W. Seong, S. K. Park, J. W. Lee, J. Lee, L. S. Kim, J. E. Lee, S. Y. Kim, J. Jeong, S. A. Han and S. W. Kim (2015). "The prevalence and spectrum of BRCA1 and BRCA2 mutations in Korean population: recent update of the Korean Hereditary Breast Cancer (KOHBRA) study." Breast Cancer Res Treat **151**(1): 157-168.

Kleibl, Z. and V. N. Kristensen (2016). "Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management." Breast **28**: 136-144.

Knappskog, S., B. Leirvaag, L. B. Gansmo, P. Romundstad, K. Hveem, L. Vatten and P. E. Lønning (2016). "Prevalence of the CHEK2 R95* germline mutation." Hered Cancer Clin Pract **14**: 19.

Kuchenbaecker, K. B., J. L. Hopper, D. R. Barnes, K. Phillips, T. M. Mooij, M. Roos-Blom, S. Jervis, F. E. van Leeuwen, R. L. Milne, N. Andrieu, D. E. Goldgar, M. Terry, M. A. Rookus, D. F. Easton, A. C.

Antoniou, a. t. BRCA1 and B. C. Consortium (2017). "Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers." JAMA **317**(23): 2402-2416.
Kurian, A. W., K. C. Ward, N. Howlader, D. Deapen, A. S. Hamilton, A. Mariotto, D. Miller, L. S. Penberthy and S. J. Katz (2019). "Genetic Testing and Results in a Population-Based Cohort of Breast Cancer Patients and Ovarian Cancer Patients." J Clin Oncol **37**(15): 1305-1315.
Laitman, Y., R. T. Borsthein, D. Stoppa-Lyonnet, E. Dagan, L. Castera, M. Goislard, R. Gershoni-Baruch, H. Goldberg, B. Kaufman, N. Ben-Baruch, J. Zidan, T. Maray, L. Soussan-Gutman and E. Friedman (2011). "Germline mutations in BRCA1 and BRCA2 genes in ethnically diverse high risk families in Israel." Breast Cancer Res Treat **127**(2): 489-495.
Leedom, T. P., H. LaDuca, R. McFarland, S. Li, J. S. Dolinsky and E. C. Chao (2016). "Breast cancer risk is similar for CHEK2 founder and non-founder mutation carriers." Cancer Genet **209**(9): 403-407.
Loman, N., O. Johannsson, U. Kristoffersson, H. Olsson and A. Borg (2001). "Family history of breast and ovarian cancers and BRCA1 and BRCA2 mutations in a population-based series of early-onset breast cancer." J Natl Cancer Inst **93**(16): 1215-1223.
Lovelock, P. K., A. B. Spurdle, M. T. Mok, D. J. Farrugia, S. R. Lakhani, S. Healey, S. Arnold, D. Buchanan, F. J. Couch, B. R. Henderson, D. E. Goldgar, S. V. Tavtigian, G. Chenevix-Trench and M. A. Brown (2007). "Identification of BRCA1 missense substitutions that confer partial functional activity: potential moderate risk variants?" Breast Cancer Res **9**(6): R82.
Lynce, F., K. L. Smith, J. Stein, T. DeMarco, Y. Wang, H. Wang, M. Fries, B. N. Peshkin and C. Isaacs (2015). "Deleterious BRCA1/2 mutations in an urban population of Black women." Breast Cancer Res Treat **153**(1): 201-209.
Ma, J., H. Cai, T. Wu, B. Sobhian, Y. Huo, A. Alcivar, M. Mehta, K. L. Cheung, S. Ganesan, A. N. Kong, D. D. Zhang and B. Xia (2012). "PALB2 interacts with KEAP1 to promote NRF2 nuclear accumulation and function." Mol Cell Biol **32**(8): 1506-1517.
Marabelli, M., S. C. Cheng and G. Parmigiani (2016). "Penetrance of ATM Gene Mutations in Breast Cancer: A Meta-Analysis of Different Measures of Risk." Genet Epidemiol **40**(5): 425-431.
Moghadasi, S., H. D. Meeks, M. P. Vreeswijk, L. A. Janssen, Å. Borg, H. Ehrencrona, Y. Paulsson-Karlsson, B. Wappenschmidt, C. Engel, A. Gehrig, N. Arnold, T. V. O. Hansen, M. Thomassen, U. B. Jensen, T. A. Kruse, B. Ejlertsen, A. M. Gerdes, I. S. Pedersen, S. M. Caputo, F. Couch, E. J. Hallberg, A. M. van den Ouweland, M. J. Collée, E. Teugels, M. A. Adank, R. B. van der Luijt, A. R. Mensenkamp, J. C. Oosterwijk, M. J. Blok, N. Janin, K. B. Claes, K. Tucker, V. Viassolo, A. E. Toland, D. E. Eccles, P. Devilee, C. J. Van Asperen, A. B. Spurdle, D. E. Goldgar and E. G. García (2018). "The BRCA1 c. 5096G>A p.Arg1699Gln (R1699Q) intermediate risk variant: breast and ovarian cancer risk estimation and recommendations for clinical management from the ENIGMA consortium." J Med Genet **55**(1): 15-20.
NCBI. (2016). "Occurence of BRCA1 c.4524G>A variant." Retrieved March 13, 2021, from https://www.ncbi.nlm.nih.gov/clinvar/variation/VCV000055221.12
NCBI. (2020). "Occurence of ATM 162T>C variant." Retrieved March 13, 2021, from https://www.ncbi.nlm.nih.gov/clinvar/variation/VCV000132757.13.
NHGRI. (2019). "Breast Cancer Information Core." from https://research.nhgri.nih.gov/bic/.
Norquist, B. M., M. I. Harrell, M. F. Brady, T. Walsh, M. K. Lee, S. Gulsuner, S. S. Bernards, S. Casadei, Q. Yi, R. A. Burger, J. K. Chan, S. A. Davidson, R. S. Mannel, P. A. DiSilvestro, H. A. Lankes, N. C. Ramirez, M. C. King, E. M. Swisher and M. J. Birrer (2016). "Inherited Mutations in Women With Ovarian Carcinoma." JAMA Oncol **2**(4): 482-490.
Oliver, A. W., S. Swift, C. J. Lord, A. Ashworth and L. H. Pearl (2009). "Structural basis for recruitment of BRCA2 by PALB2." EMBO Rep **10**(9): 990-996.
Oosthuizen, J., M. J. Kotze, N. Van Der Merwe, E. J. Myburgh, P. Bester and N. C. van der Merwe (2020). "Globally Rare BRCA2 Variants With Founder Haplotypes in the South African Population: Implications for Point-of-Care Testing Based on a Single-Institution BRCA1/2 Next-Generation Sequencing Study." Front Oncol **10**: 619469.

Petitalot, A., E. Dardillac, E. Jacquet, N. Nhiri, J. Guirouilh-Barbat, P. Julien, I. Bouazzaoui, D. Bonte, J. Feunteun, J. A. Schnell, P. Lafitte, J. C. Aude, C. Noguès, E. Rouleau, R. Lidereau, B. S. Lopez, S. Zinn-Justin and S. M. Caputo (2019). "Combining Homologous Recombination and Phosphopeptide-binding Data to Predict the Impact of BRCA1 BRCT Variants on Cancer Risk." Mol Cancer Res **17**(1): 54-69.

Plaskocinska, I., H. Shipman, J. Drummond, E. Thompson, V. Buchanan, B. Newcombe, C. Hodgkin, E. Barter, P. Ridley, R. Ng, S. Miller, A. Dann, V. Licence, H. Webb, L. T. Tan, M. Daly, S. Ayers, B. Rufford, H. Earl, C. Parkinson, T. Duncan, M. Jimenez-Linan, G. S. Sagoo, S. Abbs, N. Hulbert-Williams, P. Pharoah, R. Crawford, J. D. Brenton and M. Tischkowitz (2016). "New paradigms for BRCA1/BRCA2 testing in women with ovarian cancer: results of the Genetic Testing in Epithelial Ovarian Cancer (GTEOC) study." J Med Genet **53**(10): 655-661.

Plon, S. E., D. M. Eccles, D. Easton, W. D. Foulkes, M. Genuardi, M. S. Greenblatt, F. B. L. Hogervorst, N. Hoogerbrugge, A. B. Spurdle, S. V. Tavtigian and I. U. G. V. W. Group (2008). "Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results." Human mutation **29**(11): 1282-1291.

Rahman, N., S. Seal, D. Thompson, P. Kelly, A. Renwick, A. Elliott, S. Reid, K. Spanova, R. Barfoot, T. Chagtai, H. Jayatilake, L. McGuffog, S. Hanks, D. G. Evans, D. Eccles, D. F. Easton and M. R. Stratton (2007). "PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene." Nat Genet **39**(2): 165-167.

Renwick, A., D. Thompson, S. Seal, P. Kelly, T. Chagtai, M. Ahmed, B. North, H. Jayatilake, R. Barfoot, K. Spanova, L. McGuffog, D. G. Evans, D. Eccles, D. F. Easton, M. R. Stratton and N. Rahman (2006). "ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles." Nat Genet **38**(8): 873-875.

Shaag, A., T. Walsh, P. Renbaum, T. Kirchhoff, K. Nafa, S. Shiovitz, J. B. Mandell, P. Welcsh, M. K. Lee, N. Ellis, K. Offit, E. Levy-Lahad and M. C. King (2005). "Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population." Hum Mol Genet **14**(4): 555-563.

Shiloh, Y. (2006). "The ATM-mediated DNA-damage response: taking shape." Trends Biochem Sci **31**(7): 402-410.

Spugnesi, L., C. Balia, A. Collavoli, E. Falaschi, V. Quercioli, M. A. Caligo and A. Galli (2013). "Effect of the expression of BRCA2 on spontaneous homologous recombination and DNA damage-induced nuclear foci in Saccharomyces cerevisiae." Mutagenesis **28**(2): 187-195.

Spurdle, A. B., P. J. Whiley, B. Thompson, B. Feng, S. Healey, M. A. Brown, C. Pettigrew, C. J. Van Asperen, M. G. Ausems, A. A. Kattentidt-Mouravieva, A. M. van den Ouweland, A. Lindblom, M. H. Pigg, R. K. Schmutzler, C. Engel, A. Meindl, S. Caputo, O. M. Sinilnikova, R. Lidereau, F. J. Couch, L. Guidugli, T. Hansen, M. Thomassen, D. M. Eccles, K. Tucker, J. Benitez, S. M. Domchek, A. E. Toland, E. J. Van Rensburg, B. Wappenschmidt, Å. Borg, M. P. Vreeswijk and D. E. Goldgar (2012). "BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk." J Med Genet **49**(8): 525-532.

Stolarova, L., P. Kleiblova, M. Janatova, J. Soukupova, P. Zemankova, L. Macurek and Z. Kleibl (2020). "CHEK2 Germline Variants in Cancer Predisposition: Stalemate Rather than Checkmate." Cells **9**(12).

Tavor, S., S. Takeuchi, K. Tsukasaki, C. W. Miller, W. K. Hofmann, T. Ikezoe, J. W. Said and H. P. Koeffler (2001). "Analysis of the CHK2 gene in lymphoid malignancies." Leuk Lymphoma **42**(3): 517-520.

Tavtigian, S. V., P. J. Oefner, D. Babikyan, A. Hartmann, S. Healey, F. Le Calvez-Kelm, F. Lesueur, G. B. Byrnes, S. C. Chuang, N. Forey, C. Feuchtinger, L. Gioia, J. Hall, M. Hashibe, B. Herte, S. McKay-Chopin, A. Thomas, M. P. Vallée, C. Voegele, P. M. Webb, D. C. Whiteman, S. Sangrajrang, J. L. Hopper, M. C. Southey, I. L. Andrulis, E. M. John and G. Chenevix-Trench (2009). "Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer." Am J Hum Genet **85**(4): 427-446.

Tischkowitz, M., M. Capanu, N. Sabbaghian, L. Li, X. Liang, M. P. Vallée, S. V. Tavtigian, P. Concannon, W. D. Foulkes, L. Bernstein, J. L. Bernstein and C. B. Begg (2012). "Rare germline mutations in PALB2 and breast cancer risk: a population-based study." Hum Mutat **33**(4): 674-680.

Toss, A., E. Tenedini, C. Piombino, M. Venturelli, I. Marchi, E. Gasparini, E. Barbieri, E. Razzaboni, F. Domati, F. Caggia, G. Grandi, F. Combi, G. Tazzioli, M. Dominici, E. Tagliafico and L. Cortesi (2021). "Clinicopathologic Profile of Breast Cancer in Germline ATM and CHEK2 Mutation Carriers." Genes (Basel) **12**(5).

van der Merwe, N. C., N. Hamel, S. R. Schneider, J. P. Apffelstaedt, J. T. Wijnen and W. D. Foulkes (2012). "A founder BRCA2 mutation in non-Afrikaner breast cancer patients of the Western Cape of South Africa." Clin Genet **81**(2): 179-184.

Walsh, T., S. Casadei, M. K. Lee, C. C. Pennil, A. S. Nord, A. M. Thornton, W. Roeb, K. J. Agnew, S. M. Stray, A. Wickramanayake, B. Norquist, K. P. Pennington, R. L. Garcia, M. C. King and E. M. Swisher (2011). "Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing." Proc Natl Acad Sci U S A **108**(44): 18032-18037.

Walsh, T., J. B. Mandell, B. M. Norquist, S. Casadei, S. Gulsuner, M. K. Lee and M. C. King (2017). "Genetic Predisposition to Breast Cancer Due to Mutations Other Than BRCA1 and BRCA2 Founder Alleles Among Ashkenazi Jewish Women." JAMA Oncol **3**(12): 1647-1653.

Xia, B., Q. Sheng, K. Nakanishi, A. Ohashi, J. Wu, N. Christ, X. Liu, M. Jasin, F. J. Couch and D. M. Livingston (2006). "Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2." Mol Cell **22**(6): 719-729.

Zheng, Y., T. Walsh, S. Gulsuner, S. Casadei, M. K. Lee, T. O. Ogundiran, A. Ademola, A. G. Falusi, C. A. Adebamowo, A. O. Oluwasola, A. Adeoye, A. Odetunde, C. P. Babalola, O. A. Ojengbede, S. Odedina, I. Anetor, S. Wang, D. Huo, T. F. Yoshimatsu, J. Zhang, G. E. S. Felix, M. C. King and O. I. Olopade (2018). "Inherited Breast Cancer in Nigerian Women." J Clin Oncol **36**(28): 2820-2825.

# Chapter 4

# Pathogenic variants in hereditary cancer predisposition genes exclusively investigated for truncating variants, and variants of unknown significance

## 4.1 INTRODUCTION

The origin of carcinogenesis, and the development and maintenance thereof has been attributed to various variants including somatic single nucleotide variants (SNVs), germline variants, small insertions and deletions, structural variants, and epigenetic alterations (Helleday, Eshtad et al. 2014).

A large set of variants have been identified for breast cancer, all with varying degrees of impact. While the effect of these variants in isolation or in combination is still not fully understood, enough work has been done to group variants into different severity/penetrance sets. Thus far, three groups have been identified: High, moderate/medium, and low penetrance (Foulkes 2008). To be able to categorize cancer variants, factors that are taken into consideration include penetrance and population frequency.

Penetrance can be defined as the appearance of a disease phenotype given that a specific variant/mutation is present. If a population carries a breast cancer variant on the BRCA1 gene the penetrance can be calculated by the proportion of the population that have breast cancer while carrying this specific variant (Mahdavi, Nassiri et al. 2019). Studies have revealed that only 5% of breast cancer cases are caused by high penetrance genes (Newman, Austin et al. 1988, Hall, Lee et al. 1990).

High penetrance genes have a relatively high risk of disease occurrence when mutated. They are classified as high penetrance when the disease is four or more times more prevalent compared to what is seen in general. These genes included: BRCA1, BRCA2, CDH1, PTEN, STK11 and TP53 (Tsaousis, Papadopoulou et al. 2019, Angeli, Salvi et al. 2020).

Moderate/medium penetrance genes are observed between two and four times more than seen in the general population. They include ATM, BRIP1, CHEK2 and PALB2 (Tsaousis, Papadopoulou et al. 2019).

Finally, low penetrance genes are seen in up to twice as many cases as in the general population. Also, genes that are believed to have a clinical effect but with no sufficient data to imply their cancer risk are also lumped under low penetrance (Tsaousis, Papadopoulou et al. 2019). Some genes include but are not limited to, BARD1, BLM, CHEK1, NF1, RAD50, RAD51 and XRCC2 (Tsaousis, Papadopoulou et al. 2019). The BARD1 gene has been linked to pathogenic variants which increase the life-time breast cancer risk of patients two-fold (Ghimenti, Sensi et al. 2002, Apostolou and Fostira 2013). The RAD51 protein is known for its role in double strand DNA break repair and may play an important role as a cancer target when variants occur, its paralogs have previously been linked to breast cancer as well as ovarian cancer . There is no clear evidence to what the penetrance is for this gene at the current time.

A truncating variant causes the affected gene to produce a shortened version of the protein which may or may not be functional. Enough studies have shown that in most cases, truncating variants have a more severe impact than missense variants, not to say missense variants aren't important. Because of the often-significant effect that a truncating variant confers, it is far easier to observe this effect, while a missense variant might have no or little effect or the effects may be lost in an ocean of other missense variants which may complicate the study thereof even more.

Research has shown that different ethnic populations may carry their own specific genetic variants in key genes. This phenomenon is known as a founder mutation and occurs when a founder (early ancestor) of a population carries this variant and it is transferred to the progeny and it becomes fixed in the population when the population grows substantially (Ferla, Calò et al. 2007).

In this chapter we focus on a smaller subset of variants excluding well-known cancer-related genes (which have been discussed in the previous chapter) and only focussing on variants that cause truncated proteins. Some are known founder mutations and others may be in future be identified as founder mutations.

## 4.2 Other cancer susceptibility variants

### 4.2.1 PATHOGENIC VARIANTS IN HEREDITARY CANCER PREDISPOSITION GENES EXCLUSIVELY INVESTIGATED FOR TRUNCATING VARIANTS

**(See lower portion of Table 3.2, the discussion of the first 4 pathogenic / likely-pathogen variants is provided in Chapter 3).**

**ALK** (NM_004304.4): c.2782dup

The anaplastic lymphoma kinase (ALK) gene can be found on the shorter arm of chromosome 2 (2p23). It plays an important role in proliferation, survival, and differentiation of cells in the nervous system (Yao, Cheng et al. 2013), more specifically in the brain. The greatest anomaly linked to ALK and its role in cancer lies in the fact that it forms fusion proteins during chromosome rearrangements, usually the 3' half of ALK fuses with a 5' fragment of another gene, ALK providing its kinase catalytic domain (tyrosine kinase) while the other gene has the promoter region (Holla, Elamin et al. 2017).

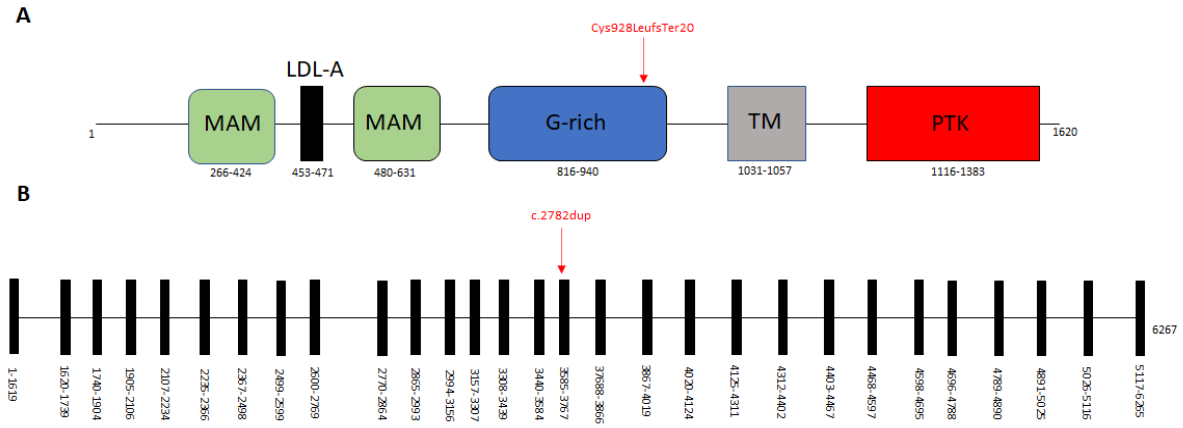


Figure 4.1: A, Protein structure of the ALK gene. MAM (merpin, A5 protein and receptor protein tyrosine phosphatase mu) domains are believed to play a role in cell-to-cell interaction, the LDL-A (low-density lipoprotein domain) role is unknown but is probably involved in some form of ligand binding, G-rich (glycine-rich) domain with an unknown function, TM (extracellular transmembrane domain), PTK (intracellular tyrosine kinase) domain has an enzyme catalytic role. B, Representation of the exons of the ALK gene. In red, the location of the variant from this study is indicated.

A variety of genes have been identified which creates functional combinations of proteins with ALK and inevitably led to some form of cancer (Table 4.1).

Table 4.1: Known genes which forms fusion proteins with ALK.

| Gene | Translocation | Linked malignancies | Reference |
|---|---|---|---|
| 5-Aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP | inv(2) (p23q35) | ALCL* | (Wlodarska, De Wolf-Peeters et al. 1998, Colleoni, Bridge et al. 2000, Trinei, |

| | | | |
|---|---|---|---|
| cyclohydrolase | | | Lanfrancone et al. 2000) |
| Clathrin heavy chain | t(2;17) (p23;q23) | ALCL*, inflammatory myofibroblastic tumour, ALK+ diffuse large B-cell lymphoma | (Touriol, Greenland et al. 2000, McManus, Catherwood et al. 2004, Yamamoto, Kohashi et al. 2006) |
| Dynactin subunit 1 | t(2;2) (p13;p23) | Lung cancer | (Wang, Krishnan et al. 2012, Iyevleva, Raskin et al. 2015) |
| Echinoderm Microtubule Associated Protein-Like 4 | inv(2) (p21p23) | Lung cancer, breast cancer | (Soda, Choi et al. 2007, Chiarle, Voena et al. 2008) |
| Eukaryotic translation elongation factor 1 gamma | t(2;11) (p23; q12.3) | ALCL* | (Palacios, Shaw et al. 2017) |
| GRIP and coiled-coil domain-containing protein 2 | t(2;2) (p23;q12) | Lung cancer | (Jiang, Wu et al. 2018) |
| Kinesin family member 5B | t(2;10) (p23;p11) | Lung cancer | (Takeuchi, Choi et al. 2009, Wong, Leung et al. 2011, Zeng, Liu et al. 2021) |
| Kinesin light chain 1 | t(2;14) (p23;q32) | Lung cancer | (Togashi, Soda et al. 2012) |
| Moesin | t(2;22) (q11;p23) | ALCL* | (Tort, Pinyol et al. 2001) |
| Myosin heavy chain 9 | t(2;22) (p23;q11) | ALCL* | (Lamant, Gascoyne et |

| | | | |
|---|---|---|---|
| | | | al. 2003) |
| Nucleophosmin 1 | t(2;5) (p23;q35) | ALCL* | (Morris, Kirstein et al. 1994, Bischof, Pulford et al. 1997) |
| Protein tyrosine phosphatase, non-receptor type 3 | t(2;9) (p23;q31) | Lung cancer | (Jung, Kim et al. 2012) |
| Ring finger protein 213 | t(2;17) (p23;q25) | ALCL* | (Cools, Wlodarska et al. 2002) |
| Striatin | del(2) (p22p23) | Lung cancer | (Kelly, Barila et al. 2014, Su, Jiang et al. 2020) |
| TNF receptor associated factor 1 | t(2;9) (p23;q33) | ALCL* | (Feldman, Vasmatzis et al. 2013) |
| TRK-fused gene | t(2;3) (p23;q21) | ALCL* | (Hernández, Pinyol et al. 1999, Hernández, Beà et al. 2002) |
| Tropomyosin 3 | t(1;2) (q25;p23) | ALCL*, inflammatory myofibroblastic tumors | (Lamant, Dastugue et al. 1999, Lawrence, Perez-Atayde et al. 2000) |
| Tropomyosin 4 | t(2;19) (p23;p13) | ALCL*, inflammatory myofibroblastic tumors | (Lawrence, Perez-Atayde et al. 2000) |

* Anaplastic Large Cell Lymphoma

Most notably, ALK has been linked to Anaplastic Large Cell Lymphoma (ALCL). ALCL is characterized as a blood cancer and develops in the white blood cells, more specifically the T cells (Medeiros and Elenitoba-Johnson 2007). In 1985, it was first described as neo-plasm Ki-1 lymphoma (Stein, Mason

et al. 1985), but has since undergone many name changes because of the complexity and variety of the lymphoma.

Today, ALCL can be subdivided into four distinct lymphomas: ALK-negative primary cutaneous ALCL, breast implant-associated ALCL, systemic ALK-negative (ALK-) or ALK-positive (ALK+) ALCL (Andraos, Dignac et al. 2021). The main characteristic of ALK-positive ALCL is its overrepresentation of the ALK gene, this is fuelled by the  fusion protein formation, made up of the kinase catalytic domain from ALK (3' region) and the promoter region from other gene (5' region) (Holla, Elamin et al. 2017). Fusion proteins with Nucleophosmin 1 (NPM-ALK) makes up the bulk of ALK+ ALCLs (75% -80%) (Duyster, Bai et al. 2001).

A recent publication found that with a slight shift in the transcription initiation site an oncogenic ALK isoform ($ALK^{ATI}$) is more readily identified (Wiesner, Lee et al. 2015). The variant that was identified in our study in patients, BRB104 and BRB28, c.2782dup, has not previously been identified and will inevitably lead to a truncated protein. This will lead to a partial or total loss of function of the protein. ALK has previously been identified as a receptor tyrosine kinase (RTK) (Morris, Naeve et al. 1997) and a variety of RTKs (EGFR, HER2, MET) have been linked to carcinogenesis when the kinase domain is mutated (Di Nicolantonio and Bardelli 2006), while this variant is located on exon 16 that falls within the glycine rich region of the extracellular part of the protein (Holla, Elamin et al. 2017). The effect of this change is still unclear and warrants further investigation.

**BUB1B** (NM_001211.5): c.2848C > T

A lot of studies have indicated the importance of chromosome instability (CIN) in the development of cancer cells. Chromosome instability gives cancer cells an adaptive advantage over normal cells and one of the main causes of CIN is irregular/compromised mitosis (Lee 2014, Koyuncu, Sharma et al. 2021). Cancer cells are constantly under genomic instability and have high division rates, so that cells that can resist/survive this instability may be selected for (Baker, Chen et al. 2005).

Cell cycle checkpoints are important stages of mitosis which ensure the integrity of the cell, one of these is the spindle assembly checkpoint, which has an important role in distributing genetic material evenly during mitosis (Suijkerbuijk, van Osch et al. 2010). Studies have shown an increase in SAC gene expression not just in breast cancer cells but other cancer cells as well (Yuan, Xu et al. 2006, Fu, Chen et al. 2016, Zhuang, Yang et al. 2018, Dong, Huang et al. 2019, Koyuncu, Sharma et al. 2021, Sekino, Han et al. 2021). The hypothesis stands that these SAC proteins help cancer cells to

negate mitotic stress that would in normal circumstances have catastrophic consequences and/or cell death (Lee 2014, Koyuncu, Sharma et al. 2021).

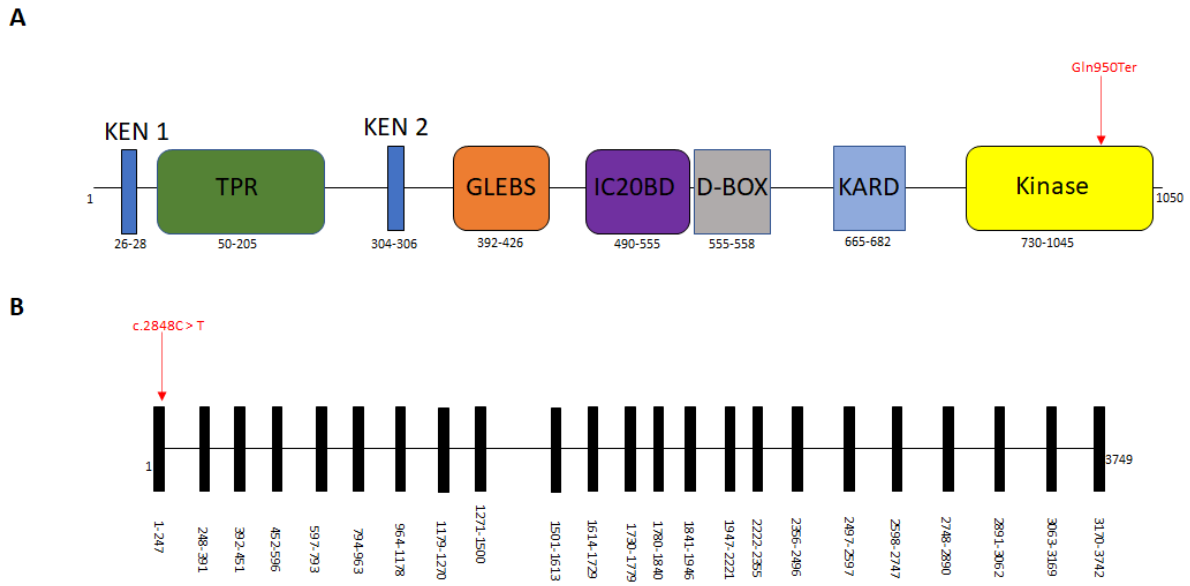


Figure 4.2: A, the protein structure of BUB1B. The protein structure contains, two KEN (Lysine-Glutamic Acid-Asparagine) boxes recognize and ensure efficient phosphorylation of CDC20, TPR (tetratrico-peptide repeat) region may play a role in binding blinkin, GLEBS (Gle2-binding-sequence) is known as the BUB3 binding site , IC20DB (formerly known as ABBA-binding site), is an important Cdc20 binding site, D-BOX (destruction box) plays a role in the degradation of the protein, KARD (kinetochore attachment regulatory domain), is important for kinetochore-microtubule monitoring and a putative kinase domain which may play some role in the stability of the protein but this is still unclear/speculative. B, all exons related to the BUB1B gene. Red labels indicate the identified variant from this study.

Activation of the SAC signaling pathway includes an array of genes which are upregulated in breast cancer cells: BubR1 (gene: BUB1B), Bub3, Mad2, and Cdc20, just to name a few (Koyuncu, Sharma et al. 2021). It has already been shown in mice that a mutated BUB1 Mitotic Checkpoint Serine/Threonine Kinase B gene (BUB1B) gene increases the cancer susceptibility as well as chromosome instability of the mutation carrier (Dai, Wang et al. 2004). BRB261 aged, 38 at time of diagnosis was the only patient (1/165, 0.01%) that carried this variant.

The BUB1B gene may be an important therapeutic target for cancer treatment. The gene, if overexpressed may protect cancer cells from chromosome instability but the absence of the BUB1B gene leads to cell death (Koyuncu, Sharma et al. 2021). Further studies are required to validate the effect of the BUB1B variant found in this study. It may be that this variant was only a passenger mutation linked to another mutation that caused the development of the cancer.

Further studies would also give us a clearer picture of the effect of the variant. Does the truncated BUB1B protein negatively affect the expression pathway by not signalling the next gene which may lead to cancer development, or does the inefficiency of the protein have an unintentional positive

effect by causing cell death in cancer cells? Recent studies have shed light on this question, Koyuncu *et al*. found that the viability of BUB1B plays an important role in cancer cells (Koyuncu, Sharma et al. 2021). They proved that the presence of BUB1B played a protective role in the cancer cell survivability, while BUB1B knockdown led to apoptosis in these cells (Koyuncu, Sharma et al. 2021).

**FANCG** (NM_004629.1): c.637-643del

Fanconi anaemia (FA) is a rare genetically inherited disease which leads mainly to progressive bone marrow failure to produce platelets, red and white blood cells (Butturini, Gale et al. 1994), the development of malignancies and to a lesser extent, endocrine abnormalities (Dillon, Feben et al. 2020). The disease may be caused by a mutation which occurs in, but not limited to, one of the following genes: FANCA, FANCB, FANCC, FANCD1/BRCA2, FANCD2, FANCE, FANCF, **FANCG**, FANCI, FANCJ, and FANCL (Morgan, Essop et al. 2005).

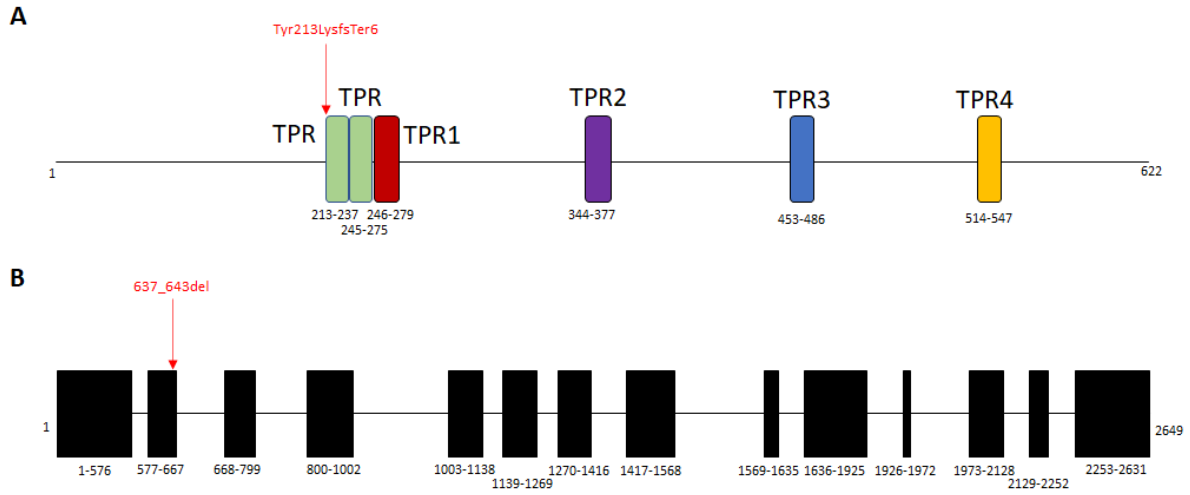


Figure 4.3: A, the protein structure of FANCG. Contains multiple TPR (tetratrico-peptide repeat) regions which function as scaffolds mediating protein-protein interactions. B, all exons related to the FANCG gene. Red labels indicate the identified variant from this study.

FANCG in conjunction with the other FA proteins plays a pivotal role in signalling the activation of FANCD2 which localizes and recruits DNA repair proteins in response to DNA damage (Qiao, Mi et al. 2004). Through this signalling FANCD2 can form a heterodimer with FANCI and recruit DNA damage response effectors to handle the damage, the pathway is then reversed when the DNA has been repaired (Nijman, Huang et al. 2005). As much as 10 percent of FA cases are attributed to mutations in FANCG (Nalepa and Clapp 2018).

World-wide carrier incidence of heterozygotic FA is calculated to be 1/300 while ethnic groups with FA founder mutations such as Ashkenazi Jews (Whitney, Saito et al. 1993, Auerbach 1997) and Afrikaners (Rosendorff, Bernstein et al. 1987) have been shown to be as high as 1/89 and 1/77

respectively. Furthermore, FA has been linked to 21 different genes, which makes diagnosis together with its similarities with other syndromes so much more difficult (Nalepa and Clapp 2018).

Mutations in FA genes not may only cause FA but these genes have been linked to a variety of cancers including bladder, breast, colorectal, connective tissue, liver, lung, ovarian, pancreatic, prostate, skin, uterine cancers, also glioblastomas, leukaemias, lymphomas, melanomas, sarcomas, and squamous cell carcinoma (Nalepa and Clapp 2018).

This variant has previously been identified as a founder mutation in the Bantu-speaking black populations of sub-Saharan Africa (Weber, Nash et al. 2000, Morgan, Essop et al. 2005). It has also been linked to endocrine complications. Researchers were able to show an increase in endocrine problems within a patient population only carrying this FA variant (Dillon, Feben et al. 2020). They were able to identify clear indicators of normal endocrine deviations including abnormal IGF-1/IGFBP-3 levels, insulin resistance, abnormal thyroid functions, and short stature. In this study, two patients BRB98 and BRB225 (2/165, 0.01%) were found to carry this variant.

Interestingly, some work has been done to identify the effectiveness of a truncated FANCG protein (Kuang, Garcia-Higuera et al. 2000). They were able to identify the complex formation and translocation of the truncated protein into the nucleus and its inability to correct the mitomycin C sensitivity (Kuang, Garcia-Higuera et al. 2000), where in a normal circumstance it would have been corrected. More importantly, another study has shown that in various regions mutated TPR region in the FANCG gene led to complete or partial loss of FANCG function (Blom, van de Vrugt et al. 2004), which has a rather important implication in this study.

These are but two examples of the effect of a truncation has on the FANCG protein. Because of the wide array of clinical effects of FA variants, which may have been coupled with the founder mutation and due to the fact that this is a truncating variant, this variant should be a very important identifier for future studies, specifically for southern Africa populations.

### 4.2.2 LESSER-KNOWN HEREDITARY VARIANTS IDENTIFIED: INTRONIC VARIANTS

**RB1** (NM_000321.2): c.1127 + 1G > A

The retinoblastoma (RB1) gene plays a pivotal role as a tumour suppressor in the form of a transcriptional co-factor (Jones, Robinson et al. 2016). It has been shown to have a large variety of targets or RB1-binding proteins to influence cell cycle, proliferation and survivability (Morris and

Dyson 2001, St-Pierre, Liu et al. 2005). RB1 has been associated with a variety of cancers, including retinoblastoma, osteosarcoma, adenocarcinomas, small cell lung cancer, breast cancer, prostate cancer, and more (Harbour and Dean 2000). RB1 acts as a negative regulator of the cell cycle and its inactivation and/or absence may have a large effect on the development of cancer (Harbour and Dean 2000, Lee, Chang et al. 2002).

Currently, there are two mechanisms by which RB1 is inactivated in breast cancer (Witkiewicz and Knudsen 2014). The first is a homozygous loss of the RB protein, most prominently seen in triple negative breast cancer (TNBC) (Witkiewicz and Knudsen 2014), the triple in the name refers to the absence of the estrogen receptor (ER), progesterone receptor (PR) and the epidermal growth factor receptor 2 (HER2). This type of cancer can only be treated through chemotherapy because of the absence of other therapeutic targets (Knudsen and Zacksenhaus 2018).

The second method involves the inactivation of RB1 through phosphorylation. The main effector of this, an aberrant CDK4/6 is the initiation point of the phosphorylation pathway (Network 2012). Because of the multitude of CDK4/6-containing complexes and the effect it may have on breast cancer a lot of research has been done on targeting it therapeutically (Condorelli, Spring et al. 2018), which additionally indicates the importance of understanding RB1 and its variants, it may even be deemed to be an important therapeutic target in the treatment of TNBC (Knudsen and Zacksenhaus 2018).

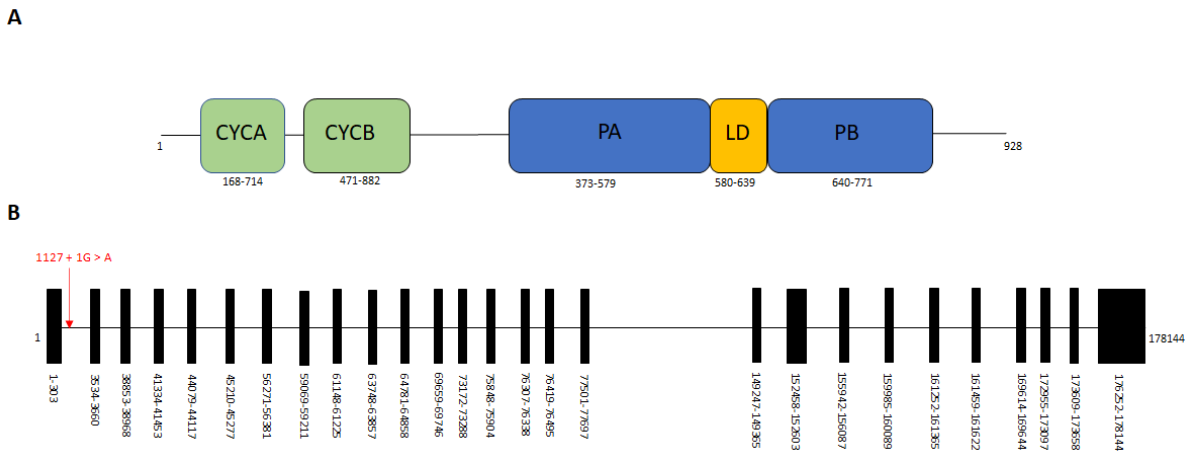


Figure 4.4: A, the protein structure of RB1. Both CYCA and CYCB (Cyclin box) regions play an important role in protein-protein interactions. PA and PB ("Pocket") are required for interactions with the E2F transcription factor, LD ("spacer") region which previously was thought to have no function, now found to be very conserved and believed to play an important role in RB1. B, all exons related to the FANCG gene. The red label indicates the identified variant from this study.

In this study we identified only one patient, BRB73 (1/165, 0.01%) with this variant. This variant has not previously been identified by other researchers but is regarded as deleterious, we can

hypothesize that it may have an impact on cell cycle control which may lead to the development of cancer. It may even, in the future, play a role as a southern Africa specific variant, but a far larger sample cohort will be required to corroborate it as a population specific variant. Interestingly, except for this variant, 21 (21/165, 12.7%) other patients were identified to carry moderately pathogenic RB1 variants. Seventeen (17/165, 10.3%) patients carried an inframe deletion, NM_000321.2:c.45_53del and three (3/165, 0.02%) patients the same missense variant, NM_000321.2:c.1574C>G.

**XPC** (NM_004628.4): c.2251-1G > C

Xeroderma pigmentosum, complementation group C (XPC) forms part of the nucleotide excision repair (NER) proteins, a group of proteins which are responsible for the repair of damage caused by heavy metals, intra-stranded DNA cross-links, organic combustion, oxidative stress, and ultraviolet radiation (Pongsavee and Wisuwan 2018, Malik, Zia et al. 2020). XPC is responsible for the recognition of UV damage on DNA, it forms a complex with RAD23 homolog B (HR23B) which binds to DNA lesions, identifies the damage, then promotes unwinding of the DNA to initiate the NER process (Schäfer 2013, Lehmann 2017). Mutations in XPC may lead to severe ocular malignant lesions and precocious skin cancers, which forms part of the clinical description of most studies done on black patients (Cartault, Nava et al. 2011).

A well-known disease linked to XPC mutations is Xeroderma pigmentosum, which is characterized by skin cancer and hypersensitivity to sunlight (Cleaver 2004). More recently, research has also linked XPC to base excision repair (BER) in fibroblasts (Fayyad, Kobaisi et al. 2020), the team was able to show a compromised BER pathway in the presence of mutated or lost XPC proteins. They further hypothesized that this impairment may even explain the diverse clinical symptoms seen in patients (Fayyad, Kobaisi et al. 2020).

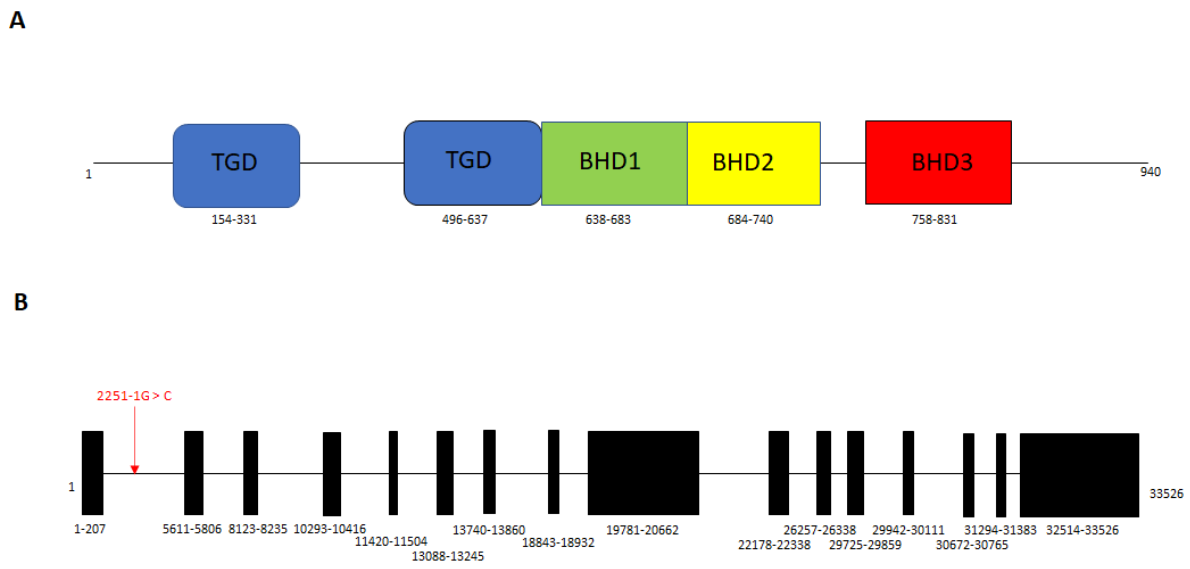


Figure 4.5: A, the protein structure of XPC. The TGD (transglutaminase-like domain and BHD1 (β-hairpin domains) plays a role in binding to double stranded DNA, BHD2 has been linked to detection of damaged DNA and the BHD3 region is used to anchor the protein to the damaged DNA. B, all exons related to the XPC gene. The red label indicates the identified variant from this study.

BRB114 and BRB161 (2/165, 0.01%???), patients from this study were the only ones to carry this variant. According to statistics, black populations are less prone to XPC variants but not much is known about black populations from sub-Saharan Africa (Cartault, Nava et al. 2011). More recently, studies from the Comoros(Cartault, Nava et al. 2011) and South African(Kgokolo, Morice-Picard et al. 2019) proved that these populations share the XPC founder mutation c.2251-1G>C. The work done supports the idea that the founder mutation may have originated in Bantu-speaking black populations from southern Africa more than a thousand years old and only later got inherited by the people from the Comoros (Kgokolo, Morice-Picard et al. 2019). The presence of a founder mutation in the black population of South Africa and the physical abnormalities it may entail, warrants farther and a more in-depth study into this variant.

## 4.3 VARIANTS OF UNKNOWN SIGNIFICANCE

Here follows a collection of variants where the significance is not understood (VUS) and may carry some weight regarding cancer susceptibility (Table 4.2). The difference to other VUS's identified in this study is that these variants were located in twelve well categorized cancer genes. Although some of these variants have previously been identified in studies, they did not seem to play a significant role in the patients. These variants are reported here to help future studies where the significance of them would be better researched.

Table 4.2: Variants of unknown clinical significance identified in a South African Breast cancer cohort of African ancestry. Variants are named according to the Human Genome Variation Society (HGVS) nomenclature, where complimentary DNA (cDNA) numbering +1 corresponds to the A of the ATG translation initiation codon.

| Gene (RefSeq)* | Variant | Exon | Predicted protein change | dbSNP | Patient | Age (yrs:mnths) |
|---|---|---|---|---|---|---|
| *ATM* (NM_000051.3) | c.131A>G | Exon 3 | p.Asp44Gly | rs150143957 | BRB146 | 52:2 |
| | | | | | BRB38 | 46:10 |
| | | | | | BRB49 | 43:4 |
| *ATM* | c.320G>A | Exon 4 | p.Cys107Tyr | rs142358238 | BRB171 | 37:4 |
| | | | | | BRB68 | 43:0 |
| *ATM* | c.1358C>T | Exon 10 | p.Pro453Leu | rs786204124 | BRB121 | 54:0 |
| | | | | | BRB170 | 40:2 |
| | | | | | BRB194 | 44:1 |
| *ATM* | c.3078G>C | Exon 21 | p.Trp1026Cys | - | BRB146 | 52:2 |
| *ATM* | c.4329C>A | Exon 29 | p.His1443Gln | rs377065665 | BRB131 | 45:3 |
| | | | | | BRB17 | 44:11 |
| | | | | | BRB229 | 38:11 |
| | | | | | BRB281 | 51:0 |
| | | | | | BRB78 | 36:4 |
| *ATM* | c.6176C>T | Exon 42 | p.Thr2059Ile | rs144761622 | BRB239 | 51:11 |
| | | | | | BRB241 | 40:1 |
| | | | | | BRB252 | 40:9 |
| *ATM* | c.6194T>C | Exon 42 | p.Ile2065Thr | rs372838622 | BRB19 | 36:0 |
| *ATM* | c.8558C>T | Exon 58 | p.Thr2853Met | rs141534716 | BRB10 | 42:8 |
| | | | | | BRB162 | 42:10 |
| | | | | | BRB203 | 39:7 |
| | | | | | BRB270 | 41:2 |
| | | | | | BRB73 | 29:11 |
| *BRCA2* (NM_000059.3) | c.4798_4800del | Exon 11 | p.Asn1600del | - | BRB193 | 43:5 |
| | | | | | BRB268 | 46:2 |
| | | | | | BRB98 | 43:3 |
| *BRCA2* | c.7762A>G | Exon 16 | p.Ile2588Val | - | BRB158 | 53:7 |
| *BRCA2* | c.8390A>G | Exon 19 | p.Asp2797Gly | - | BRB8 | 28:0 |
| *BRCA2* | c.9088A>C | Exon 23 | p.Thr3030Pro | - | BRB88 | 39:3 |
| *BRIP1* (NM_032043.2) | c.2131A>G | Exon 15 | p.Thr711Ala | rs760515227 | BRB207 | 49:9 |
| *MSH2* (NM_000251.2) | c.508C>G | Exon 3 | p.Gln170Glu | rs63750843 | BRB106 | 35:0 |
| | | | | | BRB14 | 47:8 |
| | | | | | BRB154 | 51:2 |
| | | | | | BRB238 | 44:10 |
| *MSH6* (NM_000179.2) | c.560A>G | Exon 3 | p.Lys187Arg | - | BRB246 | 42:1 |
| *MSH6* | c.2083C>T | Exon 4 | p.Leu695Phe | - | BRB182 | 41:2 |
| | | | | | BRB208 | 44:8 |
| | | | | | BRB284 | 30:10 |
| | | | | | BRB98 | 43:3 |
| *MSH6* | c.2347T>A | Exon 4 | p.Cys783Ser | rs373721483 | BRB74 | 47:4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *MSH6* | c.2962C>T | Exon 4 | p.Arg988Cys | rs61753795 | BRB62 | 43:7 |
| *MSH6* | c.3489A>C | exon 6 | p.Glu1163Asp | rs531674673 | BRB239 | 51:11 |
| | | | | | BRB270 | 41:2 |
| | | | | | BRB276 | 46:6 |
| | | | | | BRB51 | 52:6 |
| | | | | | BRB52 | 46:11 |
| | | | | | BRC134 | 45:11 |
| *NBN* (NM_002485.4) | c.706A>G | Exon 7 | p.Lys236Glu | rs1060503482 | BRB89 | 48:8 |
| *NF1* (NM_001042492.2) | c.4943C>T | Exon 37 | p.Thr1648Ile | rs376655102 | BRB174 | 37:8 |
| | | | | | BRB42 | 49:11 |
| *PALB2* (NM_024675.3) | c.23C>T | Exon 1 | p.Pro8Leu | rs150390726 | BRB55 | 32:9 |
| | | | | | BRB89 | 48:8 |
| *RAD51C* (NM_058216.2) | c.779G>A | Exon 5 | p.Arg260Gln | rs730881926 | BRB197 | 51:0 |
| *RAD51D* (NM_002878.3) | c.250A>G | Exon 3 | p.Thr84Ala | rs200018296 | BRB111 | 42:7 |
| *STK11* (NM_000455.4) | c.888G>C | Exon 7 | p.Lys296Asn | rs1555738868 | BRB199 | 45:2 |
| | | | | | BRB275 | 47:4 |
| *TP53* (NM_000546.5) | c.476C>T | Exon 5 | p.Ala159Val | rs1555526131 | BRB102 | 40:9 |
| *TP53* | c.393_395del | Exon 5 | p.Asn131del | rs879254214 | BRB234 | 39:10 |

*Reference sequences obtained from the NCBI database. For *BRCA1* the most common human transcript (NM_007294.3) is used with custom numbering of the exons (missing exon 4).

Of interest, both p.Asp44Gly and p.Glu2181Asp variants were also identified in breast cancer patients from Cameroon and Uganda (Adedokun, Zheng et al. 2020). Patients BRB68, BRB146 and BRB171 were the only patients that had additional variants in ATM in conjunction with VUS in ATM. All other VUS identified had no extra variants in the same gene.

There is another variant of note, the PALB2 N-terminus variant c.23C > T p.(P8L) detected in two patients (BRB55 & BRB89). This variant is near the coiled-coil domain of PALB2 that is involved in hetero-dimerization of BRCA1 with the protein. PALB2 is an essential component in homologous recombination-based DNA repair (HR) and loss of PALB2 function was shown to be synthetic lethal in combination with poly(ADP-ribose) polymerase inhibitors (PARPi) (Shen, Rehman et al. 2013, Smith, Hampton et al. 2015). This has led to the development of tests that exploit this weakness to assess the functional effect of PALB2 sequence variants.

Functional assays that test the vulnerability of PALB2 variants to PARP inhibitors as well as HR functionality were applied to the p.(P8L) variant. Moderate but statistically significant ($P < 0.0001$) PARPi sensitivity was observed (76% cell survival), whereas wild type PALB2 had 100% cell survival (Rodrigue, Margaillan et al. 2019). The homology-directed repair assay found p.(P8L) to have an intermediate phenotype with a 40% reduction in HR when compared to wild type PALB2 (Rodrigue,

Margaillan et al. 2019, Boonen, Vreeswijk et al. 2020), all of which appear to indicate that this variant may play a role in breast cancer.

As these are VUS's, this section can only function as a report to bring these variants under the attention of the research community.

## 4.4 DISCUSSION

This chapter set out to focus on some lesser-known truncating variants that may have the highest probability of having a detrimental effect on patients. We were able to identify and stipulate the importance of some variants that may in fact play a far greater role in the formation of cancer.

The Pan-Cancer Analysis of Whole Genomes (PCAWG) variant dataset already contains over 44 million SNVs, only about one percent of these SNVs have been identified as driver mutations (Campbell, Getz et al. 2020). The rest of these variants are classified as passenger variants, variants of which the effect on the fitness of the carrier as well as the molecular consequence is not understood at all. Some of these passengers may still have a small effect on carcinogenesis and previously were termed as mini-drivers (Castro-Giner, Ratcliffe et al. 2015) and deleterious passengers (McFarland, Korolev et al. 2013). Variants from this study may be identified as mini-drivers in the future.

Most studies to research founder mutations have been on Caucasians (Peelen, van Vliet et al. 1997, Tonin, Mes-Masson et al. 1999, Neuhausen 2000, Ferla, Calò et al. 2007) and a need for other ethnic populations has become apparent. Many African founder mutations have already been identified and some may still be unknown. The Bantu-speaking black populations from southern Africa alone has already been identified to by the founding carriers of cancer-linked BRCA2 (van der Merwe, Hamel et al. 2012), FANCG (Morgan, Essop et al. 2005, Wainstein, Kerr et al. 2013), XPC mutations (Kgokolo, Morice-Picard et al. 2019) and other diseases (Stevens, Ramsay et al. 1997, Krause, Mitchell et al. 2015). These findings should pique the interest of researchers in what may still be unknown regarding founder mutations in southern Africa which have mostly been left untapped. Furthermore, this should also show that by studying variants from Africa, light may be shed on ancestries, migrations and evolution of the people across Africa and the rest of the world.

This research may be the first step for identifying more founder mutations specific to the black South African population. The work done here could be used as a steppingstone to larger, more interesting and more comprehensive studies of breast cancer in southern Africa.

## 4.5 REFERENCES

Adedokun, B., Y. Zheng, P. Ndom, A. Gakwaya, T. Makumbi, A. Y. Zhou, T. F. Yoshimatsu, A. Rodriguez, R. K. Madduri, I. T. Foster, A. Sallam, O. I. Olopade and D. Huo (2020). "Prevalence of Inherited Mutations in Breast Cancer Predisposition Genes among Women in Uganda and Cameroon." Cancer Epidemiol Biomarkers Prev **29**(2): 359-367.

Andraos, E., J. Dignac and F. Meggetto (2021). "NPM-ALK: A Driver of Lymphoma Pathogenesis and a Therapeutic Target." Cancers (Basel) **13**(1).

Angeli, D., S. Salvi and G. Tedaldi (2020). "Genetic Predisposition to Breast and Ovarian Cancers: How Many and Which Genes to Test?" Int J Mol Sci **21**(3).

Apostolou, P. and F. Fostira (2013). "Hereditary breast cancer: the era of new susceptibility genes." BioMed research international **2013**: 747318-747318.

Auerbach, A. D. (1997). "Fanconi anemia: genetic testing in Ashkenazi Jews." Genet Test **1**(1): 27-33.

Baker, D. J., J. Chen and J. M. van Deursen (2005). "The mitotic checkpoint in cancer and aging: what have mice taught us?" Curr Opin Cell Biol **17**(6): 583-589.

Bischof, D., K. Pulford, D. Y. Mason and S. W. Morris (1997). "Role of the nucleophosmin (NPM) portion of the non-Hodgkin's lymphoma-associated NPM-anaplastic lymphoma kinase fusion protein in oncogenesis." Mol Cell Biol **17**(4): 2312-2325.

Blom, E., H. J. van de Vrugt, Y. de Vries, J. P. de Winter, F. Arwert and H. Joenje (2004). "Multiple TPR motifs characterize the Fanconi anemia FANCG protein." DNA Repair (Amst) **3**(1): 77-84.

Boonen, R., M. P. G. Vreeswijk and H. van Attikum (2020). "Functional Characterization of PALB2 Variants of Uncertain Significance: Toward Cancer Risk and Therapy Response Prediction." Front Mol Biosci **7**: 169.

Butturini, A., R. P. Gale, P. C. Verlander, B. Adler-Brecher, A. P. Gillio and A. D. Auerbach (1994). "Hematologic abnormalities in Fanconi anemia: an International Fanconi Anemia Registry study." Blood **84**(5): 1650-1655.

Campbell, P. J., G. Getz, J. O. Korbel, J. M. Stuart, L. D. Stein and T. I. T. P.-C. A. o. W. G. Consortium (2020). "Pan-cancer analysis of whole genomes." Nature **578**(1476-4687 (Electronic)).

Cartault, F., C. Nava, A. C. Malbrunot, P. Munier, J. C. Hebert, P. N'Guyen, N. Djeridi, P. Pariaud, J. Pariaud, A. Dupuy, F. Austerlitz and A. Sarasin (2011). "A new XPC gene splicing mutation has lead to the highest worldwide prevalence of xeroderma pigmentosum in black Mahori patients." DNA Repair (Amst) **10**(6): 577-585.

Castro-Giner, F., P. Ratcliffe and I. Tomlinson (2015). "The mini-driver model of polygenic cancer evolution." Nat Rev Cancer **15**(11): 680-685.

Chiarle, R., C. Voena, C. Ambrogio, R. Piva and G. Inghirami (2008). "The anaplastic lymphoma kinase in the pathogenesis of cancer." Nat Rev Cancer **8**(1): 11-23.

Cleaver, J. E. (2004). "Defective repair replication of DNA in xeroderma pigmentosum. 1968." DNA Repair (Amst) **3**(2): 183-187.

Colleoni, G. W., J. A. Bridge, B. Garicochea, J. Liu, D. A. Filippa and M. Ladanyi (2000). "ATIC-ALK: A novel variant ALK gene fusion in anaplastic large cell lymphoma resulting from the recurrent cryptic chromosomal inversion, inv(2)(p23q35)." The American journal of pathology **156**(3): 781-789.

Condorelli, R., L. Spring, J. O'Shaughnessy, L. Lacroix, C. Bailleux, V. Scott, J. Dubois, R. J. Nagy, R. B. Lanman, A. J. Iafrate, F. Andre and A. Bardia (2018). "Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer." Ann Oncol **29**(3): 640-645.

Cools, J., I. Wlodarska, R. Somers, N. Mentens, F. Pedeutour, B. Maes, C. De Wolf-Peeters, P. Pauwels, A. Hagemeijer and P. Marynen (2002). "Identification of novel fusion partners of ALK, the anaplastic lymphoma kinase, in anaplastic large-cell lymphoma and inflammatory myofibroblastic tumor." Genes Chromosomes Cancer **34**(4): 354-362.

Dai, W., Q. Wang, T. Liu, M. Swamy, Y. Fang, S. Xie, R. Mahmood, Y. M. Yang, M. Xu and C. V. Rao (2004). "Slippage of mitotic arrest and enhanced tumor development in mice with BubR1 haploinsufficiency." Cancer Res **64**(2): 440-445.
Di Nicolantonio, F. and A. Bardelli (2006). "Kinase mutations in cancer: chinks in the enemy's armour?" Curr Opin Oncol **18**(1): 69-76.
Dillon, B., C. Feben, D. Segal, J. du Plessis, D. Reynders, R. Wainwright, J. Poole and A. Krause (2020). "Endocrine profiling in patients with Fanconi anemia, homozygous for a FANCG founder mutation." Mol Genet Genomic Med **8**(8): e1351.
Dong, S., F. Huang, H. Zhang and Q. Chen (2019). "Overexpression of BUB1B, CCNA2, CDC20, and CDK1 in tumor tissues predicts poor survival in pancreatic ductal adenocarcinoma." Biosci Rep **39**(2).
Duyster, J., R. Y. Bai and S. W. Morris (2001). "Translocations involving anaplastic lymphoma kinase (ALK)." Oncogene **20**(40): 5623-5637.
Fayyad, N., F. Kobaisi, D. Beal, W. Mahfouf, C. Ged, F. Morice-Picard, M. Fayyad-Kazan, H. Fayyad-Kazan, B. Badran, H. R. Rezvani and W. Rachidi (2020). "Xeroderma Pigmentosum C (XPC) Mutations in Primary Fibroblasts Impair Base Excision Repair Pathway and Increase Oxidative DNA Damage." Front Genet **11**: 561687.
Feldman, A. L., G. Vasmatzis, Y. W. Asmann, J. Davila, S. Middha, B. W. Eckloff, S. H. Johnson, J. C. Porcher, S. M. Ansell and A. Caride (2013). "Novel TRAF1-ALK fusion identified by deep RNA sequencing of anaplastic large cell lymphoma." Genes Chromosomes Cancer **52**(11): 1097-1102.
Ferla, R., V. Calò, S. Cascio, G. Rinaldi, G. Badalamenti, I. Carreca, E. Surmacz, G. Colucci, V. Bazan and A. Russo (2007). "Founder mutations in BRCA1 and BRCA2 genes." Ann Oncol **18 Suppl 6**: vi93-98.
Foulkes, W. D. (2008). "Inherited susceptibility to common cancers." N Engl J Med **359**(20): 2143-2153.
Fu, X., G. Chen, Z. D. Cai, C. Wang, Z. Z. Liu, Z. Y. Lin, Y. D. Wu, Y. X. Liang, Z. D. Han, J. C. Liu and W. D. Zhong (2016). "Overexpression of BUB1B contributes to progression of prostate cancer and predicts poor outcome in patients with prostate cancer." Onco Targets Ther **9**: 2211-2220.
Ghimenti, C., E. Sensi, S. Presciuttini, I. M. Brunetti, P. Conte, G. Bevilacqua and M. A. Caligo (2002). "Germline mutations of the BRCA1-associated ring domain (BARD1) gene in breast and breast/ovarian families negative for BRCA1 and BRCA2 alterations." Genes Chromosomes Cancer **33**(3): 235-242.
Hall, J. M., M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey and M. C. King (1990). "Linkage of early-onset familial breast cancer to chromosome 17q21." Science **250**(4988): 1684-1689.
Harbour, J. W. and D. C. Dean (2000). "The Rb/E2F pathway: expanding roles and emerging paradigms." Genes Dev **14**(19): 2393-2409.
Helleday, T., S. Eshtad and S. Nik-Zainal (2014). "Mechanisms underlying mutational signatures in human cancers." Nat Rev Genet **15**(9): 585-598.
Hernández, L., S. Beà, B. Bellosillo, M. Pinyol, B. Falini, A. Carbone, G. Ott, A. Rosenwald, A. Fernández, K. Pulford, D. Mason, S. W. Morris, E. Santos and E. Campo (2002). "Diversity of genomic breakpoints in TFG-ALK translocations in anaplastic large cell lymphomas: identification of a new TFG-ALK(XL) chimeric gene with transforming activity." Am J Pathol **160**(4): 1487-1494.
Hernández, L., M. Pinyol, S. Hernández, S. Beà, K. Pulford, A. Rosenwald, L. Lamant, B. Falini, G. Ott, D. Y. Mason, G. Delsol and E. Campo (1999). "TRK-fused gene (TFG) is a new partner of ALK in anaplastic large cell lymphoma producing two structurally different TFG-ALK translocations." Blood **94**(9): 3265-3268.
Holla, V. R., Y. Y. Elamin, A. M. Bailey, A. M. Johnson, B. C. Litzenburger, Y. B. Khotskaya, N. S. Sanchez, J. Zeng, M. A. Shufean, K. R. Shaw, J. Mendelsohn, G. B. Mills, F. Meric-Bernstam and G. R. Simon (2017). "ALK: a tyrosine kinase target for cancer therapy." Cold Spring Harb Mol Case Stud **3**(1): a001115.

Iyevleva, A. G., G. A. Raskin, V. I. Tiurin, A. P. Sokolenko, N. V. Mitiushkina, S. N. Aleksakhina, A. R. Garifullina, T. N. Strelkova, V. O. Merkulov, A. O. Ivantsov, E. Kuligina, K. M. Pozharisski, A. V. Togo and E. N. Imyanitov (2015). "Novel ALK fusion partners in lung cancer." Cancer Lett **362**(1): 116-121.
Jiang, J., X. Wu, X. Tong, W. Wei, A. Chen, X. Wang, Y. W. Shao and J. Huang (2018). "GCC2-ALK as a targetable fusion in lung adenocarcinoma and its enduring clinical responses to ALK inhibitors." Lung Cancer **115**: 5-11.
Jones, R. A., T. J. Robinson, J. C. Liu, M. Shrestha, V. Voisin, Y. Ju, P. E. Chung, G. Pellecchia, V. L. Fell, S. Bae, L. Muthuswamy, A. Datti, S. E. Egan, Z. Jiang, G. Leone, G. D. Bader, A. Schimmer and E. Zacksenhaus (2016). "RB1 deficiency in triple-negative breast cancer induces mitochondrial protein translation." J Clin Invest **126**(10): 3739-3757.
Jung, Y., P. Kim, Y. Jung, J. Keum, S. N. Kim, Y. S. Choi, I. G. Do, J. Lee, S. J. Choi, S. Kim, J. E. Lee, J. Kim, S. Lee and J. Kim (2012). "Discovery of ALK-PTPN3 gene fusion from human non-small cell lung carcinoma cell line using next generation RNA sequencing." Genes Chromosomes Cancer **51**(6): 590-597.
Kelly, L. M., G. Barila, P. Liu, V. N. Evdokimova, S. Trivedi, F. Panebianco, M. Gandhi, S. E. Carty, S. P. Hodak, J. Luo, S. Dacic, Y. P. Yu, M. N. Nikiforova, R. L. Ferris, D. L. Altschuler and Y. E. Nikiforov (2014). "Identification of the transforming STRN-ALK fusion as a potential therapeutic target in the aggressive forms of thyroid cancer." Proc Natl Acad Sci U S A **111**(11): 4233-4238.
Kgokolo, M., F. Morice-Picard, H. R. Rezvani, F. Austerlitz, F. Cartault, A. Sarasin, M. Sathekge, A. Taieb and C. Ged (2019). "Xeroderma pigmentosum in South Africa: Evidence for a prevalent founder effect." Br J Dermatol **181**(5): 1070-1072.
Knudsen, E. S. and E. Zacksenhaus (2018). "The vulnerability of RB loss in breast cancer: Targeting a void in cell cycle control." Oncotarget **9**(57): 30940-30941.
Koyuncu, D., U. Sharma, E. T. Goka and M. E. Lippman (2021). "Spindle assembly checkpoint gene BUB1B is essential in breast cancer cell survival." Breast Cancer Res Treat **185**(2): 331-341.
Krause, A., C. Mitchell, F. Essop, S. Tager, J. Temlett, G. Stevanin, C. Ross, D. Rudnicki and R. Margolis (2015). "Junctophilin 3 (JPH3) expansion mutations causing Huntington disease like 2 (HDL2) are common in South African patients with African ancestry and a Huntington disease phenotype." Am J Med Genet B Neuropsychiatr Genet **168**(7): 573-585.
Kuang, Y., I. Garcia-Higuera, A. Moran, M. Mondoux, M. Digweed and A. D. D'Andrea (2000). "Carboxy terminal region of the Fanconi anemia protein, FANCG/XRCC9, is required for functional activity." Blood **96**(5): 1625-1632.
Lamant, L., N. Dastugue, K. Pulford, G. Delsol and B. Mariamé (1999). "A new fusion gene TPM3-ALK in anaplastic large cell lymphoma created by a (1;2)(q25;p23) translocation." Blood **93**(9): 3088-3095.
Lamant, L., R. D. Gascoyne, M. M. Duplantier, F. Armstrong, A. Raghab, M. Chhanabhai, E. Rajcan-Separovic, J. Raghab, G. Delsol and E. Espinos (2003). "Non-muscle myosin heavy chain (MYH9): a new partner fused to ALK in anaplastic large cell lymphoma." Genes Chromosomes Cancer **37**(4): 427-432.
Lawrence, B., A. Perez-Atayde, M. K. Hibbard, B. P. Rubin, P. Dal Cin, J. L. Pinkus, G. S. Pinkus, S. Xiao, E. S. Yi, C. D. Fletcher and J. A. Fletcher (2000). "TPM3-ALK and TPM4-ALK oncogenes in inflammatory myofibroblastic tumors." Am J Pathol **157**(2): 377-384.
Lee, C., J. H. Chang, H. S. Lee and Y. Cho (2002). "Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor." Genes Dev **16**(24): 3199-3212.
Lee, H. (2014). "How chromosome mis-segregation leads to cancer: lessons from BubR1 mouse models." Mol Cells **37**(10): 713-718.
Lehmann, J. (2017). Functional relevance of spontaneous alternative splice variants of xeroderma pigmentosum genes: Prognostic marker for skin cancer risk and disease outcome?
Mahdavi, M., M. Nassiri, M. M. Kooshyar, M. Vakili-Azghandi, A. Avan, R. Sandry, S. Pillai, A. K. Lam and V. Gopalan (2019). "Hereditary breast cancer; Genetic penetrance and current status with BRCA." J Cell Physiol **234**(5): 5741-5750.

Malik, S. S., A. Zia, S. Rashid, S. Mubarik, N. Masood, M. Hussain, A. Yasmin and R. Bano (2020). "XPC as breast cancer susceptibility gene: evidence from genetic profiling, statistical inferences and protein structural analysis." Breast Cancer **27**(6): 1168-1176.
McFarland, C. D., K. S. Korolev, G. V. Kryukov, S. R. Sunyaev and L. A. Mirny (2013). "Impact of deleterious passenger mutations on cancer progression." Proceedings of the National Academy of Sciences of the United States of America(1091-6490 (Electronic)).
McManus, D. T., M. A. Catherwood, P. D. Carey, R. J. Cuthbert and H. D. Alexander (2004). "ALK-positive diffuse large B-cell lymphoma of the stomach associated with a clathrin-ALK rearrangement." Hum Pathol **35**(10): 1285-1288.
Medeiros, L. J. and K. S. Elenitoba-Johnson (2007). "Anaplastic Large Cell Lymphoma." Am J Clin Pathol **127**(5): 707-722.
Morgan, N. V., F. Essop, I. Demuth, T. de Ravel, S. Jansen, M. Tischkowitz, C. M. Lewis, L. Wainwright, J. Poole, H. Joenje, M. Digweed, A. Krause and C. G. Mathew (2005). "A common Fanconi anemia mutation in black populations of sub-Saharan Africa." Blood **105**(9): 3542-3544.
Morris, E. J. and N. J. Dyson (2001). "Retinoblastoma protein partners." Adv Cancer Res **82**: 1-54.
Morris, S. W., M. N. Kirstein, M. B. Valentine, K. G. Dittmer, D. N. Shapiro, D. L. Saltman and A. T. Look (1994). "Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma." Science **263**(5151): 1281-1284.
Morris, S. W., C. Naeve, P. Mathew, P. L. James, M. N. Kirstein, X. Cui and D. P. Witte (1997). "ALK, the chromosome 2 gene locus altered by the t(2;5) in non-Hodgkin's lymphoma, encodes a novel neural receptor tyrosine kinase that is highly related to leukocyte tyrosine kinase (LTK)." Oncogene **14**(18): 2175-2188.
Nalepa, G. and D. W. Clapp (2018). "Fanconi anaemia and cancer: an intricate relationship." Nat Rev Cancer **18**(3): 168-185.
Network, T. C. G. A. (2012). "Comprehensive molecular portraits of human breast tumours." Nature **490**(7418): 61-70.
Neuhausen, S. L. (2000). "Founder populations and their uses for breast cancer genetics." Breast Cancer Res **2**(1): 77-81.
Newman, B., M. A. Austin, M. Lee and M. C. King (1988). "Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families." Proceedings of the National Academy of Sciences of the United States of America **85**(9): 3044-3048.
Nijman, S. M., T. T. Huang, A. M. Dirac, T. R. Brummelkamp, R. M. Kerkhoven, A. D. D'Andrea and R. Bernards (2005). "The deubiquitinating enzyme USP1 regulates the Fanconi anemia pathway." Mol Cell **17**(3): 331-339.
Palacios, G., T. I. Shaw, Y. Li, R. K. Singh, M. Valentine, J. T. Sandlund, M. S. Lim, C. G. Mullighan and V. Leventaki (2017). "Novel ALK fusion in anaplastic large cell lymphoma involving EEF1G, a subunit of the eukaryotic elongation factor-1 complex." Leukemia **31**(3): 743-747.
Peelen, T., M. van Vliet, A. Petrij-Bosch, R. Mieremet, C. Szabo, A. M. W. van den Ouweland, F. Hogervorst, R. Brohet, M. J. L. Ligtenberg, E. Teugels, R. van der Luijt, A. H. van der Hout, J. J. P. Gille, G. Pals, I. Jedema, R. Olmer, I. van Leeuwen, B. Newman, M. Plandsoen, M. van der Est, G. Brink, S. Hageman, P. J. W. Arts, M. M. Bakker, H. W. Willems, E. van der Looij, B. Neyns, M. Bonduelle, R. Jansen, J. C. Oosterwijk, R. Sijmons, H. J. M. Smeets, C. J. van Asperen, H. Meijers-Heijboer, J. G. M. Klijn, J. de Greve, M. C. King, F. H. Menko, H. G. Brunner, D. Halley, G. J. B. van Ommen, H. F. A. Vasen, C. J. Cornelisse, L. J. van 'tVeer, P. de Knijff, E. Bakker and D. P. (1997). "A high proportion of novel mutations in BRCA1 with strong founder effects among Dutch and Belgian hereditary breast and ovarian cancer families." Am J Hum Genet **60**(9-13).
Pongsavee, M. and K. Wisuwan (2018). "ERCC5 rs751402 polymorphism is the risk factor for sporadic breast cancer in Thailand." Int J Mol Epidemiol Genet **9**(4): 27-33.
Qiao, F., J. Mi, J. B. Wilson, G. Zhi, N. R. Bucheimer, N. J. Jones and G. M. Kupfer (2004). "Phosphorylation of fanconi anemia (FA) complementation group G protein, FANCG, at serine 7 is important for function of the FA pathway." J Biol Chem **279**(44): 46035-46045.

Rodrigue, A., G. Margaillan, T. Torres Gomes, Y. Coulombe, G. Montalban, E. S. C. S. da Costa, L. Milano, M. Ducy, G. De-Gregoriis, G. Dellaire, W. Araújo da Silva, Jr., A. N. Monteiro, M. A. Carvalho, J. Simard and J. Y. Masson (2019). "A global functional analysis of missense mutations reveals two major hotspots in the PALB2 tumor suppressor." Nucleic Acids Res **47**(20): 10662-10677.

Rosendorff, J., R. Bernstein, L. Macdougall and T. Jenkins (1987). "Fanconi anemia: another disease of unusually high prevalence in the Afrikaans population of South Africa." Am J Med Genet **27**(4): 793-797.

Schäfer, A. (2013). Clinical, functional, and genetic analysis of NER defective patients and characterization of five novel XPG mutations. Doctoral Doctoral, Georg-August University Göttingen.

Sekino, Y., X. Han, G. Kobayashi, T. Babasaki, S. Miyamoto, K. Kobatake, H. Kitano, K. Ikeda, K. Goto, S. Inoue, T. Hayashi, J. Teishima, N. Sakamoto, K. Sentani, N. Oue, W. Yasui and A. Matsubara (2021). "BUB1B Overexpression Is an Independent Prognostic Marker and Associated with CD44, p53, and PD-L1 in Renal Cell Carcinoma." Oncology **99**(4): 240-250.

Shen, Y., F. L. Rehman, Y. Feng, J. Boshuizen, I. Bajrami, R. Elliott, B. Wang, C. J. Lord, L. E. Post and A. Ashworth (2013). "BMN 673, a novel and highly potent PARP1/2 inhibitor for the treatment of human cancers with DNA repair deficiency." Clin Cancer Res **19**(18): 5003-5015.

Smith, M. A., O. A. Hampton, C. P. Reynolds, M. H. Kang, J. M. Maris, R. Gorlick, E. A. Kolb, R. Lock, H. Carol, S. T. Keir, J. Wu, R. T. Kurmasheva, D. A. Wheeler and P. J. Houghton (2015). "Initial testing (stage 1) of the PARP inhibitor BMN 673 by the pediatric preclinical testing program: PALB2 mutation predicts exceptional in vivo response to BMN 673." Pediatr Blood Cancer **62**(1): 91-98.

Soda, M., Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, M. Bando, S. Ohno, Y. Ishikawa, H. Aburatani, T. Niki, Y. Sohara, Y. Sugiyama and H. Mano (2007). "Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer." Nature **448**(7153): 561-566.

St-Pierre, B., X. Liu, L. C. Kha, X. Zhu, O. Ryan, Z. Jiang and E. Zacksenhaus (2005). "Conserved and specific functions of mammalian ssu72." Nucleic Acids Res **33**(2): 464-477.

Stein, H., D. Y. Mason, J. Gerdes, N. O'Connor, J. Wainscoat, G. Pallesen, K. Gatter, B. Falini, G. Delsol, H. Lemke and et al. (1985). "The expression of the Hodgkin's disease associated antigen Ki-1 in reactive and neoplastic lymphoid tissue: evidence that Reed-Sternberg cells and histiocytic malignancies are derived from activated lymphoid cells." Blood **66**(4): 848-858.

Stevens, G., M. Ramsay and T. Jenkins (1997). "Oculocutaneous albinism (OCA2) in sub-Saharan Africa: distribution of the common 2.7-kb P gene deletion mutation." Hum Genet **99**(4): 523-527.

Su, C., Y. Jiang, W. Jiang, H. Wang, S. Liu, Y. Shao, W. Zhao, R. Ning and Q. Yu (2020). "STRN-ALK Fusion in Lung Adenocarcinoma with Excellent Response Upon Alectinib Treatment: A Case Report and Literature Review." Onco Targets Ther **13**: 12515-12519.

Suijkerbuijk, S. J., M. H. van Osch, F. L. Bos, S. Hanks, N. Rahman and G. J. Kops (2010). "Molecular causes for BUBR1 dysfunction in the human cancer predisposition syndrome mosaic variegated aneuploidy." Cancer Res **70**(12): 4891-4900.

Takeuchi, K., Y. L. Choi, Y. Togashi, M. Soda, S. Hatano, K. Inamura, S. Takada, T. Ueno, Y. Yamashita, Y. Satoh, S. Okumura, K. Nakagawa, Y. Ishikawa and H. Mano (2009). "KIF5B-ALK, a novel fusion oncokinase identified by an immunohistochemistry-based diagnostic system for ALK-positive lung cancer." Clin Cancer Res **15**(9): 3143-3149.

Togashi, Y., M. Soda, S. Sakata, E. Sugawara, S. Hatano, R. Asaka, T. Nakajima, H. Mano and K. Takeuchi (2012). "KLC1-ALK: a novel fusion in lung cancer identified using a formalin-fixed paraffin-embedded tissue only." PLoS One **7**(2): e31323.

Tonin, P. N., A. M. Mes-Masson, S. A. Narod, P. Ghadirian and D. Provencher (1999). "Founder BRCA1 and BRCA2 mutations in French Canadian ovarian cancer cases unselected for family history." Clin Genet **55**(5): 318-324.

Tort, F., M. Pinyol, K. Pulford, G. Roncador, L. Hernandez, I. Nayach, H. C. Kluin-Nelemans, P. Kluin, C. Touriol, G. Delsol, D. Mason and E. Campo (2001). "Molecular characterization of a new ALK

translocation involving moesin (MSN-ALK) in anaplastic large cell lymphoma." Lab Invest **81**(3): 419-426.
Touriol, C., C. Greenland, L. Lamant, K. Pulford, F. Bernard, T. Rousset, D. Y. Mason and G. Delsol (2000). "Further demonstration of the diversity of chromosomal changes involving 2p23 in ALK-positive lymphoma: 2 cases expressing ALK kinase fused to CLTCL (clathrin chain polypeptide-like)." Blood **95**(10): 3204-3207.
Trinei, M., L. Lanfrancone, E. Campo, K. Pulford, D. Y. Mason, P. G. Pelicci and B. Falini (2000). "A new variant anaplastic lymphoma kinase (ALK)-fusion protein (ATIC-ALK) in a case of ALK-positive anaplastic large cell lymphoma." Cancer Res **60**(4): 793-798.
Tsaousis, G. N., E. Papadopoulou, A. Apessos, K. Agiannitopoulos, G. Pepe, S. Kampouri, N. Diamantopoulos, T. Floros, R. Iosifidou, O. Katopodi, A. Koumarianou, C. Markopoulos, K. Papazisis, V. Venizelos, I. Xanthakis, G. Xepapadakis, E. Banu, D. T. Eniu, S. Negru, D. L. Stanculeanu, A. Ungureanu, V. Ozmen, S. Tansan, M. Tekinel, S. Yalcin and G. Nasioulas (2019). "Analysis of hereditary cancer syndromes by using a panel of genes: novel and multiple pathogenic mutations." BMC Cancer **19**(1): 535.
van der Merwe, N. C., N. Hamel, S. R. Schneider, J. P. Apffelstaedt, J. T. Wijnen and W. D. Foulkes (2012). "A founder BRCA2 mutation in non-Afrikaner breast cancer patients of the Western Cape of South Africa." Clin Genet **81**(2): 179-184.
Wainstein, T., R. Kerr, C. L. Mitchell, S. Madaree, F. B. Essop, E. Vorster, R. Wainwright, J. Poole and A. Krause (2013). "Fanconi anaemia in black South African patients heterozygous for the FANCG c.637-643delTACCGCC founder mutation." S Afr Med J **103**(12 Suppl 1): 970-973.
Wang, X., C. Krishnan, E. P. Nguyen, K. J. Meyer, J. L. Oliveira, P. Yang, E. S. Yi, M. R. Erickson-Johnson, M. J. Yaszemski, A. Maran and A. M. Oliveira (2012). "Fusion of dynactin 1 to anaplastic lymphoma kinase in inflammatory myofibroblastic tumor." Hum Pathol **43**(11): 2047-2052.
Weber, W., D. J. Nash, A. G. Motulsky, M. Henneberg, M. H. Crawford, S. K. Martin, J. M. Goldsmid, G. Spedini, S. Glidewell and M. S. Schanfield (2000). "Phylogenetic relationships of human populations in sub-Saharan Africa." Hum Biol **72**(5): 753-772.
Whitney, M. A., H. Saito, P. M. Jakobs, R. A. Gibson, R. E. Moses and M. Grompe (1993). "A common mutation in the FACC gene causes Fanconi anaemia in Ashkenazi Jews." Nat Genet **4**(2): 202-205.
Wiesner, T., W. Lee, A. C. Obenauf, L. Ran, R. Murali, Q. F. Zhang, E. W. Wong, W. Hu, S. N. Scott, R. H. Shah, I. Landa, J. Button, N. Lailler, A. Sboner, D. Gao, D. A. Murphy, Z. Cao, S. Shukla, T. J. Hollmann, L. Wang, L. Borsu, T. Merghoub, G. K. Schwartz, M. A. Postow, C. E. Ariyan, J. A. Fagin, D. Zheng, M. Ladanyi, K. J. Busam, M. F. Berger, Y. Chen and P. Chi (2015). "Alternative transcription initiation leads to expression of a novel ALK isoform in cancer." Nature **526**(7573): 453-457.
Witkiewicz, A. K. and E. S. Knudsen (2014). "Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine, and therapeutic interventions." Breast Cancer Res **16**(3): 207.
Wlodarska, I., C. De Wolf-Peeters, B. Falini, G. Verhoef, S. W. Morris, A. Hagemeijer and H. Van den Berghe (1998). "The cryptic inv(2)(p23q35) defines a new molecular genetic subtype of ALK-positive anaplastic large-cell lymphoma." Blood **92**(8): 2688-2695.
Wong, D. W., E. L. Leung, S. K. Wong, V. P. Tin, A. D. Sihoe, L. C. Cheng, J. S. Au, L. P. Chung and M. P. Wong (2011). "A novel KIF5B-ALK variant in nonsmall cell lung cancer." Cancer **117**(12): 2709-2718.
Yamamoto, H., K. Kohashi, Y. Oda, S. Tamiya, Y. Takahashi, Y. Kinoshita, S. Ishizawa, M. Kubota and M. Tsuneyoshi (2006). "Absence of human herpesvirus-8 and Epstein-Barr virus in inflammatory myofibroblastic tumor with anaplastic large cell lymphoma kinase fusion gene." Pathol Int **56**(10): 584-590.
Yao, S., M. Cheng, Q. Zhang, M. Wasik, R. Kelsh and C. Winkler (2013). "Anaplastic lymphoma kinase is required for neurogenesis in the developing central nervous system of zebrafish." PLoS One **8**(5): e63757.
Yuan, B., Y. Xu, J. H. Woo, Y. Wang, Y. K. Bae, D. S. Yoon, R. P. Wersto, E. Tully, K. Wilsbach and E. Gabrielson (2006). "Increased expression of mitotic checkpoint genes in breast cancer cells with chromosomal instability." Clin Cancer Res **12**(2): 405-410.

Zeng, H., Y. Liu, W. Wang, Y. Tang, P. Tian and W. Li (2021). "A rare KIF5B-ALK fusion variant in a lung adenocarcinoma patient who responded to crizotinib and acquired the ALK L1196M mutation after resistance: a case report." Ann Palliat Med **10**(7): 8352-8357.
Zhuang, L., Z. Yang and Z. Meng (2018). "Upregulation of BUB1B, CCNB1, CDC7, CDC20, and MCM3 in Tumor Tissues Predicted Worse Overall Survival and Disease-Free Survival in Hepatocellular Carcinoma Patients." Biomed Res Int **2018**: 7897346.

# Chapter 5

# Analysis of non-coding variants

## 5.1 INTRODUCTION

Many studies have focussed on the observation of coding variants in cancer. Functionally, this is a logical thinking process to follow because of the inclusion of genes in coding sequences (CDS). Genes function as blueprints which are used in the production of proteins. While the effect of non-coding mutations only became apparent with the production of whole genome sequencing of cancer genomes (Juul, Bertl et al. 2017), only a few studies have attempted to identify non-coding variants and their particular focus were on highly recurrent variants (positive selection) (Khurana, Fu et al. 2013, Weinhold, Jacobsen et al. 2014, Lochovsky, Zhang et al. 2015) or those that were flagged as deleterious/high impacting (Mularoni, Sabarinathan et al. 2016, Hornshøj, Nielsen et al. 2018).

Efforts have been made to identify areas of non-coding regions that may play a pivotal role in gene functionality (Brandler, Antaki et al. 2018, Pena, Jiang et al. 2018, Short, McRae et al. 2018). Different non-coding regions may include regulatory regions (Brandler, Antaki et al. 2018, Pena, Jiang et al. 2018) and/or changes in 3D conformation of proteins (Zhang and Lupski 2015, Short, McRae et al. 2018). As with coding regions, non-coding variants are identified and rated on how their presence affect the viability of the individual (Wells, Heckerman et al. 2019). Genome-wide epigenomic maps have revealed thousands of non-coding elements which contains signatures synonym with enhancers, gene-regulatory elements, and promoters, which may be possible effectors of interest (Kundaje, Meuleman et al. 2015).

Research has identified various interesting variants in the non-coding region of the genome. Non-coding variants with a high penetrance may be responsible for tumorigenesis by itself (Horn, Figl et al. 2013), on the other hand those with low penetrance may only influence somatic variants on a smaller scale (Easton and Eeles 2008).

A recent study by Gan *et al*. listed transcriptional and post-translational gene regulation as important events that may be affected by non-coding variants (Gan, Carrasco Pro et al. 2018). Some of the affected regions include transcription factors (TFs), both untranslated regions (UTRs) (Schuster and Hsieh 2019), binding of microRNAs (miRNAs) and RNA binding proteins (RBPs) (Shuai, Suzuki et al. 2019, Suzuki, Kumar et al. 2019) and affecting normal splicing mechanisms (Figure 5.1).

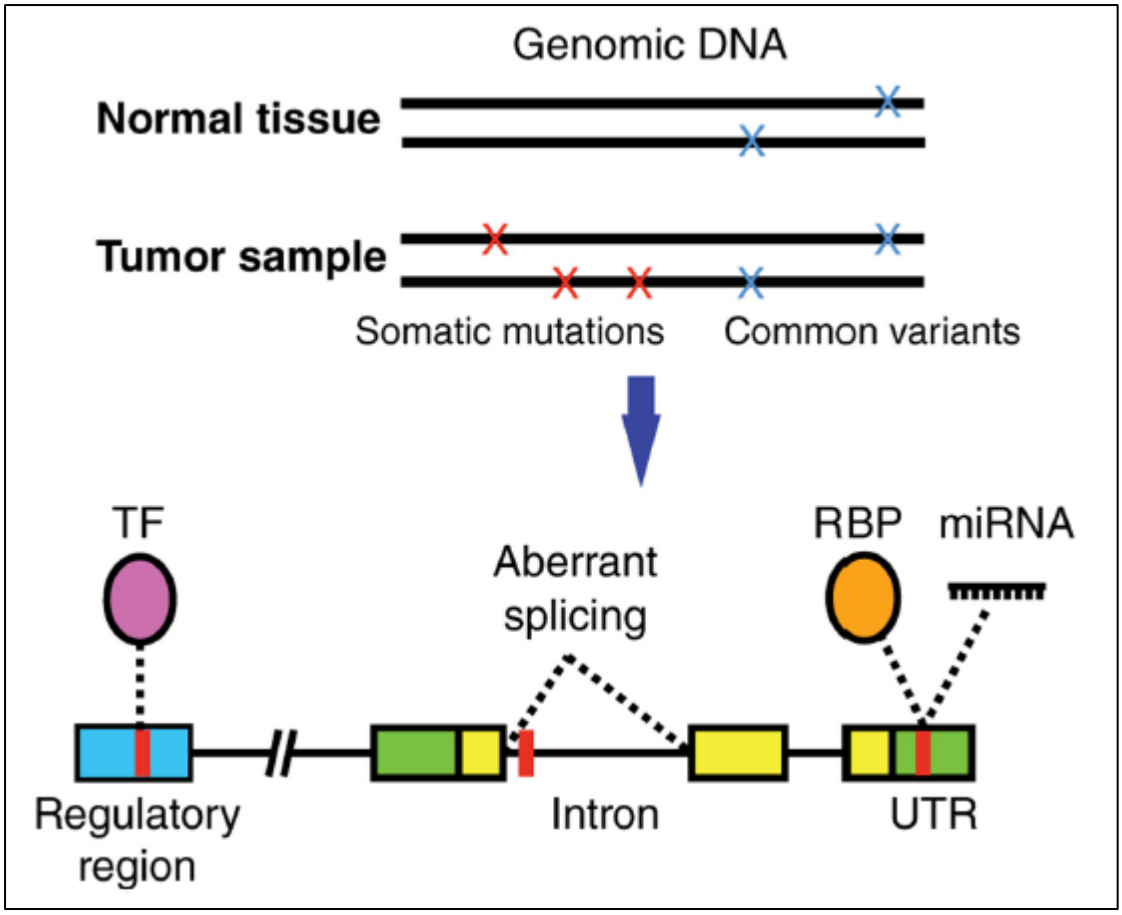


Figure 5.1: Somatic mutations can affect various regulatory mechanisms in non-coding regions of the genome (Gan, Carrasco Pro et al. 2018).

Other regions that may play important roles in cancer development could include, changes in enhancer sequences (Corona, Seo et al. 2020) and mutations that may disrupt chromatin domain structure (Hnisz, Weintraub et al. 2016). The goal is to identify driver rather than passenger mutations. Driver mutations are classified as mutations that lead to an advantageous cell growth and proliferation for tumour cells (Cuykendall, Rubin et al. 2017). Passenger mutations on the other hand are mutations that have no or a very small effect on the fitness of the cancer cell but are rather produced as a by-product of clonal expansion or genomic instability brought on by driver mutations (Vogelstein, Papadopoulos et al. 2013). Research into cancer has identified the TERT promoter as one of the most important non-coding regulatory drivers in many cancer types (Huang, Hodis et al. 2013, Vinagre, Almeida et al. 2013).

Other recurrent mutations in promoter regions around KIAA0907, SDHD, TBC1D12, WDR74 and YAE1D1 has also been linked to different cancers (Weinhold, Jacobsen et al. 2014, Araya, Cenik et al. 2016). Non-coding mutations around the DHX34 and TUBBP5 genes have respectively been identified in conjunction with lymphomas and liver cancers (Zhang, Bojorquez-Gomez et al. 2018).

Different studies around the oncogene, LIM Domain Only 1 (LMO1), have shown the presence of non-coding variants close to the transcription start site which increase the transcription of the gene 120 fold (Hu, Qian et al. 2017) and another variant in the first intron that changes the GATA transcription factor binding, converting the enhancer into a super enhancer for LMO1 (Oldridge, Wood et al. 2015). Closely related, LMO2, has also been implicated in T-cell acute lymphoblastic leukaemia (Fisch, Boehm et al. 1992, Hu, Qian et al. 2017). Up-regulation of this gene has been seen in T-cell leukaemia and researchers postulates that LMO2 upstream variants may play a important role (Elliott and Larsson 2021).

Challenges for identifying driver mutations in non-coding regions include: 1) a far larger set of mutations to confront when comparing non-coding to coding regions, 2) non-coding regions, for the most part, are incompletely annotated, 3) Non-coding regions contain areas of complex network regulatory cassettes and identification of driver mutations are based on CDS which function differently (Cuykendall, Rubin et al. 2017), 4) differentiating between driver and passenger mutations is complicated and there are far more passenger than driver mutations present in any one genome (Marx 2014), 5) moreover researchers still do not completely understand the logic of regulatory element functions (Gan, Carrasco Pro et al. 2018).

Castro-Giner *et al*. proposed an interesting concept that driver mutations may be broken down further into two groups, major and mini drivers (Castro-Giner, Ratcliffe et al. 2015). Here major drivers have a large impact while mini drivers only convey a small advantage to the cancer cell. They hypothesize that non-coding regions in fact harbour large amounts of mini drivers.

Some studies have focussed on multi-exonic non-coding RNA's (mencRNA) . These RNA's have been identified as an important alternative pathway to identify breast cancer risk (Tan, Biasini et al. 2018, Moradi Marjaneh, Beesley et al. 2020). The first step was to link any credible causal variants (CCVs) to all breast cancer susceptibility regions, this was done using stepwise multi-nomial logistic regression (Fachal, Aschard et al. 2020). Marjaneh *et al.* more closely examined these CCVs and found a set that are not linked to any regulatory regions but still had an effect. Targeted RNA sequencing and de novo transcript assembly researchers identified mencRNA's as breast cancer risk candidate genes (Moradi Marjaneh, Beesley et al. 2020). Further studies also concluded that long non-coding RNA's (lncRNA) could be a new potentially useful therapeutic target for breast cancer therapies (Moradi Marjaneh, Beesley et al. 2020). These are RNA transcripts longer than 200 bp in length which are found in intronic, intergenic and overlapping transcripts (St Laurent, Wahlestedt et al. 2015). LncRNA's have been implicated in regulation of known oncogenes (Schmitt and Chang 2016, Slack and Chinnaiyan 2019).

Various tools and methods have been developed to investigate non-coding cancer mutations. Combined Annotation Dependent Depletion (CADD) is a versatile tool used to test the level of deleteriousness of single nucleotide variants as well as insertions/deletions. The tool has the added benefit of being able to recognize both coding and non-coding variants (Rentzsch, Witten et al. 2019). The model is intentionally not trained on small variant datasets with known pathogenic status but rather on larger, less biased datasets. This helps the model to identify historical fixed variants as leaning more towards being benign/neutral rather than harmful. Each iteration of the model is trained on hundreds of genomic features and over thirty million variants. Because of the larger dataset used by CADD, it doesn't suffer any biases causes by using curated datasets of pathogenic and benign datasets (Stenson, Ball et al. 2003, Landrum, Lee et al. 2014).

Back in 2015, researchers proposed a newly developed machine learning approach to identify coding as well as non-coding variants, Functional Analysis through Hidden Markov Models (FATHMM-MKL) (Shihab, Rogers et al. 2015). At the time, only GWAVA (Ritchie, Dunham et al. 2014) and CADD (Kircher, Witten et al. 2014) were able to predict functional consequences of non-coding variants and in this study, FATHMM-MKL was found to outperform the other two approaches (Shihab, Rogers et al. 2015). FATHMM-MKL uses functional annotations from the Encyclopaedia of DNA Elements (ENCODE) consortium and nucleotide-based HMMs to predict variants.

*OncodriveFML* is a framework for detecting driver mutations in coding and non-coding regions. It analyses patterns of somatic mutations in both regions to identify positively selected mutations. Using this method, researchers were able to identified known non-coding driver mutations (Mularoni, Sabarinathan et al. 2016). Another framework, *LARVA*, tries to take limited non-coding functional annotation and overdispersion of in mutation count into account by using a comprehensive set of noncoding functional elements and modelling their mutation counts with a β-binominal distribution. Moreover, *LARVA* also concentrates on a genomic feature: replication timing, for an increased confidence in local mutation rates and mutational hotspots (Lochovsky, Zhang et al. 2015).

It is important to realize that the cancer panel used for sequencing of 94 genes used in this study focussed on coding regions, however, a brief summary of non-coding variants that were sequenced as incidentally-included regions is provided in this chapter. It should be kept in mind that most con-coding regions are not under the same restraints as coding regions and may show large numbers of mutations.

## 5.2 MATERIALS AND METHODS

Variant calling was handled by GATK HaplotypeCaller as previously discussed in Chapter 2. GATK has its own build-in variant annotation but annotation was further fine-tuned with CADD PHRED and FATHMM-MKL.

Functional annotation emphasis was placed on the annotations called by the ClinSIG (ClinVar clinical significance) column in VEP specifically. CADD PHRED and FATHMM-MKL models were further used to identify detrimental variants because of their unique ability to also predict non-coding variant significance. Focus was placed on variants that were called to be detrimental by CADD PHRED and FATHMM-MKL.

## 5.3 AIM

The aim of this chapter was to briefly report on non-coding variants that were detected in non-coding regions, and that may affect breast cancer risk in patients. These variants may by themselves may not cause a substantial increase in cancer risk but rather be passenger mutations or cause a small accumulative effect.

## 5.4 RESULTS

Initially, a total of 16279 variants were identified by VEP which included all transcript forms for the genes in the panel. We were specifically interested in the following non-coding variant consequences: 3 prime UTR, 5 prime UTR, downstream, upstream, intron, non-coding exon/transcript, TF binding site, intergenic and regulatory region variants.

Table 5.1: Initial non-coding variants identified.

| **Consequence On** | **Variant total** |
|---|---|
| 3_prime_UTR_variant | 82 |
| 5_prime_UTR_variant | 366 |
| Downstream_gene_variant | 4 095 |
| Intergenic_variant | 166 |
| Intron_variant | 1 155 |
| Intron_variant,non_coding_transcript_variant | 413 |
| Non_coding_transcript_exon_variant | 2 569 |
| Regulatory_region_variant | 3 214 |
| TF_binding_site_variant | 21 |
| Upstream_gene_variant | 4 198 |
| **Total** | 16 279 |

We further decreased these numbers by only using variants that had CADD PHRED scores > 30 and FATHMM-MKL scores > 0.5 and we only reported on canonical transcripts thereby removing multiple transcripts. After analysis with FATHMM-MKL, a set of 8 329 variants were identified as being deleterious, adding CADD PHRED scores left us with fourteen mutations. The variant annotation with CADD prediction uses a similar set of consequences and the numbers were grouped in table 5.2.

Table 5.2: Detrimental variants according to CADD PHRED.

| **Consequence** | **Variant total** |
|---|---|
| Regulatory_region_variant | 3 |
| TF_binding_site_variant | 1 |
| Non_coding_transcript_exon_variant | 7 |
| Upstream_gene_variant | 1 |
| Downstream_gene_variant | 1 |
| 5_prime_UTR_variant | 1 |
| **Total** | 14 |

A total of eight genes were affected by these detrimental variants. Significant cancer genes identified included: ATM, BRCA1, CHEK2 and CDH1. Almost all variants identified were filtered out because of an insignificant CADD PHRED score leaving us with only a handful of possible interesting variants.

**chr3:14200382: G > T (rs74737358)**

A variant linked to Xeroderma pigmentosum (XP) was previously common to a cell line with an increased susceptibility to UV rays (Li, Bales et al. 1993). A patient with the same variant in this study was found to carry XP-associated neurologic abnormalities which may be attributed to the variant (Li, Bales et al. 1993). The latest *Clinvar* record classifies this variant as benign, previous classifications have been updated through the use of *Sherloc* (semiquantitative, hierarchical evidence-based rules for locus interpretation), which is a variant classification framework to refine/updated older classifications of known variants (Nykamp, Anderson et al. 2017).

**chr1:17380483: C > T (rs111430410):**

The affected gene, succinate dehydrogenase complex iron sulphur subunit B (SDHB) plays a role in oxidation of succinate and has been linked to different forms of cancer: renal cell carcinoma, paragangliomas and pheochromocytoma (Ricketts, Woodward et al. 2008, Henderson, Douglas et al. 2009). Previously thought to be pathogenic, the latest belief is that this variant delivers a benign change to the gene through the use of *Sherloc* (Nykamp, Anderson et al. 2017).

**chr1:17380507: G > C (rs11203289)**

Also found within SDHB, reports mostly interpret this variant as a benign change. As with the above-mentioned variant, this is related to benign paragangliomas and pheochromocytomas. Of note is the linkage this variant has with Cowden syndrome but with an uncertain significance which may warrant further investigation (Ni, Zbuk et al. 2008).

**chr2:96930912: C > T (rs121908819)**

This variant simultaneously affect two different genes, cytosolic iron-sulphur assembly component 1 (CIAO1) on the forward strand (upstream variant) and transmembrane protein 127 (TMEM127) on the reverse strand (coding missense variant). Currently both variants have been classified as being benign by *Clinvar*. Work done previously has shown that the reverse strand variant may have a pathogenic effect causing an amino acid change from aspartic acid to asparagine (Yao, Schiavi et al. 2010).

**chr3:14190232: C > G (rs754673606)**

The current belief is that this variant conveys a pathogenic change according to *Clinvar*. The protein, Xeroderma pigmentosum, complementation group C (XPC), plays an important role in DNA damage sensing and binding to said DNA (Bernardes de Jesus, Bjørås et al. 2008, Sugasawa 2008). The position of this variant lies in a splice acceptor region and the change may disrupt RNA splicing, that leads to an incomplete protein and/or the absence of the XPC protein (Cartault, Nava et al. 2011). This variant has been linked to the development of Xeroderma pigmentosum (XP), a condition which leaves the carrier with increased probability of contracting skin cancers and a heightened sensitivity to ultraviolet rays because of the defect in DNA repair (Cartault, Nava et al. 2011). Of our analysed variants, this is the first one to be recognized as being pathogenic.

**chr3:52441251: A > G (rs143901408)**

The affected gene here is a well know cancer related gene, BRCA1 associated protein 1 (*BAP1*). Here, the variant change to a ‘C’ allele causes the amino acid change from tyrosine to a stop codon which is very detrimental to any protein production. A variety of cancers has been linked to BAP1 including, uveal melanoma (UM), cutaneous melanoma (CM), mesothelioma (MMe), and renal cell carcinoma (RCC). However,s the ‘G’ allele from this study leads to a benign clinical change, also no clear links have been made between BAP1 variants and breast cancer diagnosis previously.

**chr9:35077263: TGGCGGTA deletion (rs587776640)**

This specific deletion has not been reported before but enough evidence exists to show that a deletion/duplication (frameshift) in this area of the genome may have a pathogenic effect on the carrier (Morgan, Essop et al. 2005, Feben, Kromberg et al. 2015). Strong evidence links deletions in this area to the development of Fanconi anaemia (FA) (Morgan, Essop et al. 2005, Feben, Kromberg et al. 2015). Previous studies have shown that the black population of southern Africa may be the carriers of a founder mutation for FA (c.637_643delTACCGCC), researchers found the occurrence of the deletion to be as high as 1 in 40 000 births for the population (Morgan, Essop et al. 2005, Feben, Kromberg et al. 2015). This founder effect warrants further analysis of this unknown variant.

**chr9:98231100 : G > A (rs115556836)**

The variant occurs on the forward strand of chromosome 9 where the PTCH1 gene is located. Mutations in this gene have previously been linked to Gorlin syndrome or basal cell carcinoma which is a form of skin cancer (Okamoto, Naruto et al. 2014). Previously, this reference SNP has been linked to holoprosencephaly (Ming, Kaupas et al. 2002), which may warrant further studies in a different field than cancer research. But more recent information regarding the variant from this study points to a benign change with no pathogenic effect, both *ClinVar* and *Ensembl* regards this variant as non-pathogenic.

**chr13:103527930: G > C (rs9514067)**

Linked to the ERCC, a gene known to form part of a nucleotide excision repair pathway, the variant may play some significant role in cancer development. This variant was previously reported (Bodian, McCutcheon et al. 2014) but the clinical significance of the variant still remains unclear. Further research is required to identify the importance of this variant relative to cancer.

**chr16:68856041: G > A (rs33935154)**

Back in 2003, *ClinVar* recognized this coding transcript variant as being pathogenic, a protein change from alanine to threonine (Suriano, Oliveira et al. 2003). More recent research has revaluated this position and its consequences and found that it has no clinical significance (Nykamp, Anderson et al. 2017, Lee, Krempely et al. 2018). The non-coding transcript linked to this position has not been linked to any known disease and may have no effect but a in depth study will be required to verify this.

**chr17:41215947: C > T (rs41293459)**

On a molecular level, this variant is linked to a BRCA1 missense coding transcript as well as a non-coding transcript variant. The missense variant, c.5096G>A (p.Arg1699Gln) has been identified

multiple times in previous years and is believed to be linked to the development of breast and ovarian cancer (Spurdle, Whiley et al. 2012, Moghadasi, Meeks et al. 2018, Keupp, Hampp et al. 2019, Sepahi, Faust et al. 2019). The significance of the non-coding variant is unclear.

**Chr17:41226499: C > T (rs80356885)**

As with the previous BRCA1 variant, the non-coding variant is a transcript allele of a pathogenic genomic coding transcript allele. Where the exon variant, NM_007294.4:c.4524G>A (p.Trp1508Ter), has a high clinical significance (Laitman, Borsthein et al. 2011, Walsh, Casadei et al. 2011).

**chr22:29130427: G > A (rs587781269)**

The variant allele is linked to an important coding transcript stop-gained mutation, NM_007194.4(CHEK2):c.283C>T. However, the non-coding variant by itself may have no significance.

**chr22:29130456: G > A (rs17883862)**

All previous *ClinVar* calls were uncertain significance but the latest report of the variant classifies it to be a benign change (Nykamp, Anderson et al. 2017).

## 5.5 DISCUSSION

We set out to identify non-coding variants that may be linked to breast cancer significance. We identified variants from regions excluding exon regions, these included: upstream, downstream, intergenic introns and introns. We used GATK HaplotypeCaller to call a large number of variants, which was reduced during variant annotation with VEP and further with a combination between CADD-PHRED and FATHMM-MKL predictions. In the end, a total of fourteen significant non-coding pathogenic variants were identified, this is a relatively small number but is understandable in the context of only using cancer panel sequencing which removed a very large portion of non-coding regions which was left unexplored. The impact of most of these variants are untested and could only be predicted with model metrics if present, which means that these variants are difficult to call confidently.

We were able to identify non-coding variants which are in proximity or linked to important cancer genes. These genes include the likes of ATM, BRCA1, CHEK2 and CDH1. The most significant variant identified in the study was: chr3:14190232: C > G, which may be a founder mutation in the South African black population (Cartault, Nava et al. 2011). Unfortunately, most of these identified variants were either classified incorrectly as pathogenic in the past or follow-up research disclaimed their

pathogenicity, or more up-to-date modelling focussed on non-coding variants has more accurately classified the severity of these variants.

On closer examination of the fourteen pathogenically linked variants identified in this category, only two were still deemed to be pathogenic and one other as an unknown clinical significance. The chr3:14190232: C > G variant has been identified and linked to cancer previously (Cartault, Nava et al. 2011, Fassihi, Sethi et al. 2016, Kgokolo, Morice-Picard et al. 2019). Researchers believe that because of the heritage of Mahori people, it is well worth to study African populations from where this variant may have originated (Cartault, Nava et al. 2011). This is only the second time that this variant has been identified in the SA population after the study done by Kgokolo *et al*. (2019). A more in-depth study into Xeroderma pigmentosum in black South Africans will further shed light on the possibility of a South African XPC founder mutation.

Regarding the chr9:35077263: TGGCGGTA deletion, evidence exist that deletions in this region of the genome may have some impact on cancer development (Weber, Nash et al. 2000, Morgan, Essop et al. 2005, Wainstein, Kerr et al. 2013). By regarding our study as well as research done by both Weber *et al*. (2000) and Morgan *et al*. (2005) a pattern emerges where all three studies focussed on sub-Saharan populations. They believe that the Bantu-speaking population of southern Africa may be carriers of a FANCG founder mutation and even more research followed in 2013 (Wainstein, Kerr et al. 2013) to reiterate this founder mutation. These deletions are linked to the development of Fanconi anaemia, heterogeneous disorder, known to cause crosslink-induced chromosome breaks. The world-wide prevalence of the disease is relatively small (1/300 000) while the South Africa black population has a much higher prevalence (1/40 000) (Wainstein, Kerr et al. 2013). Here, we also identified a FANCG deletion in the same genome space as other researchers. This is an important find and strengthens the FANCG founder mutation idea.

Because of the connection that the chr13:103527930: G > C may have with the ERCC gene (DNA repair) it may be beneficial to research the effect of this variant because not enough information is available to make a proper evaluation of its clinical significance even though it has been reported previously (Bodian, McCutcheon et al. 2014).

Interestingly, one variant identified to a gene symbol: AC093495.4 has previously been found to be a differentially expressed long non-coding RNA (lnc-RNA) and may be linked to cholesteatoma pathogenesis (Gao, Tang et al. 2018). This paper set out to link the aberrant expression of long non-

coding RNAs in disease tissue versus normal tissue and identified a large set of lnc-RNAs that may influence the disease phenotype. The question still needs to be asked which of these variants are driver mutations. Identified variants may in future still only be identified as passenger mutations or random uninteresting variants.

We are aware of the high number of down- and upstream variants present in our study to which we have no clear answer. The Illumina TruSight Cancer panel which was used in this study covers 94 different genes and on average 50 bp down- and upstream of the exon regions. We can only speculate that firstly, in the absence of properly sequenced and documented non-coding regions of South African populations many of these positions may in fact just be normal in a South African population context and variants should be ignored, secondly, some of these regions may purely be non-coding and have no link to enhancer or exon-affecting regions and could accumulate variants which would never have a negative effect on the genome.

Recently, more studies regarding non-coding cancer variants have sprung up. One of these studies was done on the functional impact of non-coding variants, where the authors were able to identify a very impactful TERT variant, which increases the expression of the TERT (telomerase reverse transcriptase) gene. This increase plays an important role in inducing cell transformation and immortality. The authors identified another 54 additional candidates for driver mutations in non-coding regions (Mularoni, Sabarinathan et al. 2016). In this study we were not able to identify any variant linked to this gene.

The importance of the role of intronic regions in cancer is in most cases overlooked because of their absence in mature mRNAs. Studies have shown that variants in the intronic regions may lead to the alternative splice sites which leads to an alternative mature mRNA (Wang and Burge 2008). SpliceAI, an under construction deep residual neural network, has been developed to identify splice site around exonic regions up to 10 000 nucleotides far (Jaganathan, Kyriazopoulou Panagiotopoulou et al. 2019). SpliceAI has the ability to play a very important role in future non-coding studies but only when further research increases its current 41% sensitivity for variants >50 nucleotides from exon-intron boundaries.

Other studied non-coding driver mutations were found in GATA3, which in most study cases carries an indel on intron four which leads to an incorrect acceptor splice site (Hornshøj, Nielsen et al. 2018), none of these variants were identified here. As previously noted in Chapter 4, mutations in MSH6 and PMS2 warrant further research because of relatively new information that show their increased involvement in cancer development (Roberts, Jackson et al. 2018). Adding to the

importance of non-coding variants, researchers were able to identify that a simple intronic variant in the BRCA1 gene gave way to the formation of a partial exon which may have played a role in cancer formation in patients from France, Norway and USA (Høberg-Vetti, Ognedal et al. 2020).

A different study by Juul *et al*. identified sequence cassettes (DNA stretches) in non-coding regions, where mutations gave the development of cancer cells an advantage. The authors dubbed these stretches 'driver elements' (Juul, Bertl et al. 2017). They were able to identify known 'driver elements' and also a possible novel antigen-presenting gene (CD1A), where a mutated 5'UTR correlated well with a decreased survival in melanoma patients and mutations in the base-excision-repair gene (SMUG1) coincided with C-to-T mutational-signatures (Juul, Bertl et al. 2017).

But why has so little research been done on non-coding mutations? Most importantly, the focus has always been on coding regions which codes for functional proteins while up until recently non-coding regions were believed to play a very small role in functional biology of DNA. Coding regions have exhaustively been sequenced which has led to a well understood model while this is much less the case with non-coding regions. The underrepresentation of sequenced non-coding regions on the other hand now points to our lack of understanding and how complex it is. This again, shows that new models are required with non-coding elements in mind.

Different models exist which could be used to predict the importance of non-coding variants to help researchers identify areas of interest within the non-coding genome. These models are flawed in that they are inclined to focus on regulatory elements with good reason though, to date three quarters of disease-causing variants have been found there (Ritchie, Dunham et al. 2014). In the current study we used the CADD system which utilizes allelic frequencies based on machine learning patterns linked with genomic features (Kircher, Witten et al. 2014).

CADD has its own limitations, one of them being its inability to perfectly approximate a variant as pathogenic or benign (Landrum, Lee et al. 2014). This is overcome by using multiple prediction models and only select variants that are quantified the same by most of the models. A few studies have already used CADD prediction for non-coding variant analysis to some extent of success (Kircher, Witten et al. 2014, Mather, Mooney et al. 2016). Both these studies found that CADD had a low positive predictive value in identifying non-coding variants and should rather be used in conjunction with other models. While FATHMM-MKL has the option to predict both coding and non-coding variants, researchers concluded that the approach did relatively well when non-coding variants were examined but it could be improved by the inclusion of feature selection. Combinations

of feature groups could be used to more accurately predict variant significance (Shihab, Rogers et al. 2015).

Other models in circulation include an empirical scoring system, *Funseq*, which first identify sensitive regions of the non-coding genome using the 1 000 genome project data, followed by scoring variant impact in said regions (Khurana, Fu et al. 2013). Another shortfall of these models is that they do not take different tissues into account, mutations may have varying effects in different tissues (Ritchie, Dunham et al. 2014). More recently Ritchie *et al.* purposed a model which through machine learning also identifies genome regions of interest which may carry disease-causing variants but using databases of disease-causing variants (Ritchie, Dunham et al. 2014). However, according to Elliott and Larsson, 2021, all models are wrong, the mutation rate expectation models could not really reflect reality (Elliott and Larsson 2021). The understanding of genomic mutational rate heterogeneity is still in its infancy and with future studies could only improve with deeper sequencing of non-coding regions (Elliott and Larsson 2021).

To date, the most effective strategy for validating the functionality of mutations are through experimental procedures. Exciting methods like CRISPR can be used for chromatin topology manipulation, epigenome editing, gene expression regulation and live-cell chromatin imaging (Adli 2018). CRISPR has successfully been used in the past to link enhancers to candidate genes, and further combining CRISPR with other technologies such as single-cell RNA-seq (Xie, Duan et al. 2017), ATAC-seq, methylation profiling and Hi-C would greatly benefit our understanding of functional non-coding elements.

New research into model construction has surfaced that specifically focusses on elucidating detrimental non-coding variants (Li, Zhang et al. 2020). Here, the researchers were able to positively identify non-coding variants in enhancer regions for KRAS and PER2 genes. Upregulation of these genes have been implicated in human leukemia (Li, Zhang et al. 2020). This model may be useful in the future to identify other important regulatory region in the human genome.

Much work is still needed to validate non-coding mutations for their role in breast cancer and the conclusion is similar for this study. A more comprehensive study of the South African breast cancer population is required for a clearer picture, where whole genome sequencing is required for a more comprehensive overview of unknown variants and non-coding regions still unexplored.

## 5.6 REFERENCES

Adli, M. (2018). "The CRISPR tool kit for genome editing and beyond." Nature communications **9**(1): 1911-1911.

Araya, C. L., C. Cenik, J. A. Reuter, G. Kiss, V. S. Pande, M. P. Snyder and W. J. Greenleaf (2016). "Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations." Nature genetics **48**(2): 117-125.

Bernardes de Jesus, B. M., M. Bjørås, F. Coin and J. M. Egly (2008). "Dissection of the molecular defects caused by pathogenic mutations in the DNA repair factor XPC." Mol Cell Biol **28**(23): 7225-7235.

Bodian, D. L., J. N. McCutcheon, P. Kothiyal, K. C. Huddleston, R. K. Iyer, J. G. Vockley and J. E. Niederhuber (2014). "Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: implications for individual genome sequencing." PLoS One **9**(4): e94554.

Brandler, W. M., D. Antaki, M. Gujral, M. L. Kleiber, J. Whitney, M. S. Maile, O. Hong, T. R. Chapman, S. Tan, P. Tandon, T. Pang, S. C. Tang, K. K. Vaux, Y. Yang, E. Harrington, S. Juul, D. J. Turner, B. Thiruvahindrapuram, G. Kaur, Z. Wang, S. F. Kingsmore, J. G. Gleeson, D. Bisson, B. Kakaradov, A. Telenti, J. C. Venter, R. Corominas, C. Toma, B. Cormand, I. Rueda, S. Guijarro, K. S. Messer, C. M. Nievergelt, M. J. Arranz, E. Courchesne, K. Pierce, A. R. Muotri, L. M. Iakoucheva, A. Hervas, S. W. Scherer, C. Corsello and J. Sebat (2018). "Paternally inherited cis-regulatory structural variants are associated with autism." Science **360**(6386): 327-331.

Cartault, F., C. Nava, A. C. Malbrunot, P. Munier, J. C. Hebert, P. N'Guyen, N. Djeridi, P. Pariaud, J. Pariaud, A. Dupuy, F. Austerlitz and A. Sarasin (2011). "A new XPC gene splicing mutation has lead to the highest worldwide prevalence of xeroderma pigmentosum in black Mahori patients." DNA Repair (Amst) **10**(6): 577-585.

Castro-Giner, F., P. Ratcliffe and I. Tomlinson (2015). "The mini-driver model of polygenic cancer evolution." Nat Rev Cancer **15**(11): 680-685.

Corona, R. I., J. Seo, X. Lin, D. J. Hazelett, J. Reddy, M. A. S. Fonseca, F. Abassi, Y. G. Lin, P. Y. Mhawech-Fauceglia, S. P. Shah, D. G. Huntsman, A. Gusev, B. Y. Karlan, B. P. Berman, M. L. Freedman, S. A. Gayther and K. Lawrenson (2020). "Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer." Nature Communications **11**(1): 2020.

Cuykendall, T. N., M. A. Rubin and E. Khurana (2017). "Non-coding genetic variation in cancer." Current Opinion in Systems Biology **1**: 9-15.

Easton, D. F. and R. A. Eeles (2008). "Genome-wide association studies in cancer." Hum Mol Genet **17**(R2): R109-115.

Elliott, K. and E. Larsson (2021). "Non-coding driver mutations in human cancer." Nat Rev Cancer **21**(8): 500-509.

Fachal, L., H. Aschard, J. Beesley, D. R. Barnes, J. Allen, S. Kar, K. A. Pooley, J. Dennis, K. Michailidou, C. Turman, P. Soucy, A. Lemaçon, M. Lush, J. P. Tyrer, M. Ghoussaini, M. Moradi Marjaneh, X. Jiang, S. Agata, K. Aittomäki, M. R. Alonso, I. L. Andrulis, H. Anton-Culver, N. N. Antonenkova, A. Arason, V. Arndt, K. J. Aronson, B. K. Arun, B. Auber, P. L. Auer, J. Azzollini, J. Balmaña, R. B. Barkardottir, D. Barrowdale, A. Beeghly-Fadiel, J. Benitez, M. Bermisheva, K. Białkowska, A. M. Blanco, C. Blomqvist, W. Blot, N. V. Bogdanova, S. E. Bojesen, M. K. Bolla, B. Bonanni, A. Borg, K. Bosse, H. Brauch, H. Brenner, I. Briceno, I. W. Brock, A. Brooks-Wilson, T. Brüning, B. Burwinkel, S. S. Buys, Q. Cai, T. Caldés, M. A. Caligo, N. J. Camp, I. Campbell, F. Canzian, J. S. Carroll, B. D. Carter, J. E. Castelao, J. Chiquette, H. Christiansen, W. K. Chung, K. B. M. Claes, C. L. Clarke, J. M. Collée, S. Cornelissen, F. J. Couch, A. Cox, S. S. Cross, C. Cybulski, K. Czene, M. B. Daly, M. de la Hoya, P. Devilee, O. Diez, Y. C. Ding, G. S. Dite, S. M. Domchek, T. Dörk, I. Dos-Santos-Silva, A. Droit, S. Dubois, M. Dumont, M. Duran, L. Durcan, M. Dwek, D. M. Eccles, C. Engel, M. Eriksson, D. G. Evans, P. A. Fasching, O. Fletcher, G. Floris, H. Flyger, L. Foretova, W. D. Foulkes, E. Friedman, L. Fritschi, D. Frost, M. Gabrielson, M. Gago-Dominguez, G. Gambino, P. A. Ganz, S. M. Gapstur, J. Garber, J. A. García-Sáenz, M. M. Gaudet, V. Georgoulias, G. G. Giles, G. Glendon, A. K. Godwin, M. S. Goldberg, D. E. Goldgar, A.

González-Neira, M. G. Tibiletti, M. H. Greene, M. Grip, J. Gronwald, A. Grundy, P. Guénel, E. Hahnen, C. A. Haiman, N. Håkansson, P. Hall, U. Hamann, P. A. Harrington, J. M. Hartikainen, M. Hartman, W. He, C. S. Healey, B. A. M. Heemskerk-Gerritsen, J. Heyworth, P. Hillemanns, F. B. L. Hogervorst, A. Hollestelle, M. J. Hooning, J. L. Hopper, A. Howell, G. Huang, P. J. Hulick, E. N. Imyanitov, C. Isaacs, M. Iwasaki, A. Jager, M. Jakimovska, A. Jakubowska, P. A. James, R. Janavicius, R. C. Jankowitz, E. M. John, N. Johnson, M. E. Jones, A. Jukkola-Vuorinen, A. Jung, R. Kaaks, D. Kang, P. M. Kapoor, B. Y. Karlan, R. Keeman, M. J. Kerin, E. Khusnutdinova, J. I. Kiiski, J. Kirk, C. M. Kitahara, Y. D. Ko, I. Konstantopoulou, V. M. Kosma, S. Koutros, K. Kubelka-Sabit, A. Kwong, K. Kyriacou, Y. Laitman, D. Lambrechts, E. Lee, G. Leslie, J. Lester, F. Lesueur, A. Lindblom, W. Y. Lo, J. Long, A. Lophatananon, J. T. Loud, J. Lubiński, R. J. MacInnis, T. Maishman, E. Makalic, A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M. E. Martinez, K. Matsuo, T. Maurer, D. Mavroudis, R. Mayes, L. McGuffog, C. McLean, N. Mebirouk, A. Meindl, A. Miller, N. Miller, M. Montagna, F. Moreno, K. Muir, A. M. Mulligan, V. M. Muñoz-Garzon, T. A. Muranen, S. A. Narod, R. Nassir, K. L. Nathanson, S. L. Neuhausen, H. Nevanlinna, P. Neven, F. C. Nielsen, L. Nikitina-Zake, A. Norman, K. Offit, E. Olah, O. I. Olopade, H. Olsson, N. Orr, A. Osorio, V. S. Pankratz, J. Papp, S. K. Park, T. W. Park-Simon, M. T. Parsons, J. Paul, I. S. Pedersen, B. Peissel, B. Peshkin, P. Peterlongo, J. Peto, D. Plaseska-Karanfilska, K. Prajzendanc, R. Prentice, N. Presneau, D. Prokofyeva, M. A. Pujana, K. Pylkäs, P. Radice, S. J. Ramus, J. Rantala, R. Rau-Murthy, G. Rennert, H. A. Risch, M. Robson, A. Romero, M. Rossing, E. Saloustros, E. Sánchez-Herrero, D. P. Sandler, M. Santamariña, C. Saunders, E. J. Sawyer, M. T. Scheuner, D. F. Schmidt, R. K. Schmutzler, A. Schneeweiss, M. J. Schoemaker, B. Schöttker, P. Schürmann, C. Scott, R. J. Scott, L. Senter, C. M. Seynaeve, M. Shah, P. Sharma, C. Y. Shen, X. O. Shu, C. F. Singer, T. P. Slavin, S. Smichkoska, M. C. Southey, J. J. Spinelli, A. B. Spurdle, J. Stone, D. Stoppa-Lyonnet, C. Sutter, A. J. Swerdlow, R. M. Tamimi, Y. Y. Tan, W. J. Tapper, J. A. Taylor, M. R. Teixeira, M. Tengström, S. H. Teo, M. B. Terry, A. Teulé, M. Thomassen, D. L. Thull, M. Tischkowitz, A. E. Toland, R. Tollenaar, I. Tomlinson, D. Torres, G. Torres-Mejía, M. A. Troester, T. Truong, N. Tung, M. Tzardi, H. U. Ulmer, C. M. Vachon, C. J. van Asperen, L. E. van der Kolk, E. J. van Rensburg, A. Vega, A. Viel, J. Vijai, M. J. Vogel, Q. Wang, B. Wappenschmidt, C. R. Weinberg, J. N. Weitzel, C. Wendt, H. Wildiers, R. Winqvist, A. Wolk, A. H. Wu, D. Yannoukakos, Y. Zhang, W. Zheng, D. Hunter, P. D. P. Pharoah, J. Chang-Claude, M. García-Closas, M. K. Schmidt, R. L. Milne, V. N. Kristensen, J. D. French, S. L. Edwards, A. C. Antoniou, G. Chenevix-Trench, J. Simard, D. F. Easton, P. Kraft and A. M. Dunning (2020). "Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes." Nat Genet **52**(1): 56-73.

Fassihi, H., M. Sethi, H. Fawcett, J. Wing, N. Chandler, S. Mohammed, E. Craythorne, A. M. Morley, R. Lim, S. Turner, T. Henshaw, I. Garrood, P. Giunti, T. Hedderly, A. Abiona, H. Naik, G. Harrop, D. McGibbon, N. G. Jaspers, E. Botta, T. Nardo, M. Stefanini, A. R. Young, R. P. Sarkany and A. R. Lehmann (2016). "Deep phenotyping of 89 xeroderma pigmentosum patients reveals unexpected heterogeneity dependent on the precise molecular defect." Proc Natl Acad Sci U S A **113**(9): E1236-1245.

Feben, C., J. Kromberg, R. Wainwright, D. Stones, J. Poole, T. Haw and A. Krause (2015). "Hematological consequences of a FANCG founder mutation in Black South African patients with Fanconi anemia." Blood cells, molecules & diseases **54**(3): 270-274.

Fisch, P., T. Boehm, I. Lavenir, T. Larson, J. Arno, A. Forster and T. H. Rabbitts (1992). "T-cell acute lymphoblastic lymphoma induced in transgenic mice by the RBTN1 and RBTN2 LIM-domain genes." Oncogene **7**(12): 2389-2397.

Gan, K. A., S. Carrasco Pro, J. A. Sewell and J. I. Fuxman Bass (2018). "Identification of Single Nucleotide Non-coding Driver Mutations in Cancer." Frontiers in Genetics **9**: 16.

Gao, J., Q. Tang, X. Zhu, S. Wang, Y. Zhang, W. Liu, Z. Gao and H. Yang (2018). "Long noncoding RNAs show differential expression profiles and display ceRNA potential in cholesteatoma pathogenesis." Oncology reports **39**(5): 2091-2100.

Henderson, A., F. Douglas, P. Perros, C. Morgan and E. R. Maher (2009). "SDHB-associated renal oncocytoma suggests a broadening of the renal phenotype in hereditary paragangliomatosis." Fam Cancer **8**(3): 257-260.
Hnisz, D., A. S. Weintraub, D. S. Day, A. L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker and R. A. Young (2016). "Activation of proto-oncogenes by disruption of chromosome neighborhoods." Science **351**(6280): 1454-1458.
Høberg-Vetti, H., E. Ognedal, A. Buisson, T. B. A. Vamre, S. Ariansen, J. M. Hoover, G. E. Eide, G. Houge, T. Fiskerstrand, B. I. Haukanes, C. Bjorvatn and P. Knappskog (2020). "The intronic BRCA1 c.5407-25T>A variant causing partly skipping of exon 23—a likely pathogenic variant with reduced penetrance?" European Journal of Human Genetics **28**(8): 1078-1086.
Horn, S., A. Figl, P. S. Rachakonda, C. Fischer, A. Sucker, A. Gast, S. Kadel, I. Moll, E. Nagore, K. Hemminki, D. Schadendorf and R. Kumar (2013). "TERT promoter mutations in familial and sporadic melanoma." Science **339**(6122): 959-961.
Hornshøj, H., M. M. Nielsen, N. A. Sinnott-Armstrong, M. P. Świtnicki, M. Juul, T. Madsen, R. Sallari, M. Kellis, T. Ørntoft, A. Hobolth and J. S. Pedersen (2018). "Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival." NPJ Genom Med **3**: 1.
Hu, S., M. Qian, H. Zhang, Y. Guo, J. Yang, X. Zhao, H. He, J. Lu, J. Pan, M. Chang, G. Du, T. N. Lin, S. K. Kham, T. C. Quah, H. Ariffin, A. M. Tan, Y. Cheng, C. Li, A. E. Yeoh, C. H. Pui, A. J. Skanderup and J. J. Yang (2017). "Whole-genome noncoding sequence analysis in T-cell acute lymphoblastic leukemia identifies oncogene enhancer mutations." Blood **129**(24): 3264-3268.
Huang, F. W., E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin and L. A. Garraway (2013). "Highly recurrent TERT promoter mutations in human melanoma." Science **339**(6122): 957-959.
Jaganathan, K., S. Kyriazopoulou Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders and K. K. Farh (2019). "Predicting Splicing from Primary Sequence with Deep Learning." (1097-4172 (Electronic)).
Juul, M., J. Bertl, Q. Guo, M. M. Nielsen, M. Świtnicki, H. Hornshøj, T. Madsen, A. Hobolth and J. S. Pedersen (2017). "Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate." eLife **6**: e21778.
Keupp, K., S. Hampp, A. Hübbel, M. Maringa, S. Kostezka, K. Rhiem, A. Waha, B. Wappenschmidt, R. Pujol, J. Surrallés, R. K. Schmutzler, L. Wiesmüller and E. Hahnen (2019). "Biallelic germline BRCA1 mutations in a patient with early onset breast cancer, mild Fanconi anemia-like phenotype, and no chromosome fragility." Mol Genet Genomic Med **7**(9): e863.
Kgokolo, M., F. Morice-Picard, H. R. Rezvani, F. Austerlitz, F. Cartault, A. Sarasin, M. Sathekge, A. Taieb and C. Ged (2019). "Xeroderma pigmentosum in South Africa: Evidence for a prevalent founder effect." Br J Dermatol **181**(5): 1070-1072.
Khurana, E., Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. S. Evani, P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. H. Gumus, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. M. Lipkin, D. G. MacArthur, G. Marth, D. Muzny, T. H. Pers, G. R. S. Ritchie, J. A. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu, E. T. Dermitzakis, H. Yu, M. A. Rubin, C. Tyler-Smith and M. Gerstein (2013). "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science **342**(6154): 1235587.
Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper and J. Shendure (2014). "A general framework for estimating the relative pathogenicity of human genetic variants." Nat Genet **46**(3): 310-315.
Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A.

Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang and M. Kellis (2015). "Integrative analysis of 111 reference human epigenomes." Nature **518**(7539): 317-330.

Laitman, Y., R. T. Borsthein, D. Stoppa-Lyonnet, E. Dagan, L. Castera, M. Goislard, R. Gershoni-Baruch, H. Goldberg, B. Kaufman, N. Ben-Baruch, J. Zidan, T. Maray, L. Soussan-Gutman and E. Friedman (2011). "Germline mutations in BRCA1 and BRCA2 genes in ethnically diverse high risk families in Israel." Breast Cancer Res Treat **127**(2): 489-495.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church and D. R. Maglott (2014). "ClinVar: public archive of relationships among sequence variation and human phenotype." Nucleic Acids Res **42**(Database issue): D980-985.

Lee, K., K. Krempely, M. E. Roberts, M. J. Anderson, F. Carneiro, E. Chao, K. Dixon, J. Figueiredo, R. Ghosh, D. Huntsman, P. Kaurah, C. Kesserwan, T. Landrith, S. Li, A. R. Mensenkamp, C. Oliveira, C. Pardo, T. Pesaran, M. Richardson, T. P. Slavin, A. B. Spurdle, M. Trapp, L. Witkowski, C. S. Yi, L. Zhang, S. E. Plon, K. A. Schrader and R. Karam (2018). "Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants." Human mutation **39**(11): 1553-1568.

Li, K., Y. Zhang, X. Liu, Y. Liu, Z. Gu, H. Cao, K. E. Dickerson, M. Chen, W. Chen, Z. Shao, M. Ni and J. Xu (2020). "Noncoding Variants Connect Enhancer Dysregulation with Nuclear Receptor Signaling in Hematopoietic Malignancies." Cancer Discov **10**(5): 724-745.

Li, L., E. S. Bales, C. A. Peterson and R. J. Legerski (1993). "Characterization of molecular defects in xeroderma pigmentosum group C." Nat Genet **5**(4): 413-417.

Lochovsky, L., J. Zhang, Y. Fu, E. Khurana and M. Gerstein (2015). "LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations." Nucleic Acids Res **43**(17): 8123-8134.

Marx, V. (2014). "Cancer genomes: discerning drivers from passengers." Nat Methods **11**(4): 375-379.

Mather, C. A., S. D. Mooney, S. J. Salipante, S. Scroggins, D. Wu, C. C. Pritchard and B. H. Shirts (2016). "CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel." Genetics in medicine : official journal of the American College of Medical Genetics **18**(12): 1269-1275.

Ming, J. E., M. E. Kaupas, E. Roessler, H. G. Brunner, M. Golabi, M. Tekin, R. F. Stratton, E. Sujansky, S. J. Bale and M. Muenke (2002). "Mutations in PATCHED-1, the receptor for SONIC HEDGEHOG, are associated with holoprosencephaly." Hum Genet **110**(4): 297-301.

Moghadasi, S., H. D. Meeks, M. P. Vreeswijk, L. A. Janssen, Å. Borg, H. Ehrencrona, Y. Paulsson-Karlsson, B. Wappenschmidt, C. Engel, A. Gehrig, N. Arnold, T. V. O. Hansen, M. Thomassen, U. B. Jensen, T. A. Kruse, B. Ejlertsen, A. M. Gerdes, I. S. Pedersen, S. M. Caputo, F. Couch, E. J. Hallberg, A. M. van den Ouweland, M. J. Collée, E. Teugels, M. A. Adank, R. B. van der Luijt, A. R. Mensenkamp, J. C. Oosterwijk, M. J. Blok, N. Janin, K. B. Claes, K. Tucker, V. Viassolo, A. E. Toland, D. E. Eccles, P. Devilee, C. J. Van Asperen, A. B. Spurdle, D. E. Goldgar and E. G. García (2018). "The BRCA1 c. 5096G>A p.Arg1699Gln (R1699Q) intermediate risk variant: breast and ovarian cancer risk estimation and recommendations for clinical management from the ENIGMA consortium." J Med Genet **55**(1): 15-20.

Moradi Marjaneh, M., J. Beesley, T. A. O'Mara, P. Mukhopadhyay, L. T. Koufariotis, S. Kazakoff, N. I. Hussein, L. Fachal, N. Bartonicek, K. M. Hillman, S. H. Kaufmann, H. Sivakumaran, C. E. Smart, A. E.

McCart Reed, K. Ferguson, J. M. Saunus, S. R. Lakhani, D. R. Barnes, A. C. Antoniou, M. E. Dinger, N. Waddell, D. F. Easton, A. M. Dunning, G. Chenevix-Trench, S. L. Edwards and J. D. French (2020). "Non-coding RNAs underlie genetic predisposition to breast cancer." Genome biology **21**(1): 7-7.
Morgan, N. V., F. Essop, I. Demuth, T. de Ravel, S. Jansen, M. Tischkowitz, C. M. Lewis, L. Wainwright, J. Poole, H. Joenje, M. Digweed, A. Krause and C. G. Mathew (2005). "A common Fanconi anemia mutation in black populations of sub-Saharan Africa." Blood **105**(9): 3542-3544.
Mularoni, L., R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez and N. López-Bigas (2016). "OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations." Genome Biology **17**(1): 128.
Ni, Y., K. M. Zbuk, T. Sadler, A. Patocs, G. Lobo, E. Edelman, P. Platzer, M. S. Orloff, K. A. Waite and C. Eng (2008). "Germline mutations and variants in the succinate dehydrogenase genes in Cowden and Cowden-like syndromes." American journal of human genetics **83**(2): 261-268.
Nykamp, K., M. Anderson, M. Powers, J. Garcia, B. Herrera, Y. Y. Ho, Y. Kobayashi, N. Patil, J. Thusberg, M. Westbrook and S. Topper (2017). "Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria." Genet Med **19**(10): 1105-1117.
Okamoto, N., T. Naruto, T. Kohmoto, T. Komori and I. Imoto (2014). "A novel PTCH1 mutation in a patient with Gorlin syndrome." Human Genome Variation **1**(1): 14022.
Oldridge, D. A., A. C. Wood, N. Weichert-Leahey, I. Crimmins, R. Sussman, C. Winter, L. D. McDaniel, M. Diamond, L. S. Hart, S. Zhu, A. D. Durbin, B. J. Abraham, L. Anders, L. Tian, S. Zhang, J. S. Wei, J. Khan, K. Bramlett, N. Rahman, M. Capasso, A. Iolascon, D. S. Gerhard, J. M. Guidry Auvil, R. A. Young, H. Hakonarson, S. J. Diskin, A. T. Look and J. M. Maris (2015). "Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism." Nature **528**(7582): 418-421.
Pena, L. D. M., Y. H. Jiang, K. Schoch, R. C. Spillmann, N. Walley, N. Stong, S. Rapisardo Horn, J. A. Sullivan, A. McConkie-Rosell, S. Kansagra, E. C. Smith, M. El-Dairi, J. Bellet, M. A. Keels, J. Jasien, P. G. Kranz, R. Noel, S. K. Nagaraj, R. K. Lark, D. S. G. Wechsler, D. Del Gaudio, M. L. Leung, L. G. Hendon, C. C. Parker, K. L. Jones, D. B. Goldstein and V. Shashi (2018). "Looking beyond the exome: a phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases." Genet Med **20**(4): 464-469.
Rentzsch, P., D. Witten, G. M. Cooper, J. Shendure and M. Kircher (2019). "CADD: predicting the deleteriousness of variants throughout the human genome." Nucleic Acids Res **47**(D1): D886-d894.
Ricketts, C., E. R. Woodward, P. Killick, M. R. Morris, D. Astuti, F. Latif and E. R. Maher (2008). "Germline SDHB mutations and familial renal cell carcinoma." J Natl Cancer Inst **100**(17): 1260-1262.
Ritchie, G. R., I. Dunham, E. Zeggini and P. Flicek (2014). "Functional annotation of noncoding sequence variants." Nat Methods **11**(3): 294-296.
Roberts, M. E., S. A. Jackson, L. R. Susswein, N. Zeinomar, X. Ma, M. L. Marshall, A. R. Stettner, B. Milewski, Z. Xu, B. D. Solomon, M. B. Terry, K. S. Hruska, R. T. Klein and W. K. Chung (2018). "MSH6 and PMS2 germ-line pathogenic variants implicated in Lynch syndrome are associated with breast cancer." Genet Med **20**(10): 1167-1174.
Schmitt, A. M. and H. Y. Chang (2016). "Long Noncoding RNAs in Cancer Pathways." Cancer Cell **29**(4): 452-463.
Schuster, S. L. and A. C. Hsieh (2019). "The Untranslated Regions of mRNAs in Cancer." Trends Cancer **5**(4): 245-262.
Sepahi, I., U. Faust, M. Sturm, K. Bosse, M. Kehrer, T. Heinrich, K. Grundman-Hauser, P. Bauer, S. Ossowski, H. Susak, R. Varon, E. Schröck, D. Niederacher, B. Auber, C. Sutter, N. Arnold, E. Hahnen, B. Dworniczak, S. Wang-Gorke, A. Gehrig, B. H. F. Weber, C. Engel, J. R. Lemke, A. Hartkopf, H. P. Nguyen, O. Riess and C. Schroeder (2019). "Investigating the effects of additional truncating variants in DNA-repair genes on breast cancer risk in BRCA1-positive women." BMC Cancer **19**(1): 787.
Shihab, H. A., M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. Day, T. R. Gaunt and C. Campbell (2015). "An integrative approach to predicting the functional effects of non-coding and coding sequence variation." Bioinformatics **31**(10): 1536-1543.

Short, P. J., J. F. McRae, G. Gallone, A. Sifrim, H. Won, D. H. Geschwind, C. F. Wright, H. V. Firth, D. R. FitzPatrick, J. C. Barrett and M. E. Hurles (2018). "De novo mutations in regulatory elements in neurodevelopmental disorders." Nature **555**(7698): 611-616.
Shuai, S., H. Suzuki, A. Diaz-Navarro, F. Nadeu, S. A. Kumar, A. Gutierrez-Fernandez, J. Delgado, M. Pinyol, C. López-Otín, X. S. Puente, M. D. Taylor, E. Campo and L. D. Stein (2019). "The U1 spliceosomal RNA is recurrently mutated in multiple cancers." Nature **574**(7780): 712-716.
Slack, F. J. and A. M. Chinnaiyan (2019). "The Role of Non-coding RNAs in Oncology." Cell **179**(5): 1033-1055.
Spurdle, A. B., P. J. Whiley, B. Thompson, B. Feng, S. Healey, M. A. Brown, C. Pettigrew, C. J. Van Asperen, M. G. Ausems, A. A. Kattentidt-Mouravieva, A. M. van den Ouweland, A. Lindblom, M. H. Pigg, R. K. Schmutzler, C. Engel, A. Meindl, S. Caputo, O. M. Sinilnikova, R. Lidereau, F. J. Couch, L. Guidugli, T. Hansen, M. Thomassen, D. M. Eccles, K. Tucker, J. Benitez, S. M. Domchek, A. E. Toland, E. J. Van Rensburg, B. Wappenschmidt, Å. Borg, M. P. Vreeswijk and D. E. Goldgar (2012). "BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk." J Med Genet **49**(8): 525-532.
St Laurent, G., C. Wahlestedt and P. Kapranov (2015). "The Landscape of long noncoding RNA classification." Trends Genet **31**(5): 239-251.
Stenson, P. D., E. V. Ball, M. Mort, A. D. Phillips, J. A. Shiel, N. S. Thomas, S. Abeysinghe, M. Krawczak and D. N. Cooper (2003). "Human Gene Mutation Database (HGMD): 2003 update." Hum Mutat **21**(6): 577-581.
Sugasawa, K. (2008). "XPC: its product and biological roles." Adv Exp Med Biol **637**: 47-56.
Suriano, G., C. Oliveira, P. Ferreira, J. C. Machado, M. C. Bordin, O. De Wever, E. A. Bruyneel, N. Moguilevsky, N. Grehan, T. R. Porter, F. M. Richards, R. H. Hruban, F. Roviello, D. Huntsman, M. Mareel, F. Carneiro, C. Caldas and R. Seruca (2003). "Identification of CDH1 germline missense mutations associated with functional inactivation of the E-cadherin protein in young gastric cancer probands." Hum Mol Genet **12**(5): 575-582.
Suzuki, H., S. A. Kumar, S. Shuai, A. Diaz-Navarro, A. Gutierrez-Fernandez, P. De Antonellis, F. M. G. Cavalli, K. Juraschka, H. Farooq, I. Shibahara, M. C. Vladoiu, J. Zhang, N. Abeysundara, D. Przelicki, P. Skowron, N. Gauer, B. Luu, C. Daniels, X. Wu, A. Forget, A. Momin, J. Wang, W. Dong, S. K. Kim, W. A. Grajkowska, A. Jouvet, M. Fèvre-Montange, M. L. Garrè, A. A. Nageswara Rao, C. Giannini, J. M. Kros, P. J. French, N. Jabado, H. K. Ng, W. S. Poon, C. G. Eberhart, I. F. Pollack, J. M. Olson, W. A. Weiss, T. Kumabe, E. López-Aguilar, B. Lach, M. Massimino, E. G. Van Meir, J. B. Rubin, R. Vibhakar, L. B. Chambless, N. Kijima, A. Klekner, L. Bognár, J. A. Chan, C. C. Faria, J. Ragoussis, S. M. Pfister, A. Goldenberg, R. J. Wechsler-Reya, S. D. Bailey, L. Garzia, A. S. Morrissy, M. A. Marra, X. Huang, D. Malkin, O. Ayrault, V. Ramaswamy, X. S. Puente, J. A. Calarco, L. Stein and M. D. Taylor (2019). "Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma." Nature **574**(7780): 707-711.
Tan, J., A. Biasini, R. Young and A. Marques (2018). An unexpected contribution of lincRNA splicing to enhancer function, bioRxiv.
Vinagre, J., A. Almeida, H. Populo, R. Batista, J. Lyra, V. Pinto, R. Coelho, R. Celestino, H. Prazeres, L. Lima, M. Melo, A. G. da Rocha, A. Preto, P. Castro, L. Castro, F. Pardal, J. M. Lopes, L. L. Santos, R. M. Reis, J. Cameselle-Teijeiro, M. Sobrinho-Simoes, J. Lima, V. Maximo and P. Soares (2013). "Frequency of TERT promoter mutations in human cancers." Nat Commun **4**: 2185.
Vogelstein, B., N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz and K. W. Kinzler (2013). "Cancer genome landscapes." Science **339**.
Wainstein, T., R. Kerr, C. L. Mitchell, S. Madaree, F. B. Essop, E. Vorster, R. Wainwright, J. Poole and A. Krause (2013). "Fanconi anaemia in black South African patients heterozygous for the FANCG c.637-643delTACCGCC founder mutation." S Afr Med J **103**(12 Suppl 1): 970-973.
Walsh, T., S. Casadei, M. K. Lee, C. C. Pennil, A. S. Nord, A. M. Thornton, W. Roeb, K. J. Agnew, S. M. Stray, A. Wickramanayake, B. Norquist, K. P. Pennington, R. L. Garcia, M. C. King and E. M. Swisher

(2011). "Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing." Proc Natl Acad Sci U S A **108**(44): 18032-18037.
Wang, Z. and C. B. Burge (2008). "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." (1469-9001 (Electronic)).
Weber, W., D. J. Nash, A. G. Motulsky, M. Henneberg, M. H. Crawford, S. K. Martin, J. M. Goldsmid, G. Spedini, S. Glidewell and M. S. Schanfield (2000). "Phylogenetic relationships of human populations in sub-Saharan Africa." Hum Biol **72**(5): 753-772.
Weinhold, N., A. Jacobsen, N. Schultz, C. Sander and W. Lee (2014). "Genome-wide analysis of noncoding regulatory mutations in cancer." Nat Genet **46**(11): 1160-1165.
Wells, A., D. Heckerman, A. Torkamani, L. Yin, J. Sebat, B. Ren, A. Telenti and J. di Iulio (2019). "Ranking of non-coding pathogenic variants and putative essential regions of the human genome." Nat Commun **10**(1): 5241.
Xie, S., J. Duan, B. Li, P. Zhou and G. C. Hon (2017). "Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells." Mol Cell **66**(2): 285-299.e285.
Yao, L., F. Schiavi, A. Cascon, Y. Qin, L. Inglada-Pérez, E. E. King, R. A. Toledo, T. Ercolino, E. Rapizzi, C. J. Ricketts, L. Mori, M. Giacchè, A. Mendola, E. Taschin, F. Boaretto, P. Loli, M. Iacobone, G. P. Rossi, B. Biondi, J. V. Lima-Junior, C. E. Kater, M. Bex, M. Vikkula, A. B. Grossman, S. B. Gruber, M. Barontini, A. Persu, M. Castellano, S. P. Toledo, E. R. Maher, M. Mannelli, G. Opocher, M. Robledo and P. L. Dahia (2010). "Spectrum and prevalence of FP/TMEM127 gene mutations in pheochromocytomas and paragangliomas." Jama **304**(23): 2611-2619.
Zhang, F. and J. R. Lupski (2015). "Non-coding genetic variants in human disease." Hum Mol Genet **24**(R1): R102-110.
Zhang, W., A. Bojorquez-Gomez, D. O. Velez, G. Xu, K. S. Sanchez, J. P. Shen, K. Chen, K. Licon, C. Melton, K. M. Olson, M. K. Yu, J. K. Huang, H. Carter, E. K. Farley, M. Snyder, S. I. Fraley, J. F. Kreisberg and T. Ideker (2018). "A global transcriptional network connecting noncoding mutations to changes in tumor gene expression." Nature genetics **50**(4): 613-620.

# Chapter 6

# Concluding Discussion

The aim of this work was to create a better representation of the germline variants found in South African women with breast cancer. Most studies on breast cancer has been primarily focussed on Caucasians and Asians, and thus variants in these populations are better described and their effects better understood. A recent demographic study of the different ethnical groups in South Africa identified the African population of the country to make up 81 percent of the total population. To our knowledge, only two studies have used multigene panels to investigate germline variants in breast cancer in Africa (Cameroon / Uganda /Sudan, in collaboration with American teams).

First and foremost, we wanted to focus on identifying known pathogenic/likely pathogenic variants in important cancer susceptibility genes. These genes included: BRCA1, BRCA2, TP53 (high penetrance), ATM, CHEK2, BRIP1, PALB2, RAD50, NBN (medium penetrance) and PTEN, RAD51C, BARD1, STK11, CDH1 (low penetrance), MLH1, MSH2, MSH6 and PMS2 (mismatch repair pathway). Breast cancer studies in South Africa have been highly focused on BRCA1 and BRCA2 with little information on other cancer susceptibility genes. This leaves a fairly large gap in our understanding regarding many other cancer susceptibility genes in African populations.

In light of more recent studies (2021-2022) regarding the classification of important cancer genes we were able to identify some genes that we required to include in future studies and genes from this study that should be excluded. These papers were released to the public in the latter part of the thesis submission processed and unfortunately did not form part of this study. According to these articles the core clinically significant genes are ATM, BARD1, BRCA1, BRCA2, CHEK2, PALB2, RAD51C, RAD51D and TP53. Genes that statically were not significant and should be reconsider for inclusion are BRIP1, CDH1, MLH1, MSH2, MSH6, NBN, NF1, PMS2, PTEN, RAD50 and STK11 (Dorling, Carvalho et al. 2021, Foulkes 2021, Hu, Hart et al. 2021).

Our second objective was to investigate other hereditary cancer predisposition genes. Rather than presenting an unfocused study with all the different types of variants, we opted to only lift out truncating variants which would be the most pathogenically relevant to the research community.

Lastly, we analysed non-coding variants that may harbour a pathogenic effect. Recently, more research has shown that non-coding variants may play a significant role in cancer development.

The sample set of this study consisted out of 165 females of African ancestry (self-reported) all previously diagnosed with breast cancer. Peripheral blood was collected from these females between the years of 1993 and 2001, the ages of the patients ranged from 21 to 85 years. All collections were done at the Oncology Clinic at Steve Biko Hospital, Pretoria, South Africa. These samples were collected with the patients' consent and stored as frozen blood samples.

All patients were previously checked for BRCA1 and BRCA2 deleterious variants using SSCP/Heteroduplex analyses and multiplex ligation-dependent probe amplification (MLPA), except for four (BRB130, BRB290, BRC134 and BRC210). In the absence of the initial screening for BRCA1 and BRCA2 deleterious variants, variants from other genes would have been responsible for the development of breast cancer in these patients.

Our decision on the use of the Illumina TruSight Cancer sequencing panel for the analysis was based on the fact that it delivered a fairly comprehensive spread across 94 genes including our cancer susceptibility genes of interest, and also 284 SNPs related to cancer.

All processing and analysis was performed in-house. Processes were executed on a Linux cluster housed at the Centre for Bioinformatics and Computational Biology, University of Pretoria, South Africa. A pipeline setup was done using BCBIO to follow the GATK best practices with BWA-MEM alignment. Variant calling was carried out using the HaplotypeCaller, afterwards functional annotation was done using VEP using default parameters and dbNSFP.

Variants were filtered by removing any variants that had a minor allele frequency (MAF) of ≥ 1% in the 1 000 Genomes African database. To select the most significant variants from the remaining set, we retained in-frame insertions or deletions, truncating, nonsense, frameshift, or splice-site variants, and non-synonymous variants predicted to be deleterious by at least 3/5 in silico functional effect predictors: LRT_pred, MutationTaster, PROVEAN, CADD and FATHMM.

We were able to identify pathogenic/likely pathogenic variants in 7.9% of the cohort which included important breast cancer susceptibility genes: ALK, ATM, BUB1B, BRCA1, BRCA2, CHEK2, FANCG, PALB2, RB1 and XPC. Even though these samples were pre-screened as BRCA1 and BRCA2 negative, we were still able to identify a deleterious BRCA1 variant in patient BRB264. Two patients, BRB130 and BRB290, where the BRCA status was unknown beforehand were found to carry a deleterious BRCA1 and BRCA2 variant respectively.

Furthermore, 27 variants of unknown significance were identified, these variants were located in important cancer susceptibility genes and may play a role in promoting cancer. The significance may

be unknown but well worth documenting for future studies. Six of these variants have not even been described by previous studies.

In the last chapter we could identify a small number of non-coding variants that may have some role in cancer susceptibility. It should be kept in mind, this study used targeted panels centred around known genes, so a large proportion of the non-coding genome was not even included in this study. To improve the investigation of non-coding variants, whole genome sequencing would be required. Panel sequencing focusses on specific areas of the genome while excluding a large proportion of the non-coding genome.

Importantly, African data shows a serious lack of information regarding cancer susceptibility genes other than BRCA1 and BRCA2. Thus, we highlighted other genes with their linked variants that may have an impact on cancer development. It is generally unclear from our study if breast cancer was caused by a single variant or an accumulation of different cancer-related variants.

The significance of a select few VUS and possible deleterious non-coding variants from this study Is generally not clear at this stage. In future studies some of these variants may resurface and this study would then play a role in strengthening claims of future studies regarding those variants. It may be that some of these variants may turn up to be an undiscovered founder mutation or highly pathogenic.

Little is still known about non-coding variants and how much of an effect they may have on cancer development. This study would help support future studies in their findings and function as a roadmap on how variants were called initially and how it was improved in follow up studies.

This study serves as the first attempt at identifying germline cancer susceptibility variants using a multigene panel in a south African population where the individuals were of an African ancestry. This is a novel approach for this country and the population under scrutiny. It may also contribute to the identification of pathogenic breast cancer variants in African females of southern African descent. Identification of variants linked to a specific ethnic group may even help understand migration of these population into/out and across the African continent.

As mentioned previously, most breast cancer susceptibility testing in South Africa has focussed on BRCA1 and BRCA2. This study is an example where 98% of the patients would test negative for breast cancer susceptibility using classical methods. The work done here should alert the South

African research community that other tests than those investigating BRCA genes are required for the early detection of important breast cancer variants.

The sample size of this study was not as large as we would have liked, due to financial constraints. Possibly because of the limited samples used, no radical new conclusions were made. To complicate this even further, females from this study were all of African ancestry. However South Africa has several ethnolinguistic groups, and there is a very wide range of throughout the rest of Africa, and differences between these groups may be significant.

Breast cancer susceptibility testing in South Africa has very recently been expanding with multigene panels being implemented by commercial pathology and specialized genomics companies. However, these costs are still relatively expensive in terms of the average South African's financial situation. The implementation of these tests in the public health sector will be highly challenging, due to the weak economic status of public health in South Africa. Decreasing the cost of genomic testing in Africa for a wide range of conditions will be a crucial activity.

In conclusion this study pioneers multigene panel sequencing for breast cancer in South Africa and will hopefully play an important role in stimulating and leveraging more and larger as a gateway into better understanding breast cancer in southern Africa. Only a relatively small subset of samples were investigated here and in doing so we were able to identify a range of pathogenic and likely pathogenic variants, including deleterious variants that have not been described previously. Africa is a treasure trove for cancer research and improving our understanding of cancer from this genomically-neglected continent is paramount.

Dorling, L., S. Carvalho, J. Allen, A. González-Neira, C. Luccarini, C. Wahlström, K. A. Pooley, M. T. Parsons, C. Fortuno, Q. Wang, M. K. Bolla, J. Dennis, R. Keeman, M. R. Alonso, N. Álvarez, B. Herraez, V. Fernandez, R. Núñez-Torres, A. Osorio, J. Valcich, M. Li, T. Törngren, P. A. Harrington, C. Baynes, D. M. Conroy, B. Decker, L. Fachal, N. Mavaddat, T. Ahearn, K. Aittomäki, N. N. Antonenkova, N. Arnold, P. Arveux, M. Ausems, P. Auvinen, H. Becher, M. W. Beckmann, S. Behrens, M. Bermisheva, K. Białkowska, C. Blomqvist, N. V. Bogdanova, N. Bogdanova-Markov, S. E. Bojesen, B. Bonanni, A. L. Børresen-Dale, H. Brauch, M. Bremer, I. Briceno, T. Brüning, B. Burwinkel, D. A. Cameron, N. J. Camp, A. Campbell, A. Carracedo, J. E. Castelao, M. H. Cessna, S. J. Chanock, H. Christiansen, J. M. Collée, E. Cordina-Duverger, S. Cornelissen, K. Czene, T. Dörk, A. B. Ekici, C. Engel, M. Eriksson, P. A. Fasching, J. Figueroa, H. Flyger, A. Försti, M. Gabrielson, M. Gago-Dominguez, V. Georgoulias, F. Gil, G. G. Giles, G. Glendon, E. B. G. Garcia, G. I. G. Alnæs, P. Guénel, A. Hadjisavvas, L. Haeberle, E. Hahnen, P. Hall, U. Hamann, E. F. Harkness, J. M. Hartikainen, M. Hartman, W. He, B. A. M. Heemskerk-Gerritsen, P. Hillemanns, F. B. L. Hogervorst, A. Hollestelle, W. K. Ho, M. J. Hooning, A. Howell, K. Humphreys, F. Idris, A. Jakubowska, A. Jung, P. M. Kapoor, M. J. Kerin, E. Khusnutdinova, S. W. Kim, Y. D. Ko, V. M.

Kosma, V. N. Kristensen, K. Kyriacou, I. M. M. Lakeman, J. W. Lee, M. H. Lee, J. Li, A. Lindblom, W. Y. Lo, M. A. Loizidou, A. Lophatananon, J. Lubiński, R. J. MacInnis, M. J. Madsen, A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M. E. Martinez, T. Maurer, D. Mavroudis, C. McLean, A. Meindl, A. R. Mensenkamp, K. Michailidou, N. Miller, N. A. Mohd Taib, K. Muir, A. M. Mulligan, H. Nevanlinna, W. G. Newman, B. G. Nordestgaard, P. S. Ng, J. C. Oosterwijk, S. K. Park, T. W. Park-Simon, J. I. A. Perez, P. Peterlongo, D. J. Porteous, K. Prajzendanc, D. Prokofyeva, P. Radice, M. U. Rashid, V. Rhenius, M. A. Rookus, T. Rüdiger, E. Saloustros, E. J. Sawyer, R. K. Schmutzler, A. Schneeweiss, P. Schürmann, M. Shah, C. Sohn, M. C. Southey, H. Surowy, M. Suvanto, S. Thanasitthichai, I. Tomlinson, D. Torres, T. Truong, M. Tzardi, Y. Valova, C. J. van Asperen, R. M. Van Dam, A. M. W. van den Ouweland, L. E. van der Kolk, E. M. van Veen, C. Wendt, J. A. Williams, X. R. Yang, S. Y. Yoon, M. P. Zamora, D. G. Evans, M. de la Hoya, J. Simard, A. C. Antoniou, Å. Borg, I. L. Andrulis, J. Chang-Claude, M. García-Closas, G. Chenevix-Trench, R. L. Milne, P. D. P. Pharoah, M. K. Schmidt, A. B. Spurdle, M. P. G. Vreeswijk, J. Benitez, A. M. Dunning, A. Kvist, S. H. Teo, P. Devilee and D. F. Easton (2021). "Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women." N Engl J Med **384**(5): 428-439.

Foulkes, W. D. (2021). "The ten genes for breast (and ovarian) cancer susceptibility." (1759-4782 (Electronic)).

Hu, C., S. N. Hart, R. Gnanaolivu, H. Huang, K. Y. Lee, J. Na, C. Gao, J. Lilyquist, S. Yadav, N. J. Boddicker, R. Samara, J. Klebba, C. B. Ambrosone, H. Anton-Culver, P. Auer, E. V. Bandera, L. Bernstein, K. A. Bertrand, E. S. Burnside, B. D. Carter, H. Eliassen, S. M. Gapstur, M. Gaudet, C. Haiman, J. M. Hodge, D. J. Hunter, E. J. Jacobs, E. M. John, C. Kooperberg, A. W. Kurian, L. Le Marchand, S. Lindstroem, T. Lindstrom, H. Ma, S. Neuhausen, P. A. Newcomb, K. M. O'Brien, J. E. Olson, I. M. Ong, T. Pal, J. R. Palmer, A. V. Patel, S. Reid, L. Rosenberg, D. P. Sandler, C. Scott, R. Tamimi, J. A. Taylor, A. Trentham-Dietz, C. M. Vachon, C. Weinberg, S. Yao, A. Ziogas, J. N. Weitzel, D. E. Goldgar, S. M. Domchek, K. L. Nathanson, P. Kraft, E. C. Polley and F. J. Couch (2021). "A Population-Based Study of Genes Previously Implicated in Breast Cancer." (1533-4406 (Electronic)).