

https://dspace.org/

The contribution of common genetic variants to breast cancer risk in South African black populations

Hayat, Mahtaab

2023-08
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/43876
Hayat, Mahtaab. (2023). The contribution of common genetic variants to breast cancer risk in South African black populations. [Masters dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace.

Downloaded from a DSpace repository. https://dspace.org/

The contribution of common genetic variants to breast cancer risk in South African black populations

Mahtaab Hayat



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Supervisor:

Professor Christopher Mathew

Co-supervisors:

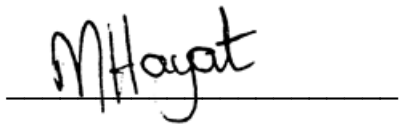
Dr Jean-Tristan Brandenburg and Professor Michèle Ramsay

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa, in fulfilment of the requirements of the degree of Doctor of Philosophy in Human Genetics

Johannesburg, 2023

Declaration

I, Mahtaab Hayat, declare that this thesis is my own work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University. Co-author contributions to each publication or manuscript are listed under Author contributions.



Mahtaab Hayat, signed on the 31st day of August 2023

Dedication

بسم الله

هَٰذَا مِن فَضْلِ رَبِّي

Presentations arising from this thesis

Conference proceedings:

Hayat M, Chen WC, Brandenburg JT, Ramsay M, Mathew C on behalf of the H3A Consortium and the ERICA-SA study. The contribution of common genetic variants to breast cancer risk in South African Black populations, Poster presentation at the SASHG Biennial Congress, 3-6 August 2019, Cape Town, South Africa.

Hayat M, Chen WC, Brandenburg JT, Lewis CM, Ramsay M, AWI-Gen, ERICA-SA, Mathew C. The contribution of common genetic variants to breast cancer risk in South African Black populations. Poster presentation at the Wits Faculty of Health Sciences Biennial Research Day 2020, 15 October 2020, Virtual.

Hayat M, Chen WC, Brandenburg JT, Lewis CM, Ramsay M, AWI-Gen, ERICA-SA, Mathew CG. The Contribution of Common Genetic Variants to Breast Cancer Risk in South African Black Populations. Oral presentation at the AORTIC Cancer Genomics Conference, 18-19 February 2021, Virtual.

Hayat M, Chen WC, Brandenburg JT, Lewis CM, Ramsay M, AWI-Gen, ERICA-SA, Mathew CG. The Contribution of Common Genetic Variants to Breast Cancer Risk in South African Black Populations. Oral presentation at the AORTIC International Conference on Cancer in Africa, 5-10 November 2021, Virtual.

Hayat M, Chen WC, Brandenburg JT, Lewis CM, Ramsay M, Mathew CG for ERICA-SA. The Contribution of Common Genetic Variants to Breast Cancer Risk in South African Black Populations. Oral presentation at MBRT Postgraduate Research Day, 9 December 2021, Virtual.

Hayat M, Chen WC, Ramsay M, Brandenburg JT, Mathew CG for ERICA-SA. The contribution of common genetic variants to breast cancer risk in South African black populations. Poster presented at the Molecular Bioscience Research Thrust Research Day, 1 December 2022. Wits University, Johannesburg, South Africa.

Hayat M, Chen WC, Brandenburg JT, Ramsay M, Mathew CG. The contribution of common genetic variants to breast cancer risk in South African black populations. Oral presentation at the International Congress of Human Genetics, 22-26 February 2023. Cape Town, South Africa.

Publications arising from this thesis

Hayat M, Chen WC, Brandenburg JT, Babb de Villiers C, Ramsay M, Mathew CG (2021) Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations. *JCO Global Oncology*, **7(1462-1471)**, https://doi.org/10.1200/go.21.00089.

Related publications

Swiel Y, Brandenburg JT, Hayat M, Chen WC, Cox M & Hazelhurst S (2022) FPGA Acceleration of GWAS Permutation Testing. BioRxiv, preprint. https://doi.org/10.1101/2022.03.11.483235

Grants and Awards

2018: Recipient of Boehringer Ingelheim Fonds travel grant (€1,860)

2019: Recipient of Wits Health Science Faculty MFREF (R8000)

2023: Recipient of SASHG travel grant for ICHG 2023 (registration and accommodation covered and R4000 for flight ticket)

Abstract

Breast cancer is the second most common cancer in South African black women. The contribution of common genetic variants to breast cancer risk is well studied in non-African populations, but little is known about their role in resident African populations, and there are no published genome-wide association studies (GWAS) on breast cancer in Africa. This PhD thesis aimed to determine the contribution of common genetic variants to breast cancer in a South African black population.

A GWAS was carried out in 2,573 black female breast cancer patients from the Johannesburg Cancer Study and 744 population-matched, female controls from the AWI-Gen study. All participants were from Soweto, Johannesburg, South Africa. Samples were genotyped on the H3Africa SNP array. Replication testing was done of existing loci from European and African American (AA) populations in the resident African data, and loci from the resident African data in European and AA populations. A meta-analysis was carried out with an AA population. Finally, existing polygenic risk scores (PRSs) were tested in the resident African dataset.

Three variants at two loci were strongly associated with breast cancer in this study. Two variants (rs77422433, p-value=2.89×10^{-08}, odds ratio (OR):0.46, 95% confidence interval (95%CI): 0.40-0.52 and rs112410019, p-value=3.01×10^{-08}, OR: 0.47, 95%CI: 0.41-0.53) were located within the DNA repair gene *XRCC5*. These variants were not previously associated with breast cancer, suggesting that it may be an African specific risk locus. The second locus is on chromosome 16 in *CES5A* (rs3859109, p-value=4.54×10^{-08}, OR=0.70, 95%CI: 0.68-0.73), and had not previously been associated with breast cancer. None of these SNPs were replicated in European and AA populations. The meta-analysis with AA data revealed strong association of an intergenic SNP with breast cancer (rs139299680, $p_{meta}=7.25 \times 10^{-08}$) on chromosome 3. A polygenic risk score (PRS) developed in European populations demonstrated poor transferability to this African dataset.

This GWAS is the first to be conducted in a resident black African population. This study suggests that there may be African-specific genetic risk factors for African breast cancer, and that large genome-wide studies in African populations are essential to develop a comprehensive understanding of the genetics of breast cancer in Africa.

Acknowledgements

First and foremost, thanks to the Almighty for granting me this opportunity.

To my Ammi and my family, for standing by me as I completed this degree. For listening to my presentations and reading my work even when nothing made sense to you.

To my supervisor, Professor Chris Mathew: thank you for your guidance, patience and support during these years. This study would not be what it is without you.

To Dr JT Brandenburg, one of my co-supervisors: thank you for your constant check-ins, for your support and for motivating me to complete my analyses and write-up! Your support was much appreciated.

To Prof Michèle Ramsay, my other co-supervisor: thank you for your wise input, your words of encouragement and your support throughout the PhD journey.

Carl: although you were not a formal supervisor, your support, guidance and friendship have been invaluable since the beginning of my postgraduate journey. I do not have the words to express my gratitude.

My friends and colleagues at the SBIMB – you are all amazing. Thank you for always being around for a chat, vent or to bounce ideas with.

My friends and family, who asked, when are you submitting? Here it is. Finally. I'm done being a student now.

MZ, this will be the only piece of my academic work you won't get to read. Be with peace.

Thank you to all three of my examiners for taking the time to review my work. Your time and feedback is much appreciated.

Contents

Declaration i

Dedication ii

Presentations arising from this thesis iii

Publications arising from this thesis v

Related publications vi

Grants and Awards vii

Abstract viii

Acknowledgements x

List of tables xvii

List of abbreviations xviii

Chapter 1: Introduction 1

1.1. Breast Cancer Overview 1

1.2. Types of breast cancer 3

1.3. Treatment of breast cancer 3

1.4. Screening for breast cancer 5

1.5. Epidemiology of Breast Cancer 5

1.6. Genetic epidemiology of breast cancer 6

1.7. Genome-Wide Association Studies 9

1.8. Breast Cancer GWAS in non-African populations 12

1.9. GWAS in African American populations 14

1.10. Breast Cancer GWAS in sub-Saharan Africa 15

1.11. Polygenic Risk Scores 16

1.12. Study Aim and Objectives 18

Chapter 2: Breast Cancer Genetics in sub-Saharan African Populations 19
2.1 Introduction 19
2.2 Methods 21
2.2.1 Article selection and data extraction 23
2.3 Results 23
2.3.1 The contribution of rare variants to breast cancer risk in Sub-Saharan Africa 25
2.3.2 The contribution of common variants to breast cancer risk in Sub-Saharan Africa30
2.4 Discussion 32
2.4.1 Contribution of rare variants to breast cancer risks 32
2.4.2 Genetic association studies of common variants.....33
2.4.3 Current limitations and future prospects for breast cancer genetic research in Sub-Saharan Africa.....34
2.5 Recent genetic studies 37
2.5.1 Rare variant studies 37
2.5.2 Common variant studies.....38
2.5.3 Summary of the new studies 39
Chapter 3: Methods and Materials 40
3.1. Overview 40
3.2. Study design 40
3.3. Study sample 41
3.4. Sampling and DNA isolation..... 42
3.5. Genotyping and data quality control 43
3.5.1. Data analysis workflow.....44
3.5.2. Genotype data and quality control 45
3.5.3. Principal component analysis and admixture analysis.....48

3.5.4. Estimation of Heritability....49
3.6. Genetic imputation 49
3.7. GWAS LMM modelling 51
3.8. Permutation testing for top hits 52
3.9. Fine mapping and functional assessment of associated variants 52
3.10. Replication 53
3.11. Meta-analysis 53
3.12. Polygenic risk scores 53
Chapter 4: Results 55
4.1. Study Samples and Genotyping 55
4.2. Adjustment for population structure 56
4.2.1. Initial GWAS analysis 56
4.2.2. Analysis of genetic differences between male and female controls 57
4.2.3. Analysis of batch effects 59
4.2.4. Admixture analysis 61
4.2.5. Principal Component Analysis 62
4.3. Estimation of heritability 65
4.4. GWAS results 65
4.5. Fine mapping and functional analysis 76
4.5.1. FINEMAP: 95% credible set 76
4.5.2. eQTL results from FUMA 77
4.5.3. Functional annotation 78
4.6. Replication 79
4.6.1. Replication of known breast cancer risk SNPs in the African breast cancer GWAS 79
4.6.2. Replication of South African breast cancer SNPs in other populations 82

4.7. Meta-analysis 84
4.8. Polygenic Risk Score 87
Chapter 5: Discussion 91
5.1. Study design 91
5.2. Admixture analysis, population substructure and batch effects 93
5.2.1. Removal of male controls from the analysis 93
5.2.2. Removal of the JCS female controls from the analysis 94
5.2.3. Removing individuals with East Asian and European ancestry 95
5.3. Heritability of African breast cancer 95
5.4. Two loci associated with African breast cancer 97
5.5. Suggestive significant hits 99
5.6. Functional analysis 100
5.7. Replication in the African breast cancer GWAS 101
5.8. Meta-analysis 103
5.9. Polygenic Risk Scores 104
5.10. Conclusion 105
References 108
Appendices 136
Appendix A: JCS questionnaire 136
Appendix B: Ethics certificates 140
Appendix C: Plagiarism declaration 146
Appendix D: Turn-It-In support letter and report 147
Appendix E: Declaration: Student's contribution to article and agreement of co-authors 187
Appendix F: Published article (included in chapter 2) 189
Appendix G: Letter of permission to include adaptation of GWAS workflow 199

List of Figures

Figure 1.1: Estimated age-standardised incidence rates (ASR) (World) in 2020 for breast cancer, all ages, females only. 1
Figure 1.2: Ancestry of GWAS participants over time, as compared with the global population 10
Figure 2.1: Process of selection of papers 24
Figure 2.2: Variant distribution in genes and countries. 30
Figure 3.1: Map of Gauteng province showing case and control recruitment sites in Johannesburg, Gauteng. 41
Figure 3.2: The GWAS workflow summary 44
Figure 3.3: Schema of the H3Agwas pipeline 46
Figure 3.4: Overview of the H3Agwas QC workflow 47
Figure 4.1: Admixture plot for cases and all control sites 55
Figure 4.2 Manhattan plot including male and female control samples 57
Figure 4.3: Manhattan plots showing differences between male and female controls. 58
Figure 4.4: Manhattan plots showing inflation of p-values in the dataset after removing male controls 60
Figure 4.5: Admixture plot 61
Figure 4.6 Eigenvalue curve of breast cancer cases and controls 62
Figure 4.7: Principal Component Analysis plots of cases and controls 63
Figure 4.8: Quantile-quantile plots 64
Figure 4.9: Manhattan plot of 2,573 breast cancer cases against 744 AWI-Gen controls. 67
Figure 4.10: Breast cancer GWAS regional plots using LocusZoom 73
Figure 4.11: Genotype cluster plot of rs3859109 74
Figure 4.12: Genotype plot for rs2143559. 75
Figure 4.13: Manhattan plot of meta-analysis 86
Figure 4.14: Regional plot using LocusZoom of the top marker in the meta-analysis of the resident African data and BCAC African American data. 87
Figure 4.15: Polygenic Risk Score of 313 SNP model applied to African breast cancer dataset. 88

Figure 4.16: Polygenic Risk Score of 3,820 SNP model applied to African breast cancer dataset. 89
Figure 4.17: Polygenic risk score from the African American BCAC summary statistics applied to the African breast cancer dataset..... 90
Figure 4.18: Polygenic risk score generated from the European BCAC summary statistics applied to the African breast cancer dataset 90

List of tables

Table 2.1. Keywords and combinations of keywords used to search the literature using PubMed 21
Table 2.2. Studies of rare variants in breast cancer genes in Sub-Saharan Africa 26
Table 2.3. Predicted Pathogenic Variants Detected in Two or More African Patients 28
Table 2.4. Association Studies of Common Variants in SSA breast cancer 31
Table 2.5. Predicted Pathogenic Rare Variants Detected in two or more African breast cancer patients in recent studies 38
Table 2.6. New Association Studies of Common Variants in SSA breast cancer 39
Table 4.1. Number of samples with 5% European and East Asian ancestry 61
Table 4.2. Top SNPs associated with African breast cancer (p-value<$5x10^{-06}$). 68
Table 4.3. Credible set of causal variants in chromosomes 2 and 16 76
Table 4.4. Results of eQTL analysis in breast tissue mapped to breast cancer from FUMA 78
Table 4.5. SNPs predicted to be deleterious by PolyPhen and SIFT in SNP-nexus 79
Table 4.6. Replicating top hits (p-value<$5x10^{-06}$) from BCAC African American hits in African breast cancer GWAS 80
Table 4.7. Test for replicating top hits (p-value<$5x10^{-08}$) from BCAC European hits in the African breast cancer GWAS. 81
Table 4.8. Replicating top hits from the African breast cancer GWAS in the BCAC African American GWAS (p-value<0.05) 82
Table 4.9. Replicating top hits from the African breast cancer GWAS in the BCAC European GWAS 83
Table 4.10. Meta-analysis with African breast cancer data and BCAC African American data 85

List of abbreviations

1KGP	1000 Genomes project
AA	African American
AGIN	Agincourt (population)
AMBER	African American Breast Cancer Epidemiology and Risk
AUC	Area under the curve
AWI-Gen	Africa Wits-INDEPTH partnership for Genomic study
AWS	Amazon web service
BC	Breast cancer
BCAC	Breast Cancer Association Consortium
CEU	Northern European (population)
CHB	Han Chinese (population)
CI	Confidence interval
COGS	Collaborative Oncologic Gene-environment Study
COVID-19	Corona virus disease
DIK	Dikgale (population)
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
eQTL	Expression quantitative trait locus
ER	Oestrogen receptor
ERICA-SA	Evolving Risk Factors for Cancer in African populations
EUR	European
FDR	False discovery rate
FPGA	Field-Programmable Gate Arrays
FUMA	Functional mapping and annotation
gDNA	Genomic DNA
GLM	Generalized linear models

GREML	GRM restricted maximum likelihood
GRM	Genomic relatedness matrix
GRS	Genetic risk score
GTC	Genotype Call
GTex	Genotype-Tissue Expression (project)
GWAS	Genome-wide association studies
H3Africa	Human, Heredity and Health in Africa
HER2	Human epidermal growth factor receptor 2
HICs	High-income countries
HIV/AIDS	Human Immunodeficiency Virus/Acquired Immunodeficiency syndrome
HREC	Human Research Ethics Committee
HWE	Hardy-Weinberg equilibrium
IAAP	Illumina Array Analysis Platform
IARC	International Agency for Research on Cancer
JCS	Johannesburg Cancer Study
LD	Linkage disequilibrium
LMICs	Low-middle income countries
LMM	Linear mixed model
LOF	Loss-of-function
MAF	Minor allele frequency
MLPA	Multiplex ligation-dependent probe amplification
MRI	Magnetic resonance imaging
NCI	National Cancer Institute
NGS	Next-generation sequencing
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis

PCR	Polymerase chain reaction
PhD	Doctor of Philosophy
PMID	PubMed ID
PR	Progesterone receptor
PRS	Polygenic risk score
PRSs	Polygenic risk scores
QC	Quality control
QRT-PCR	Quantitative real-time PCR
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
ROOT	Breast Cancer in the African Diaspora
SAMRC	South African Medical Research Council
SBIMB	Sydney Brenner Institute for Molecular Bioscience
SNP	Single-nucleotide polymorphism
SNPs	Single-nucleotide polymorphisms
SSA	Sub-Saharan Africa
SSCP	Single strand conformation polymorphism
SWT	Soweto (population)
TE	Tris EDTA
TNBC	Triple-negative breast cancer
UKMRC	United Kingdom Medical Research Council
US	United States
WGS	Whole genome sequencing
YRI	Yoruban (population)

Chapter 1: Introduction

1.1. Breast Cancer Overview

Breast cancer was the most common cancer to be diagnosed in the world in 2020 and is the most common cause of death in women worldwide (Sung *et al.*, 2021). The International Agency for Research on Cancer (IARC) estimated that in 2020 there were more than 2.26 million new breast cancer cases, and more than half a million deaths due to breast cancer worldwide (Figure 1.1) (Sung *et al.*, 2021). The IARC reports that in 2020 there were 186,598 breast cancer cases recorded in Africa which accounted for approximately 8% of the cases worldwide, and this is predicted to increase due to urbanisation affecting lifestyle changes (Lukong *et al.*, 2017). However, the IARC cancer incidence report includes data for less than 50% of sub-Saharan African countries as not all of them have high-quality cancer registries (Wild *et al.*, 2020). Thus it is likely that the incidence of breast cancer in Africa is under-reported owing to the lack of infrastructure for cancer diagnosis in resource-constrained countries (Agyepong *et al.*, 2017; Cumber *et al.*, 2017; Wild *et al.*, 2020; Anyigba *et al.*, 2021).

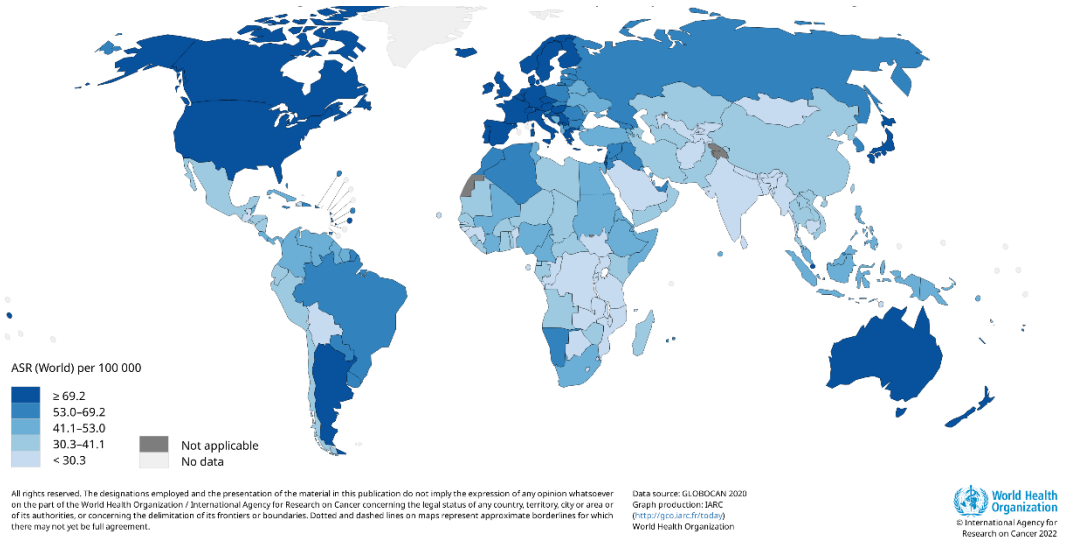


Figure 0.1: Estimated age-standardised incidence rates (ASR) (World) in 2020 for breast cancer, all ages, females only. Image shows high rates of BC in first world countries. Rates in South Africa are increasing, compared to many other SSA countries. GLOBOCAN 2021. (Sung et al., 2021).

While countries in sub-Saharan Africa are heavily burdened with communicable diseases such as HIV/AIDS, malaria and tuberculosis, there is an added strain on the healthcare system with the increase of non-communicable diseases such as respiratory diseases, diabetes, cardiovascular diseases and cancer (Roth *et al.*, 2018; Gouda *et al.*, 2019). This results in women having limited access to medical facilities, with restricted and delayed options for treatment and therefore, high mortality rates (Barr *et al.*, 2016; Agyepong *et al.*, 2017). Genetic research on cancer is given less priority given pressing concerns regarding HIV/AIDS, malaria and tuberculosis infections in sub-Saharan Africa (Ngoma and Ngoma, 2019).

The mortality from breast cancer remains one of the highest in the world. The age-standardised breast cancer mortality rate in sub-Saharan Africa is 19.3 per 100 000 people which is higher compared to the European rate of 14.8 per 100,000 people and even higher when compared to North America's mortality rate of 12.5 per 100,000 people (Sung *et al.*, 2021). High-income countries (HICs), before the COVID-19 pandemic, had substantially reduced their burden of communicable diseases and focused most of their healthcare staff and funding on non-communicable diseases such as cancer. However, low-and-middle-income countries (LMICs) that include most of the countries in Africa are having to cope with high rates of infectious diseases such as tuberculosis, malaria, HIV/AIDS and Ebola (Barr *et al.*, 2016; Agyepong *et al.*, 2017). These diseases consume most healthcare resources, leaving little funding, time and staff dedicated to cancer care, leading to higher early mortality. For example, women in public hospitals in South Africa face extensive delays in receiving breast cancer treatment, which is associated with worse survival (Pumpalova *et al.*, 2022). These delays are attributed to various factors, including ethnicity, level of education and living distance from a hospital.

This introduction will discuss the epidemiology of breast cancer, the types of breast cancer and the genetic risk factors associated with breast cancer. It also reviews the approaches used to identify genetic variants that are associated with breast cancer, and the progress that has been made in European versus African populations. Finally, the potential clinical uses of the data generated from genetic studies such as polygenic risk scores are discussed.

1.2. Types of breast cancer

Breast cancer is a very heterogeneous disease, and can be classified histologically and using molecular markers. Histologically, there are two main types of cancers: lobular and ductal breast cancer, with ductal carcinoma being more common (Tavassoli and Devilee, 2003; Loibl *et al.*, 2021). Classifying breast cancer further according to a molecular subtype is based on the presence or absence of three main receptors that are expressed by breast cancer cells: the hormone receptors oestrogen receptor (ER) and progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) (Perou *et al.*, 2000; Sørlie *et al.*, 2001; Loibl *et al.*, 2021). Luminal A breast cancers are ER+/PR+/HER2-, luminal B are ER+/PR-/HER2+, HER2 positive are ER-/PR-, and basal-like are ER-/PR-/HER2-.

Triple-negative breast cancers (TNBC) are ER-/PR-/HER2-. A proliferation marker protein Ki-67 helps to differentiate between luminal A breast cancers and luminal B, HER2+ and TNBC breast cancers. The advent of high throughput DNA sequencing technology has enabled a more detailed classification of breast cancers based on somatic mutations in tumour tissue that define the molecular pathways involved. The classification of tumours can inform treatment strategies, as discussed in section 1.3.

1.3. Treatment of breast cancer

The main forms of therapy for breast cancer are either surgery, adjuvant therapy or targeted therapy. The type of treatment largely depends on the type of breast cancer the patient presents with and how advanced the disease is. A bilateral mastectomy at the time of diagnosis as a form of treatment is an option which increases survival rates (Bedrosian *et al.*, 2010). If a mastectomy is not opted for, breast-conserving surgery is carried out where the tumour, if localised, will be removed (Loibl *et al.*, 2021; Trayes and Cokenakes, 2021). Adjuvant therapy can be given before or after surgery and includes radiotherapy, chemotherapy, hormone therapy or immunotherapy. Targeted therapies based on the genomic changes in the tumour can also be carried out (Trayes and Cokenakes, 2021). All therapies aim to reduce the size of the

tumour and the number of cancerous cells in a patient so that the tumour can be removed surgically (Loibl *et al.*, 2021; Trayes and Cokenakes, 2021).

Radiation can be administered pre- or post-surgery. It is generally used for localised cancers and has been found to reduce the length of treatment. Radiation is reported to induce fewer side effects when compared to other therapies, such as chemotherapy (Shah *et al.*, 2020). Chemotherapy is used after surgery but it can also be administered before surgery to shrink the size of a tumour, or to target metastatic cells (Loibl *et al.*, 2021). It is usually involved in interfering with DNA replication by preventing mitosis from occurring and therefore preventing cancer cells from dividing. Both radiation and chemotherapy can be combined with another type of therapy to increase the effectiveness of the treatment.

Hormone treatment, as the name suggests, targets the hormone receptors of cancer with the exception of triple-negative breast cancer, or suppresses the production of oestrogen (Abe *et al.*, 2011; Trayes and Cokenakes, 2021). Cancer cells evade the immune system which enables them to proliferate in an uncontrolled manner. Immunotherapy can allow the body to identify and target cancer cells through various approaches including immune checkpoint blockades and adoptive T-cell therapy. This allows the immune system to kill cancer cells on its own (Barzaman *et al.*, 2021).

Another form of targeted therapy involves the inhibition of poly (ADP-ribose) polymerase (PARP). Poly (ADP-ribose) polymerase 1 (PARP1) is important in the maintenance of genome integrity and DNA repair. Briefly, inhibition of PARP prevents the enzyme from repairing DNA in cancer cells, prompting cancer cell death (Curtin and Szabo, 2020). This inhibition works particularly well in cancers with mutations in *BRCA1* and *BRCA2* which are already partially deficient in DNA repair, making it an effective form of therapy for *BRCA* mutation carriers. In addition, breast cancers positive for HER2 are targeted with Herceptin (trastuzumab), an anti-

HER2 monoclonal antibody (Slamon *et al.*, 2011). This targeted form of therapy has been successful and improves the over-all survival of patients treated with it (Cameron *et al.*, 2017).

1.4. Screening for breast cancer

The most effective way to improve outcomes for breast cancer patients is early detection. Screening is key to early detection and can significantly reduce mortality in patients. Types of screening tools include mammography, ultrasonography and molecular resonance imaging (MRI) screening (Loibl *et al.*, 2021). Mammography can reduce mortality by 20% (Oeffinger *et al.*, 2015). MRI screening has greatly improved the detection of breast cancer, particularly in young women and women with denser breast tissue (Sardanelli *et al.*, 2011). In addition to these screening methods, genetic screening of high-risk genes can also be carried out, which will be discussed in section 1.6.

1.5. Epidemiology of Breast Cancer

The epidemiology of breast cancer is complex and the causes are multifactorial, with genetic risk factors and the environment playing important roles in cancer development (Momenimovahed and Salehiniya, 2019; Loibl *et al.*, 2021). Many risk factors, such as older age, lifestyle, environmental factors genetic risk factors and family history (which includes shared environmental and genetic factors) contribute to the development of breast cancer. Environmental risk factors include consumption of alcohol, smoking, obesity, low physical activity, hormonal replacement therapies and radiation exposure (Wild *et al.*, 2020; Loibl *et al.*, 2021). Genetic risk factors contributing to the development of breast cancer include rare pathogenic variants (Dorling *et al.*, 2021) or many common variants that, in combination, increase the risk of developing breast cancer (Maxwell and Nathanson, 2013).

Increased age is an important risk factor for the development of breast cancer, with the risk doubling with each decade of life until menopause when the risk increases more slowly

(McPherson *et al.*, 2000; Veronesi *et al.*, 2005; Rojas and Stuckey, 2016). Alcohol consumption is considered a carcinogenic agent, even if consumed at low levels (Shield *et al.*, 2016). Evidence suggests that consuming 10-15 grams of alcohol per day can increase the risk for the disease (Freudenheim, 2020). Obesity is another risk factor for breast cancer. A high BMI, particularly in postmenopausal women, increases the risk of developing breast cancer (Veronesi *et al.*, 2005; Engin, 2017; Picon-Ruiz *et al.*, 2017; Gravena *et al.*, 2018). Obesity also negatively affects the response to treatment and survival of patients (Engin, 2017; Lee *et al.*, 2019).

Reproductive and hormonal risk factors include early age of menarche, later age to menopause, advanced age at first birth, fewer children, less breastfeeding, menopausal hormone therapy, and oral contraceptives (Hong *et al.*, 2022). The use of hormone replacement therapies has been shown to increase the risk of developing breast cancer in menopausal and non-menopausal women. A substantial study, including over 100,000 menopausal women with cancer and 400,000 controls, supports this risk for menopausal women, specifically, in developed countries (Group on Hormonal Factors in Breast Cancer, 2021). This study found that the use of both oestrogen and progesterone for more than five years greatly increases the risk of developing breast cancer (odds ratio: 2.08, 95% confidence interval (CI): 2.02-2.15). It should be noted that this study included women only from developed countries. Similarly, two studies found that the long-term use (more than five years) of oestrogen and progesterone significantly increases the risk of breast cancer development in non-menopausal women with odds ratios of 1.28 (95% CI: 1.13-1.45) (Chlebowski *et al.*, 2020) and 1.79 (95% CI: 1.73-1.85) (Vinogradova *et al.*, 2020). Genetic risk factors are also important in breast cancer and will be discussed further in section 1.6.

1.6. Genetic epidemiology of breast cancer

In the diagnosis of breast cancer, hereditary or familial links were identified as risk factors and when further investigated, provided strong evidence for a genetic contribution to breast cancer (Claus *et al.*, 1990; Easton *et al.*, 1993). This led to genetic linkage studies and positional cloning

being carried out which identified the first genes associated with breast cancer: *BRCA1* (Miki *et al.*, 1994) and *BRCA2* (Wooster *et al.*, 1995). Patients with inherited BC, particularly those with mutations in *BRCA1*, often present with TNBC (Foulkes *et al.*, 2010). Mutations in these genes have a dominant effect and confer a high risk of breast cancer development, but appeared to account for a small number of all breast cancer cases (Claus *et al.*, 1990; Easton *et al.*, 1993; Melchor and Benítez, 2013). The remaining cases were thought to occur sporadically (Claus *et al.*, 1990; Easton *et al.*, 1993).

Most of the genes associated with inherited breast cancers (high or moderately high risk of breast cancer) encode proteins involved in the detection or repair of DNA damage. The BRCA1 and BRCA2 proteins, for example, play a role in repairing double-stranded breaks in DNA, preventing an uncontrolled overgrowth of cells and therefore act as tumour suppressors. Loss-of-function (LOF) mutations in these genes increase genomic instability, leading to breast cancer development (Roy *et al.*, 2011; Savage and Harkin, 2015; Sadeghi *et al.*, 2020). The PALB2 protein, encoded by the *PALB2* gene, is recruited by BRCA1 when double-stranded breaks in DNA are being repaired and the BRCA2 protein is subsequently recruited (Toh *et al.*, 2020). As technology improved, other high-risk genes with loss of function mutations associated with breast cancer were discovered including *CHEK2* (Meijers-Heijboer *et al.*, 2002), *PALB2* (Rahman *et al.*, 2007) and *ATM* (Renwick *et al.*, 2006). *ATM* encodes a protein which plays a key role in cell cycle checkpoint signalling pathways associated with the cellular response to DNA damage, and *CHEK2* plays a critical role in halting cell cycle progression in response to DNA damage (Audrey *et al.*, 2022).

The discovery of these genes has had very important consequences for genetic testing for breast cancer risk. Soon after the US Supreme Court ruled that genes cannot be patented (Justia US Supreme Court, 2013), gene panels were created to test for the risk of breast cancer that included the *BRCA1* and *BRCA2* genes and other genes that were strongly associated with breast cancer. However, key questions and considerations had to be addressed when designing and administering these panels.

Easton *et al.*, (2015) discussed how to decide which genes confer a high risk of breast cancer and should therefore be included in gene panels to test for breast cancer risk. The authors concluded that more association studies needed to be carried out so that tests could be as robust as possible. Six years later, a gene panel with 34 susceptibility genes for breast cancer was sequenced in over 60,000 cases and more than 53,000 controls to establish which of the expanding number of candidate genes were clearly associated with an increased risk and to determine the level of risk conferred by mutations in the individual genes (Dorling *et al.*, 2021). The study found that protein-truncating variants in the *BRCA1*, *BRCA2, ATM, CHEK2* and *PALB2* genes conferred a high risk of breast cancer (p-value<0.0001) with odds ratios (ORs) ranging between 2.10 and 10.57. Protein-truncating variants in *BARD1*, *RAD51C*, *RAD51D* and *TP53* conferred a moderate risk (ORs: 1.8-2.25, p-value<0.05). In addition, rare, missense variants in *ATM*, *BRCA1*, *CDH1*, *TP53* and *CHEK2* showed evidence of breast cancer risk. New panels should be designed to include genes that previously were not included.

Panels like these can be used as screening tools to identify carriers of high-risk variants such that physicians can offer prophylactic treatment to patients with breast cancer to reduce the risk of developing breast cancer a second time. Despite the importance of these high-risk genes for genetic testing, they have been reported to be associated with the development of breast cancer in only 5-7% of cases (Melchor and Benítez, 2013; Lukong *et al.*, 2017), leaving the remaining patients without insight into the causality of their cancer. The majority of these remaining cases are sporadic and occur in the general population without evidence of significant family history. They generally result from a combination of low effect genetic risk variants individually, in combination with environmental risk factors. The genetic risk is conferred by the presence of multiple genetic variants such as single-nucleotide polymorphisms (SNPs), which are relatively common in the population being studied, with allele frequencies of 1% or more. Each variant individually confers only a small increase in the risk of disease, but the presence of many such variants together significantly increases the risk of developing breast cancer (Weeks and Lathrop, 1995; Maxwell and Nathanson, 2013).

The contribution of common variants to breast cancer risk has been investigated through the use of genome-wide association studies (GWAS), with at least 172 known genetic loci associated at genome-wide significance with the risk of breast cancer (Lilyquist *et al.*, 2018). Genome-wide heritability estimates suggest that the known loci explain only about 40% of the heritability conferred by all common variants on genome-wide SNP arrays (Mavaddat *et al.*, 2019). Identifying these genes and variants can be used to translate to the clinical landscape by informing early detection and prevention practices for those at higher genetic risk for developing breast cancer (Dalivandan *et al.*, 2021).

1.7. Genome-Wide Association Studies

Genome-wide association studies (GWAS) are used to identify common genetic variants that collectively contribute to disease risk (Bush and Moore, 2012). GWAS can lead to identifying new genes and variants associated with a phenotype and thus may contribute to the biological mechanism of the susceptibility to disease. The GWAS approach draws on the common disease/common variant hypothesis that numerous relatively common variants distributed across the genome increase the risk of disease (Duncan *et al.*, 2019a).

GWAS do not assume as to which genes or variants are associated with a disorder. Genetic variants across the genome are genotyped in great numbers and each variant is tested for association with the phenotype. A case-control study design is commonly used and the strength of GWAS is that millions of genetic variants spread across the entire genome are screened to detect variants which differ in frequency between cases and controls, thereby identifying variants that are associated with the disease phenotype (Dewan et al., 2006; Burton et al., 2007).

However, if using a genotyping array, not every variant is genotyped in the genome, therefore, GWAS rely on the use of linkage disequilibrium (LD) to provide better coverage from the subset

of SNPs which are genotyped. LD is the measure of non-independence of two alleles at different locations, i.e. when two alleles tend to be inherited or observed together more often than expected (Pritchard and Przeworski, 2001). Many SNPs are inherited together in a block (or haplotype) so that genotyping one of the variants (a tag SNP) in the block can inform the genotypes of the other SNPs in the block without having to genotype them all. A reference panel is used to infer the additional genotypes through the process of imputation. Genotyping SNPs remains the preferred method for GWAS at present as it is cheaper and much less computationally intensive than candidate gene or regional DNA sequencing (Hindorff *et al.*, 2018). Alternatively, whole genome sequencing (WGS) can be used where every base in the genome is sequenced to identify not only SNPs but also structural variants (Nakagawa and Fujita, 2018; Rossing *et al.*, 2019).

The majority of GWAS across all phenotypes have been performed in non-African populations. Figure 1.2 illustrates the relative proportions of GWAS participants from different regions/ethnicities. Over 70% of all GWAS have been performed using cohorts with predominantly European participants. Although there is substantial emerging literature on GWAS in African-American populations (i.e. partial African ancestry) and other ethnic groups, there remains a paucity of studies in populations resident in sub-Saharan Africa (Huo *et al.*, 2016; Palmer *et al.*, 2020; Hayat *et al.*, 2021).

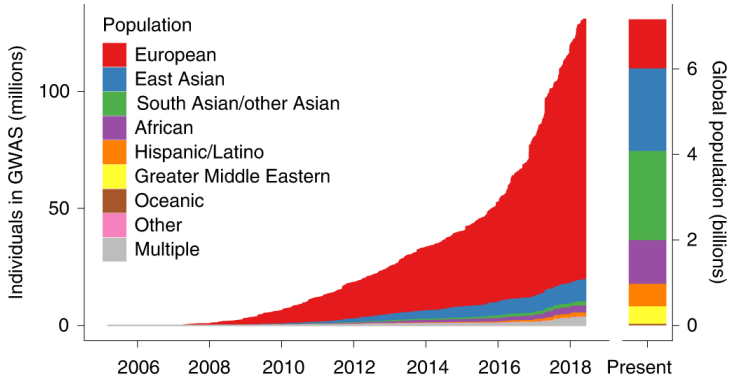


Figure 0.2: Ancestry of GWAS participants over time, as compared with the global population. Approximately 79% of all GWAS participants are of European descent, despite Europeans making up only 16% of the world population (Martin et al., 2019).

Many studies have explored whether genetic signals associated with a wide variety of phenotypes in European populations transfer to populations of African ancestry. A detailed study of the transferability of associated DNA markers for lipid traits, for example, detected in predominantly European populations found that many well-established signals did not replicate in African populations (Kuchenbaecker *et al.*, 2019). However, a recent study of lipid traits which included large sample sizes across multiple African populations showed that transferability to African populations increased with the size of the replication cohort, as did loci with stronger effect sizes in the European cohorts (Choudhury *et al.*, 2022).

Nonetheless, there are several other possible explanations for the lack of replication of genetic signals discovered in European studies to African populations. One is that some of the causal variants arose after the out-of-Africa migrations. Another is that, in general, there is lower LD across African genomes so the causal variants may be in weaker LD with SNPs that tag the association well in European populations. Also, some of the causal variants may have different frequencies or be specific to African populations and thus are not detected in European GWAS (Choudhury *et al.*, 2020; Sengupta *et al.*, 2021). This highlights the pressing need to perform GWAS in resident African populations using SNP arrays designed to consider these factors and capture African genetic diversity.

To help address these issues, the Human, Heredity and Health in Africa (H3Africa) consortium was started, funded by the National Institutes of Health (USA) and the Wellcome Trust (UK). (https://h3africa.org/). H3Africa projects have collected samples for over 70,000 participants across Africa (Mulder *et al.*, 2018). One of these is the Africa Wits-INDEPTH partnership for Genomic studies (AWI-Gen) (https://h3africa.org/index.php/awi-gen/). The purpose of the AWI-Gen study is to investigate genomic and environmental factors that influence body composition, body fat distribution and cardiometabolic disease risk in middle-aged African populations. AWI-Gen has recruited and genotyped more than 12,000 resident African participants from six different study sites across four African countries using the H3Africa Custom SNP genotyping array, which is yielding new insights into genetic factors associated with African cardiometabolic

diseases and related traits (Boua *et al.*, 2022; Choudhury *et al.*, 2022). One of the study sites was from the Soweto district of Johannesburg in South Africa and provided a set of female controls for the genetic studies to be described in this thesis.

1.8. Breast Cancer GWAS in non-African populations

GWAS in non-African populations have identified many variants that affect breast cancer risk. The first GWAS in breast cancer was carried out in 2007 by Easton *et al.* (Easton *et al.*, 2007). This analysis was done in stages, where first, a large number of SNPs were genotyped in a small cohort. Any SNPs of interest were selected and genotyped in a larger sample set. In the first stage, a panel of 266,722 SNPs was genotyped in 408 cases with a family history of breast cancer and 400 controls of Caucasian ancestry. In the second stage, 12,711 SNPs from stage 1 were genotyped in 3,990 cases and 3,916 controls. In the third stage, 30 significant SNPs were tested for confirmation in 21,860 cases and 22,578 controls, which identified 5 novel, independent loci strongly associated with breast cancer, four of which contained plausible candidate genes (*FGFR2, LSP1, TNRC9* and *MAP3K1*). The *FGFR2* locus was consistently replicated in further studies (Hunter *et al.*, 2007; Stacey *et al.*, 2007; Antoniou *et al.*, 2008; Garcia-Closas *et al.*, 2008), showing an increased risk for breast cancer and supporting the validity of association studies in discovering risk genes.

A study by the Breast Cancer Association Consortium (BCAC), which included 37,012 cases and 40,069 controls in the second stage of association testing, discovered two new risk loci on chromosome 3p24 (close to *NECK10* and *SLC4A7*) and chromosome 17q23.2 (close to *COX11*) (Ahmed *et al.*, 2009). The Shanghai Breast Cancer study discovered an association at the *ESR1* locus which was also replicated in European women (Zheng *et al.*, 2009). Shortly after this, five new risk loci were identified, including one in the gene *LSP1* (Turnbull *et al.*, 2010). Subsequently, an association at a locus on 9q31.2 was found in a study of 1,694 European cases and 2,365 European controls (Fletcher *et al.*, 2011). Ghoussaini *et al.* investigated 72 previously associated hits from two different studies (Ghoussaini *et al.*, 2013). They identified three new

loci; at 12p11 (near *PTHLH*, associated with overall breast cancer) and at 12p21 and 21q21 (near *NRIP1*) associated with ER-positive breast cancer.

More loci associated with breast cancer were discovered by using larger sample sizes, genotyping more SNPs and doing meta-analyses across studies. This approach was exemplified by the Collaborative Oncologic Gene-environment Study (COGS) (Michailidou *et al.*, 2013). The COGS study included multiple consortia and was designed to study prostrate, breast and ovarian cancer using an SNP genotyping array from Illumina (iCOGS) that genotyped 200,000 SNPs. These SNPs were genotyped in 45,290 cases and 41,880 controls of European ancestry from 41 studies in the Breast Cancer Association Consortium (BCAC). This identified 41 new breast cancer risk loci (Michailidou *et al.*, 2013), bringing the total number of replicated loci to 68. Michailidou *et al.* later performed a meta-analysis that included 11 GWAS to identify 15 new breast cancer risk loci (Michailidou *et al.*, 2015).

A more recent breast cancer GWAS included 570,000 SNPs using the Illumina OncoArray BeadChip (Amos *et al.*, 2017). The genotyped SNPs were imputed to approximately 11.8 million SNPs for analysis (Michailidou *et al.*, 2017). The results of 61,282 breast cancer cases and 45,494 controls were used in a meta-analysis with results from iCOGS and 11 other GWAS studies (Michailidou *et al.*, 2017). There were 122,977 cases and 105,974 controls of European ancestry, and 14,068 cases and 13,104 controls of East Asian ancestry. This study identified 65 new risk loci at genome-wide significance and replicated 49 previously identified loci. It should be noted that the majority of these studies included mostly European individuals and some individuals of East Asian ancestry. A review by Lilyquist *et al.*, (2018) summarises all 172 risk SNPs, which account for approximately 18% of relative risk. More recently, a fine-mapping study which combined association data with functional annotations identified 205 independent risk loci which co-localised with 191 potential target genes (Fachal *et al.*, 2020).

1.9. GWAS in African American populations

Given the differences in genomic architecture between individuals of European and African ancestry, and within African populations themselves (Choudhury *et al.*, 2020), the paucity of data in these populations should lead to more GWAS being done to discover common and African-specific breast cancer risk variants. While there is no breast cancer GWAS comprising only resident sub-Saharan African populations (Hayat *et al.*, 2021), several association studies have been carried out in African American women. Some of these investigated known risk loci and candidate genes in African American populations. In 2013, 67 known breast cancer risk loci were investigated in a modest sample size of 1,231 cases and 2,069 controls (Long *et al.*, 2013). Only seven SNPs were potentially associated (p-value<0.05) with breast cancer risk, suggesting that not all breast cancer risk loci in European women are associated with African American women.

A study that included both African American women (cases=1,004) and European women (cases=1,718) was carried out by Haiman *et al.* which identified a new locus: *TERT-CLPTM1L* (Haiman *et al.*, 2011). This was a multi-ethnic study and African American-only GWAS analysis had not yet been done at that time. Three African American consortia for breast cancer were created: GWAS in Breast Cancer in the African Diaspora (ROOT), African American Breast Cancer (AABC), and African American Breast Cancer Epidemiology and Risk (AMBER) (Chen *et al.*, 2011; Huo *et al.*, 2012; Palmer *et al.*, 2014). The three consortia included a combined total of just under 11,000 cases and approximately 12,000 controls. The ROOT and AABC consortia were included in a GWAS (Huo *et al.*, 2012; Chen *et al.*, 2013), where no variants were found to be associated with breast cancer at a genome-wide significance level. Three SNPs were found to be associated with breast cancer at genome-wide significance in a meta-analysis (Huo *et al.*, 2016). An SNP upstream of *TNFSF10* (rs13074711) was significantly associated (p-value<$5x10^{-08}$) with ER-negative breast cancer, as was rs10069690 in *TERT*. The third SNP, rs12998806 on chromosome 2, was significantly associated with the risk of ER-positive breast cancer.

Finally, Ruiz-Narvaez *et al.* (2016) used admixture mapping to identify two known loci associated with breast cancer, as well as two novel, regional loci associated with breast cancer in women of African ancestry (Ruiz-Narváez *et al.*, 2016). The two novel associations, 4p16.1 (p-value4.00x10^{-06}) and 17q25.1 (p-value1.6x10^{-04}) were associated with ER-positive breast cancer by admixture mapping criteria.

Several recent studies have been carried out that include women of African ancestry. A cross-ancestry approach which used a meta-analysis of African ancestry cohorts with European ancestry cohorts from the very large BCAC study reported the discovery of four loci associated with breast cancer risk, with relatively modest contributions from the African cohorts (Adedokun *et al.*, 2021). A candidate gene study from the AMBER consortium reported limited evidence of association with anti-Mullerian hormone-related genes (Nichols *et al.*, 2021).

1.10. Breast Cancer GWAS in sub-Saharan Africa

In contrast to the existing GWAS carried out in individuals of African ancestry, there are no reports of GWAS done on breast cancer in resident sub-Saharan African populations. Several candidate gene studies and attempts at replication of specific SNPs associated with breast cancer have been done (Hayat *et al.*, 2021). However, these studies had several limitations. They lacked the power to detect associations with modest effect sizes, and they limited their search to a few SNPs or a few select genes. In addition, no GWAS have been reported on breast cancer in South African black women. A review (Hayat *et al.*, 2021) summarises the genetic studies done on breast cancer in sub-Saharan African populations and highlights that there are no GWAS in resident sub-Saharan African populations. These studies will be covered in detail in Chapter 2.

1.11. Polygenic Risk Scores

An exciting and relatively recent development of post-GWAS analysis has been the development of polygenic risk scores (Torkamani *et al.*, 2018). A GWAS identifies many SNPs that, individually, have a small effect on the phenotype, limiting their predictive power (Choi *et al.*, 2020). GWAS were found to identify mutations associated with a phenotype, however, there was always the matter of missing heritability, the SNPs that were associated with a phenotype at a significant p-value did not completely explain the heritability of the trait (Maher, 2008). In 2010, Yang *et al.* found that by combining many SNPs that have a small effect, the risk of developing a particular phenotype can be estimated (Yang *et al.*, 2010). This led to the development of polygenic risk score methods (Dudbridge, 2013). Conventionally, a polygenic risk score computes a score by taking the sum of the risk alleles of each individual in a GWAS and applying a weight by the size of the effect of an individual SNP on the phenotype (Dudbridge, 2013). Studies have shown that polygenic risk score calculations using SNPs other than those that pass the genome-wide significance p-valuethreshold can be more predictive (Mavaddat *et al.*, 2019). Different methods have been developed to create polygenic risk scores, such as clumping and pruning methods whereby SNPs are selected based on LD (i.e. SNPs in a specified window are removed if they fall below a certain r^2 value) and calculating the score by including the SNPs that meet a particular p-valuethreshold for association with the phenotype. This method is used by PLINK (Chang *et al.*, 2015) and PRSice-2 (Choi and O'Reilly, 2019) to compute polygenic risk scores. LDpred uses a different method (Vilhjálmsson *et al.*, 2015) by calculating the score without using a subset of SNPs and only using summary statistics (effect sizes) and LD information.

Using the clumping and pruning method in an African population is the more suitable option. African populations have higher levels of genomic diversity and smaller LD blocks compared to non-African popuatlions. Populations may also exhibit different MAF for SNPs that are associated with a trait in a European population. To illustrate this, Yengo, *et al* (2020) demonstrated that LD and allele frequency differences explained 70-80% of the loss of value of the relative accuracy of European derived PRSs in an African population (Wang *et al.*, 2020). Therefore, taking into account LD patterns in a population can greatly aid in selecting

appropriate SNPs to create a risk score as opposed to only selecting those with a low p-value. A new tool called PRS-SCx which has been developed for multi-ethnic PRSs leverages on LD patterns which highlights the importance of taking this into account when developing a PRS and demonstrating its efficacy (Ruan *et al.*, 2022).

It must be noted that most GWAS have been carried out in non-African populations, and therefore most polygenic risk scores show better predictability in these populations. Applying polygenic risk scores trained in a European population to another European population generally is expected to provide good predictability of risk. However, applying the same risk score to an African population produces poor levels of risk prediction (Cavazos and Witte, 2021). A study of lipid traits showed that polygenic risk scores derived from data from African American individuals improved prediction in sub-Saharan African populations compared to European and multi-ancestry scores (Kamiza *et al.*, 2022). This again highlights the need for more GWAS in sub-Saharan African populations so that risk scores can be developed in African populations to improve predictability.

The potential of a polygenic risk score for the stratification of disease risk is exemplified by studies in breast cancer. Mavaddat *et al*. developed a polygenic risk score using 313 SNPs and estimated that the lifetime absolute risk by age 80 years for ER-positive breast cancer ranged from 2% for women in the lowest centile of the polygenic risk score to 31% in the highest centile (Mavaddat *et al.*, 2019). Risk prediction in breast cancer can be further improved by the inclusion of a range of risk factors such as family history, rare pathogenic variants in cancer susceptibility genes, polygenic risk, lifestyle, hormonal and imaging risk factors in the analysis, and a web-based tool (CanRisk) has been developed for this purpose (Carver *et al.*, 2021).

This PhD research project is motivated by the lack of knowledge on the genetics of breast cancer in resident African populations, specifically in the South African black population. The information derived from this study has the potential to identify genetic risk loci for breast

cancer in an African population, some of which may be African-specific. This in turn could lead to a better understanding of breast cancer aetiology in South African black women. Furthermore, the results from this study could form the basis for the future development of polygenic risk scores to identify, in combination with other known risk factors, individuals who are at increased risk of developing breast cancer.

1.12. Study Aim and Objectives

The aim of this study was to identify common genetic variants that contribute to the risk of developing breast cancer in South African black populations. The specific objectives were:

1. To identify genetic risk factors for breast cancer in a South African black population through a GWAS approach using the H3Africa SNP array.
2. To perform fine-mapping of African breast cancer risk loci to identify potentially causal genes and variants.
3. To test known breast cancer genetic risk loci identified in other populations for association with breast cancer in South African black populations.
4. To test novel genetic risk factors associated with breast cancer in South African black populations for replication in other populations.
5. To test breast cancer PRSs generated in other populations in a South African black population.
6. To carry out a meta-analysis with data from an African American population.

In chapter 2 of this thesis, all association studies and GWAS for breast cancer that have been done in sub-Saharan Africa will be reviewed. This is followed by Chapter 3 which describes the methods used to carry out this GWAS in a South African black population. Chapter 4 will present the results of the study, and finally, in chapter 5 the value and limitations of the results will be discussed and a way forward suggested for future studies.

Chapter 2: Breast Cancer Genetics in sub-Saharan African Populations

This chapter is composed of two parts. Sections 2.1 – 2.4 are the scoping review I published on this topic in JCO Global Oncology in 2021, which covered papers published up to 01 April 2020. Section 2.5 is an updated description of papers published between 02 April 2020 and 22 July 2022.

2.1 Introduction

Breast cancer is the most common cancer among African women, and is the second most common cause of cancer deaths in Africa as estimated by Globocan 2018 (Bray *et al.*, 2018). Cancer incidence is increasing in Africa (Bray *et al.*, 2018), and is associated largely with lifestyle changes resulting from increased urbanisation (Lukong *et al.*, 2017). Ferlay *et al* (Ferlay *et al.*, 2019) predicted a 49% increase in breast cancer incidence across Africa between 2018 and 2040, an increase faster than any other region in the world. The International Association of Cancer Registries (IACR) cancer incidence report includes data for only 20 of 46 Sub-Saharan African countries (Wild *et al.*, 2020), and it is likely that the incidence of breast cancer is under-reported owing to the lack of infrastructure for cancer diagnosis in resource-constrained countries in Sub-Saharan Africa (Cumber *et al.*, 2017). Also, owing to the large burden of communicable and other non-communicable diseases in SSA, most resources are directed at controlling these diseases (Roth *et al.*, 2018; Gouda *et al.*, 2019). Cancer research, especially molecular research, in Sub-Saharan Africa is often deprioritized against other pressing public health concerns (Ngoma and Ngoma, 2019).

The epidemiology of breast cancer is complex and the causes are multifactorial, with genetic risk factors and the environment playing important roles in cancer development (Momenimovahed and Salehiniya, 2019). Genetic risk factors contributing to the development of breast cancer include rare, pathogenic variants (Dorling *et al.*, 2021) or many common variants that, in

combination, increase the risk of developing breast cancer (Maxwell and Nathanson, 2013). Familial or hereditary breast cancer is associated with rare variants which significantly increase the risk of developing breast cancer, and occur in genes such as *BRCA1, BRCA2*, *PTEN*, *FGFR2*, *CHEK2*, *TP53*, *PALB2*, *ATM* and *XRCC2*. Hereditary breast cancer accounts for only 5-7% of all cases of breast cancer (Lukong *et al.*, 2017). The contribution of rare variants with moderate to high penetrance for breast cancer risk has been reviewed recently by Nguyen *et al.*, (2019). The other category is breast cancer which occurs in the general population without evidence of significant family history and results from a combination of genetic and environmental risk factors. The genetic risk is conferred by the presence of multiple genetic variants such as single nucleotide polymorphisms (SNPs), which are relatively common in the population being studied, with allele frequencies of 1% or more.

Each variant individually confers only a small increase in the risk of disease but the presence of many such variants together significantly increases the risk of developing breast cancer (Weeks and Lathrop, 1995; Maxwell and Nathanson, 2013). The contribution of common variants to breast cancer risk has been investigated through the use of genome-wide association studies (GWAS), with at least 172 known genetic loci associated at genome-wide significance with the risk of breast cancer (Lilyquist *et al.*, 2018).

The majority of studies into the molecular aetiology of breast cancer have been done in non-African populations, although there is a substantial emerging literature on the genetics of breast cancer in African-American populations (Huo *et al.*, 2016; Palmer *et al.*, 2020). However, published data on resident black African populations from continental Africa are much more limited. This scoping review aims to summarise genetic studies on breast cancer risk that have been conducted in resident black Sub-Saharan African populations and to identify potential gaps in knowledge and research opportunities.

2.2 Methods

The research question for this review was: "What knowledge exists on the association of germline genetic variants with familial or sporadic breast cancer in black populations resident in sub-Saharan Africa?". This review follows the framework for scoping studies according to Arksey, H, et al (2005) (Arksey and O'Malley, 2005). We chose to carry out a scoping review rather than a systematic review based on the criteria described in Munn *et al.*, (2018). A scoping review is more appropriate since there is a limited body of evidence in Sub-Saharan Africa, and our purpose was to scope the available evidence on breast cancer genetics in Sub-Saharan Africa and to identify gaps in knowledge and limitations in methodological approaches.

PubMed and SCOPUS were used to search the literature and only original research articles written in English were included. All studies published on or before the 1st of April 2020 were included. Review articles were excluded. Only studies including data from Sub-Saharan African populations were included. A comprehensive combination of keywords was used to identify genetic studies on participants resident in Sub-Saharan African countries (Table 2.1).

Table 0.1. Keywords and combinations of keywords used to search the literature using PubMed

Search no.	Keywords	Field searched
#1	breast cancer and variant	Title/Abstract
#2	breast cancer and variants	Title/Abstract
#3	breast cancer and mutation	Title/Abstract
#4	breast cancer and mutations	Title/Abstract
#5	breast cancer and germline	Title/Abstract
#6	breast cancer and hereditary	Title/Abstract
#7	breast cancer and inherited	Title/Abstract
#8	breast cancer and genetic	Title/Abstract
#9	breast cancer and genetics	Title/Abstract
#10	breast cancer and genomic	Title/Abstract
#11	breast cancer and genomics	Title/Abstract
#12	breast cancer and genome-wide	Title/Abstract

#13	breast cancer and GWAS	Title/Abstract
#14	breast cancer and associated	Title/Abstract
#15	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14	
#16	(africa) OR (african) OR (angola) OR (benin) OR (botswana) OR (burkina faso) OR (burundi) OR (cameroon) OR (central african republic) OR (chad) OR (comoros) OR (congo) OR (democratic republic of congo) OR (DRC) OR (djibouti) OR (equitorial guinea) OR (eritrea) OR (ethiopia) OR (gabon) OR (gambia) OR (ghana) OR (guinea) OR (bissau) OR (ivory coast) OR (kenya) OR (lesotho) OR (liberia) OR (malawi) OR (mali) OR (mauritania) OR (mozambique) OR (namibia) OR (niger) OR (nigeria) OR (rwanda) OR (senegal) OR (sierra leone) OR (somalia) OR (south africa) OR (sudan) OR (swaziland) OR (tanzania) OR (togo) OR (uganda) OR (zaire) OR (zambia) OR (zimbabwe) OR (central africa) OR (west africa) OR (east africa) OR (southern africa) OR (sub saharan africa), Field: Title/Abstract	
#17	#15 AND #16	

The entire search looked like this:

((africa[Title/Abstract]) OR (african[Title/Abstract])) OR (angola[Title/Abstract])) OR (benin[Title/Abstract])) OR (botswana[Title/Abstract])) OR (burkina faso[Title/Abstract])) OR (burundi[Title/Abstract])) OR (cameroon[Title/Abstract])) OR (central african republic[Title/Abstract])) OR (chad[Title/Abstract])) OR (comoros[Title/Abstract])) OR (congo[Title/Abstract])) OR (democratic republic of congo[Title/Abstract])) OR (DRC[Title/Abstract])) OR (djibouti[Title/Abstract])) OR (equitorial guinea[Title/Abstract])) OR (eritrea[Title/Abstract])) OR (ethiopia[Title/Abstract])) OR (gabon[Title/Abstract])) OR (gambia[Title/Abstract])) OR (ghana[Title/Abstract])) OR (guinea[Title/Abstract])) OR (bissau[Title/Abstract])) OR (ivory coast[Title/Abstract])) OR (kenya[Title/Abstract])) OR (lesotho[Title/Abstract])) OR (liberia[Title/Abstract])) OR (malawi[Title/Abstract])) OR (mali[Title/Abstract])) OR (mauritania[Title/Abstract])) OR (mozambique[Title/Abstract])) OR (namibia[Title/Abstract])) OR (niger[Title/Abstract])) OR (nigeria[Title/Abstract])) OR (rwanda[Title/Abstract])) OR (senegal[Title/Abstract])) OR (sierra leone[Title/Abstract])) OR (somalia[Title/Abstract])) OR (south africa[Title/Abstract])) OR (sudan[Title/Abstract])) OR (swaziland[Title/Abstract])) OR (tanzania[Title/Abstract])) OR (togo[Title/Abstract])) OR (uganda[Title/Abstract])) OR (zaire[Title/Abstract])) OR (zambia[Title/Abstract])) OR (zimbabwe[Title/Abstract])) OR (central africa[Title/Abstract])) OR (west africa[Title/Abstract])) OR (east africa[Title/Abstract])) OR (southern africa[Title/Abstract])) OR (sub saharan africa[Title/Abstract])) AND ((((((((((((breast cancer[Title/Abstract]) AND (variant[Title/Abstract])) OR ((breast cancer[Title/Abstract]) AND (variants[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (mutation[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (mutations[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (germline[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (hereditary[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (inherited[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (genetic[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (genetics[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (GWAS[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (genome wide[Title/Abstract]))) OR ((breast cancer[Title/Abstract]) AND (association[Title/Abstract])))

2.2.1 Article selection and data extraction

The selection of appropriate studies was carried out according to the following criteria: original research articles that studied inherited variants associated with breast cancer in resident Sub-Saharan African populations. Screening of titles and abstracts for all papers was done independently by the lead authors (MH & WCC) and papers were included based on consensus agreements between them. If there was uncertainty in deciding whether a study was relevant, the full article was read to make a final decision. Studies of South African women of European, Indian or mixed ancestry were excluded. If the abstract or full text of a study could not be accessed, it was excluded from this study. The following information was extracted: PMID, year of study, population, sample size, key findings (variants, genes), the predicted pathogenicity of variants, methodology and a brief summary of the study. Rare variants were considered of interest if predicted to be deleterious by variant effect predictor tools and/or significantly enriched in breast cancer cases compared to controls.

2.3 Results

Using the comprehensive list of keywords to search, PubMed identified a total of 1,029 articles (Figure 2.1). Scanning the titles and the abstracts led to exclusion of 1002 studies as they were not breast cancer genetic studies, did not include samples from Sub-Saharan Africa, or were reviews, leaving 27 which met the broad inclusion criteria. A further four studies were excluded after full-text evaluation as they were conducted in African-American populations and our aim was specifically to review studies of breast cancer genetics in sub-Saharan African populations. One study was removed as it only investigated tumour DNA. The search in SCOPUS did not yield any additional eligible articles.

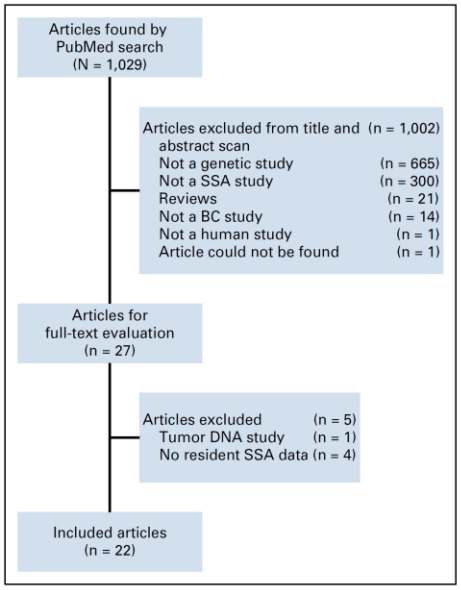


Figure 0.1: Process of selection of papers.
Process of selection of papers. The selection criteria were germline genetic studies focusing on BC and carried out in a sub-Saharan African population(s). BC, breast cancer; SSA, sub-Saharan Africa.

The remaining 22 papers were included and assessed in this review (Gao *et al.*, 2000; Okobia *et al.*, 2008, 2009; Zhang *et al.*, 2009, 2012; Diez *et al.*, 2011; van der Merwe *et al.*, 2012; Fackenthal *et al.*, 2012; Mvila *et al.*, 2014; Francies *et al.*, 2015; Zheng *et al.*, 2018; Zoure *et al.*, 2018; Habyarimana, *et al.*, 2018a, 2018b, 2018c; 2018; Ahmed *et al.*, 2019; Kiendrebeogo *et al.*, 2019; Sagna *et al.*, 2019; Diop *et al.*, 2019; Adedokun *et al.*, 2020; Ndiaye *et al.*, 2020; Torrorey-Sawe *et al.*, 2020). Of these, 16 were rare variant studies (Gao *et al.*, 2000; Zhang *et al.*, 2009, 2012; Diez *et al.*, 2011; van der Merwe *et al.*, 2012; Fackenthal *et al.*, 2012; Mvila *et al.*, 2014; Francies *et al.*, 2015; Zheng *et al.*, 2018; Zoure *et al.*, 2018; Habyarimana, Attaleb, *et al.*, 2018b; Diop *et al.*, 2019; Sagna *et al.*, 2019; Adedokun *et al.*, 2020; Torrorey-Sawe *et al.*, 2020; Ndiaye *et al.*, 2020), and the remaining six investigated a few common variants in small sample sizes (Okobia *et al.*, 2008, 2009; Habyarimana, Attaleb, *et al.*, 2018a; Habyarimana, Bakri, *et al.*, 2018; Ahmed *et al.*, 2019; Kiendrebeogo *et al.*, 2019). The studies used either protein truncation tests, variant-specific PCR assays, sequencing of genes, exons or coding regions or gene panel sequencing to identify variants. There were no GWAS among the 22 papers. The studies in this review included populations from 10 of the 46 Sub-Saharan African countries: Burkina Faso (Zoure *et al.*, 2018; Kiendrebeogo *et al.*, 2019; Sagna *et al.*, 2019), Cameroon (Adedokun *et al.*, 2020), Democratic

Republic of Congo (DRC) (Mvila *et al.*, 2014), Ethiopia (Ahmed *et al.*, 2019), Kenya (Torrorey-Sawe *et al.*, 2020), Nigeria (Gao *et al.*, 2000; Okobia *et al.*, 2008, 2009; Zhang *et al.*, 2009, 2012; Fackenthal *et al.*, 2012; Zheng *et al.*, 2018), Rwanda (Habyarimana, Attaleb, *et al.*, 2018b, 2018a; Habyarimana, Bakri, *et al.*, 2018), Senegal (Diez *et al.*, 2011; Diop *et al.*, 2019; Ndiaye *et al.*, 2020), South Africa (van der Merwe *et al.*, 2012; Francies *et al.*, 2015) and Uganda (Adedokun *et al.*, 2020).

2.3.1 The contribution of rare variants to breast cancer risk in Sub-Saharan Africa

Among the 22 studies identified, 16 were rare variant studies which involved the screening of exons of genes of interest for variants with moderate to high penetrance and low population frequencies. A summary of the rare variant studies including countries of origin of participants and sample sizes is shown in Table 2.2. Probable pathogenic variants which have been detected in two or more patients are listed in Table 2.3. Some of the rare variant studies (6/16) genotyped only a few specific variants which had been identified previously in other populations. In 14 of the 16 studies variants were identified in one or more known breast cancer susceptibility genes, with sample sizes ranging from 1 to 1,136 breast cancer patients. The largest and most comprehensive of these, which built on previous studies in the Nigerian population, used the BROCA gene panel (Shirts *et al.*, 2016) to sequence the exons of 25 genes for variants in 1,136 Nigerian women with breast cancer and in 997 Nigerian female controls (Zheng *et al.*, 2018). Variants predicted to be deleterious were detected in 16 genes, including *BRCA1*, *BRCA2*, *PALB2* and *TP53*. Deleterious variants were detected in 14.7% of breast cancer patients and 1.8% of controls. Carriers of mutations in *BRCA1* and *TP53* had a significantly younger age at diagnosis than non-carriers, and *BRCA1* mutation carriers were more likely to have triple-negative breast cancer (TNBC). The majority of patients for whom tumour stage was available were stage III (41.8%) or stage IV (42.6%), but no association of tumour stage with mutation status was reported. Four other previous studies of *BRCA1* and *BRCA2* in Nigerian breast cancer patients are summarised in Table 2.2 and mutations detected are listed in Table 2.2.

Table 0.2. Studies of rare variants in breast cancer genes in Sub-Saharan Africa

Methodology	Country	BC cases	Controls	Variants detected[a]	Reference
Whole exome sequencing	Kenya	13	0	1	(Torrorey-Sawe *et al.*, 2020)
BRCA1/BRCA2, Sanger sequencing	Senegal	27	90	1	(Ndiaye *et al.*, 2020)
CHEK2 sequencing	Rwanda	41	42	0	(Habyarimana, *et al.*, 2018c)
BRCA1/*BRCA2* sequencing	Nigeria	434	0	29	(Fackenthal *et al.*, 2012)
BRCA1/2 testing , founder variants, PTT	South Africa (Xhosa)	16	0	1	(van der Merwe *et al.*, 2012)
Allele-specific PCR	Nigeria	365	117	1	(Zhang *et al.*, 2009)
SSCP and protein truncation test	Nigeria	70	0	3	(Gao *et al.*, 2000)
BRCA2 exon sequencing	Senegal	33	96	1	(Diop *et al.*, 2019)
Gene panel sequencing	Cameroon & Uganda	196	185	34	(Adedokun *et al.*, 2020)
Sequencing of coding regions of genes	Burkina Faso	9	0	0	(Sagna *et al.*, 2019)
BROCA gene panel sequencing	Nigeria	1136	611	106	(Zheng *et al.*, 2018)
NGS of *BRCA1/2* and *PALB2*; MLPA	South Africa (Black)	85	0	6	(Francies *et al.*, 2015)
BRCA1/*BRCA2* sequencing	DRC	3	0	1	(Mvila *et al.*, 2014)
BRCA1 sequencing	Senegal	1	0	1	(Diez *et al.*, 2011)
BRCA1 sequencing (exons 2, 5, 11)	Burkina Faso	15	0	0	(Zoure *et al.*, 2018)
SNAPSHOT and sequencing	Nigeria	356	0	12	(Zhang *et al.*, 2012)

BC, breast cancer; MLPA, multiplex ligation–dependent probe amplification; PCR, polymerase chain reaction; SSA, sub-Saharan Africa;
SSCP; single-strand conformation polymorphism.
[a]Number of pathogenic variants detected.

A South African study sequenced *BRCA1*, *BRCA2* and *PALB2* from several ancestral groups, which included 85 South African black breast cancer patients (Francies *et al.*, 2015), and found pathogenic variants in *BRCA1*, *BRCA2* and *PALB2* in this ethnic subset. The group of TNBC patients had a higher mutation frequency than the group of patients diagnosed before the age of 50. In another South African study which included 16 women of Xhosa ancestry with breast cancer, the same deleterious variant in *BRCA2* (p.Ile1924ArgTer) was detected in four of 16 patients and is suggested to be a founder variant in both black and mixed ancestry populations in the Western Cape Province (van der Merwe *et al.*, 2012). A study of breast cancer patients from Cameroon and Uganda sequenced 12 known or candidate breast cancer genes in 196 cases and 185 controls unselected for age at diagnosis or family history (Adedokun *et al.*, 2020) and identified pathogenic or likely pathogenic variants in 15.8% of breast cancer cases and 1.6% of controls. Their findings included two patients with a mutation in *ATM* and single patients with mutations in *BARD1*, *CDH1* and *TP53*. One of three studies in Senegal identified a duplication of 10 nucleotides in the *BRCA1* gene, c.815_824dup10, in 15 index cases from families with hereditary breast cancer (Ndiaye *et al.*, 2020). This duplication was located on the same microsatellite haplotype in all cases and had previously been reported in breast cancer cases from France, Spain and the United States, particularly in people of African or Hispanic descent, and is thus a founder mutation likely to be of West African origin. Whole exome sequencing of 13 breast cancer patients in Kenya (Torrorey-Sawe *et al.*, 2020) detected five variants of uncertain clinical significance in *BRCA1* and *BRCA2* as well as a pathogenic variant in *BRCA2*. A study in the Democratic Republic of Congo detected a pathogenic *BRCA1* variant in all three affected members of a breast cancer family (Mvila *et al.*, 2014). However sequencing of *BRCA1* and *BRCA2* in breast cancer patients in Burkina Faso identified only benign variants in these genes (Zoure *et al.*, 2018; Sagna *et al.*, 2019), and a screen for germline variants in *CHEK2* in Rwandan breast cancer cases and controls were negative (Habyarimana, *et al.*, 2018c).

In summary, a total of 36 predicted pathogenic variants were detected in two or more breast cancer patients. These variants, together with their country of origin, are listed in Table 2.3. Some *BRCA1* variants were detected in up to 15 patients, and, as noted above, there are suggestive founder pathogenic variants in the Senegalese and South African populations.

Overall, pathogenic variants were detected in at least two patients in the *ATM*, *BARD1, BRCA1, BRCA2, PALB2* and *RAD51C* genes. The study of Zheng *et al.*, (2018) in Nigeria suggests that variants in several other genes may also contribute to breast cancer risk, but the mutations were too rare to draw firm conclusions on causality.

Table 0.3. Predicted Pathogenic Variants Detected in Two or More African Patients

Gene	**Variant**	**Variant effect**	**Number of cases**	**Country**	**Reference**
BRCA1	c.80+1G>A	splice site	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.133_134del	p.Lys45IleTer	5	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.191G>A	p.Cys64Tyr	3	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1			4	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.303T>G	p.Tyr101Ter	4	Nigeria	(Zhang *et al.*, 2012)
BRCA1			10	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.815_824dup	p.Thr276Ter	15	Senegal	(Ndiaye *et al.*, 2020)
BRCA1	c.838dup	p.Ala280GlyTer	3	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.1504_1508del	p.Leu502AlaTer	2	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1			3	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.1623dup	p.Asn542GluTer	3	Nigeria	(Zheng *et al.*, 2018)
BRCA1			4	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1	c.2192_2196del	p.Lys731ArgTer	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.2329del	p.Tyr777MetTer	4	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.2864C>A	p.Ser955Ter	3	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.3268C>T	p.Gln1090Ter	2	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1			4	Nigeria	(Zheng *et al.*, 2018)
BRCA1			2	Nigeria	(Gao *et al.*, 2000)
BRCA1	c.3637_3638del	p.Glu1213ArgTer	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.4122_4123del	p.Ser1374ArgTer	3	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1			3	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.4240dup	p.Leu1414ProTer	2	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1	c.4484G>T	p.Arg1495Met	3	Cameroon	(Adedokun *et al.*, 2020)

BRCA1	c.4986+6T>C	splice site	2	Cameroon	(Adedokun *et al.*, 2020)
BRCA1	c.4992C>T	splice site	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.5095C>T	p.Arg1699Trp	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.5177_5180del	p.Arg1726LysTer	2	Nigeria	(Zheng *et al.*, 2018)
BRCA1	c.5324T>G	p.Met1775Arg	3	Nigeria	(Fackenthal *et al.*, 2012)
BRCA1			7	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.93G>A	p.Trp31Ter	2	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.1310_1313del	p.Lys437IleTer	3	Nigeria	(Fackenthal *et al.*, 2012)
BRCA2			4	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.1796_1800del	p.Ser599Ter	3	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.2402_2412del	p.Asn801IleTer	2	Nigeria	(Fackenthal *et al.*, 2012)
BRCA2			3	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.2808_2811del	p.Ala938ProTer	2	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.5241_5242insTA	p.Ser1748Ter	2	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.5286T>G	p.Tyr1762Ter	2	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.5771_5774del	p.Ile1924ArgTer	4	South Africa (black)	(van der Merwe *et al.*, 2012)
BRCA2	c.8351+2T>C	splice site	3	Nigeria	(Zheng *et al.*, 2018)
BRCA2	c.8817_8820del	p.Lys2939AsnTer	2	Nigeria	(Fackenthal *et al.*, 2012)
BARD1	c.1396-1G>A	p.Tyr101Ter	2	Nigeria	(Zheng *et al.*, 2018)
ATM	c.7913G>A	p.Trp2638Ter	2	Cameroon	(Adedokun *et al.*, 2020)
PALB2	c.2485C>T	p.Gln829Ter	2	Nigeria	(Zheng *et al.*, 2018)
RAD15C	c.186_187del	p.Gln62HisTer	2	Nigeria	(Zheng *et al.*, 2018)

The number of predicted pathogenic variants found in each gene and the Sub-Saharan African country where the studies were conducted are summarised in Figure 2.2. The preponderance of variants of Nigerian origin reflects the much larger and more comprehensive variant studies that have been conducted on this population (Zheng *et al.*, 2018).

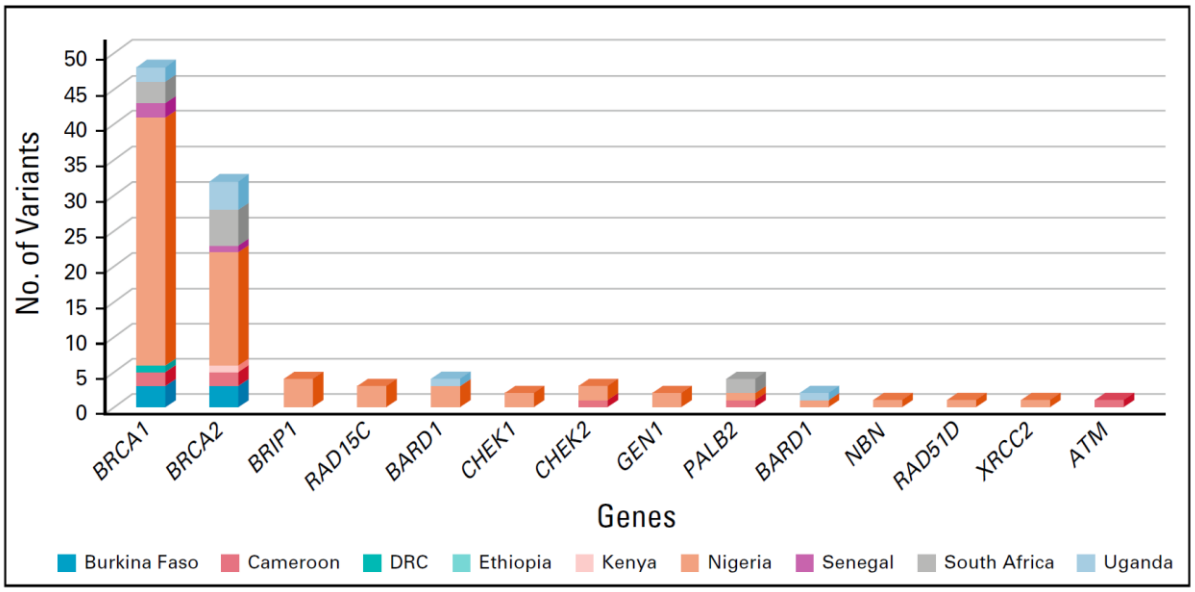


Figure 0.2: Variant distribution in genes and countries.
This shows how many predicted pathogenic variants were identified in breast cancer susceptibility genes in each reported sub-Saharan African country. DRC, Democratic Republic of Congo.

2.3.2 The contribution of common variants to breast cancer risk in Sub-Saharan Africa

Our search did not find any GWAS of breast cancer done exclusively in resident Sub-Saharan African populations. The genetic association studies carried out in Sub-Saharan African populations were candidate gene studies which investigated a limited number of common variants previously associated with breast cancer in other populations. These were genotyped in relatively small numbers of breast cancer cases and controls and tested for association with breast cancer and are summarised in Table 2.4. Okobia *et al.*, (2009) genotyped a *CYP1B1* polymorphism Val432Leu (c.1294C>G) in breast cancer cases and controls from Southern Nigeria and found an association of the heterozygous Val/Leu genotype with breast cancer (OR = 1.59, 95% CI 1.01-2.52). The same group genotyped the c.668A>G SNP (Gln223Arg) in the leptin receptor gene (*LEPR)* and reported a suggestive association of the 223Arg allele with disease risk in pre-menopausal women (OR = 1.8, 95% CI 1.0 – 3.2, p-value0.07) (Okobia *et al.*, 2008). A study of the glutathione S-transferase null variants in the *GSTM1* and *GSTT1* genes in cases and controls from Burkina Faso found no association of the *GSTM1-null* variant with breast cancer while the *GSTT1-null* allele was associated with an increase in breast cancer risk

(OR = 2.42, 95% CI 1.17-5.02, p-value0.01) (Kiendrebeogo *et al.*, 2019). Investigation of an initiation codon variant (Met1Arg, rs2228570) in the Vitamin D receptor (*VDR*) gene in 392 Ethiopian breast cancer cases and controls reported an association of the Arg/Arg (G/G) genotype with an increased risk of breast cancer (OR=1.44, 95% CI 1.01-2.06, p-value0.04) (Ahmed *et al.*, 2019). A lack of association of a functional SNP in the *GPX1* gene (p.Pro198Leu) and the *TP53* functional SNP (p.Pro72Arg) with breast cancer was reported in the Rwandese population (Habyarimana, *et al.*, 2018a, 2018b), but the sample sizes were very small.

Table 0.4. Association Studies of Common Variants in SSA breast cancer

Gene	Methodology	Variant	Country	BC cases	Controls	Odds ratios (95% CI)[a]	Reference
LEPR	PCR-RFLP	p.Gln223Arg	Nigeria	209	209	1.8 (1.0-3.2)	(Okobia *et al.*, 2008)
CYP1B1	PCR-RFLP	p.Leu432Val	Nigeria	250	250	1.59 (1.0-2.52)	(Okobia *et al.*, 2009)
GSTT1	PCR	Deletion (null)	Burkina Faso	80	100	2.42 (1.17-5.02)	(Kiendrebeogo *et al.*, 2019)
VDR	PCR-RFLP	p.Met51Thr	Ethiopia	392	193	1.44 (1.01-2.06)	(Ahmed *et al.*, 2019)
GPX1	Exon sequencing	p.Pro198Leu	Rwanda	41	42	Not reported[b]	(Habyarimana, Bakri, *et al.*, 2018)
TP53	Allele-specific PCR	p.Pro72Arg	Rwanda	40	39	Not reported[b]	(Habyarimana, Attaleb, *et al.*, 2018a)

BC, breast cancer; OR, odds ratio; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; SSA, sub-Saharan African.
[a]ORs and 95% CIs for the strongest results.
[b]No association with BC overall. No ORs were reported.

2.4 Discussion

Our initial PubMed and SCOPUS search yielded 1,029 articles from which 22 were selected for this review. These studies investigated the association of germline variants with breast cancer in resident Sub-Saharan African populations. A large majority of studies focused on the detection and analysis of rare variants in known breast cancer susceptibility genes in cases and, in some studies, also in controls. Our search did not find any GWAS studies for breast cancer in resident African populations, but we did find a small number of association studies of candidate genes and common variants in cases and controls. Although the focus of our review was on sub-Saharan Africa, we discuss the findings below in the context of the extensive literature on breast cancer genetics in African-American populations, which provide important insights relevant to continental Africa.

2.4.1 Contribution of rare variants to breast cancer risks

The studies included in this review either tested for specific variants known to be associated with breast cancer risk in non-African populations or used DNA sequencing of the coding regions of known breast cancer genes to explore the contributions of rare germline variants to the aetiology of breast cancer in their local populations. These germline variants are rare in the general population but have moderate to high penetrance, thus individuals with these variants are at substantially higher risk of developing breast cancer compared to the general population (Kuchenbaecker *et al.*, 2017). Most of the studies focused on the key breast cancer predisposition genes, *BRCA1* and *BRCA2*, which have important roles in DNA repair pathways (Deng and Brodie, 2000; Shahid *et al.*, 2014).

The variants of particular interest listed in Table 2.3 indicate that at least 6 genes, *BRCA1*, *BRCA2, PALB2, ATM, BARD1* and *RAD51C*, collectively make a significant contribution to breast cancer risk in the African populations that were studied. The variants disrupt normal protein function either by early termination of mRNA translation by the introduction of stop-codons or by missense variants which may affect key functional elements of the protein (Tarsounas and

Sung, 2020). However, we note that no functional assays were done in any of the studies reviewed here, and all potential functional effects were predicted *in silico*. There is also the problem of variants of unknown significance (VUS) involving missense variants in known breast cancer susceptibility genes which fail current predictive algorithms for pathogenicity, which will be amplified by the inclusion of diverse and under-studied African genomes (Sirugo, Williams and Tishkoff, 2019). The strategy of, for example, Zheng *et al.*, (2018), in which genes were sequenced in large numbers of cases and controls is helpful in this regard, as it demonstrated a significant excess of many variants in cases vs controls, but the cost of sequencing large numbers of samples may be unaffordable for many African researchers. The contribution of other genes in which potentially pathogenic variants were found is not yet clear as the number of cases with variants in each of these genes was very small. However, for some of these genes, such as *ATM*, there is strong epidemiological and molecular evidence of causality in other populations (Ahmed and Rahman, 2006). Another question of interest is whether specific pathogenic variants are found in more than one African country or ethnic group, such as is the case for European populations. There is currently insufficient evidence to resolve this question in African populations, but it could be addressed by larger-scale studies, such as the Nigerian study of Zheng *et al.*, (2018), being carried out in other African populations.

2.4.2 Genetic association studies of common variants

All the genetic association studies for common variants which we identified involved genotyping a small number of SNPs in one or two candidate genes and small sample sizes. Some marginally significant associations were reported, but such studies would have very low power to detect the typically small effect sizes for individual variants in genetic susceptibility to disease in a complex trait. A further difficulty with candidate gene studies is that they do not provide the means for correction of population structure in the samples, which is an important requirement for a robust genetic association study.

While the above-mentioned studies are limited in scope and of low statistical power, recently there have been several well-powered genetic studies in breast cancer cases and controls of African ancestry, not resident in Africa. For example, a meta-analysis of two GWAS and a

replication study in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor (ER) negative breast cancer (Huo *et al.*, 2016). This study involved genotyping of more than 6,657 breast cancer cases and 7,713 controls on a genome-wide SNP array. Breast cancer risk variants identified in other populations were tested for association in another large study of women of African ancestry (Feng *et al.*, 2017), and identified at least 12 regions which showed evidence of association in this population. These findings suggest that there is at least partial transferability of breast cancer risk loci between different populations and ethnicities. This phenomenon was also observed for other cancers such as prostate and oesophageal cancers where some but not all risk loci were transferable across African, European and Asian populations (Shriner *et al.*, 2009; Chang *et al.*, 2011; Chen *et al.*, 2019). These studies suggest that the GWAS approach may be fruitful for studying the genetics of breast cancer in resident African populations.

2.4.3 Current limitations and future prospects for breast cancer genetic research in Sub-Saharan Africa

Whole genome sequencing of many ethnolinguistic groups in Africa has confirmed the extraordinary degree of genomic diversity across the African continent (Choudhury *et al.*, 2020). A comprehensive analysis of the contribution of germline variants to breast cancer risk in continental Africa will therefore require studies in a wide range of populations and regions. Another important consideration is the study design, which will depend on whether the aim is to identify rare variants of moderate to high penetrance in genes such as *BRCA1* and *BRCA2*, likely associated with familial cancers, or to identify variants with low penetrance which collectively make a large contribution to breast cancer risk in the general population.

In order to identify rare and novel germline variants of intermediate to high penetrance, DNA sequencing, rather than genotyping of known risk variants across the regions and populations of Africa will be required since the limited data available suggests extensive diversity of breast cancer risk variants across the African continent. The study design should also include an equivalent set of well-matched controls which, together with algorithms for predicting functional

effects *in silico*, will help to identify variants that are likely to be pathogenic. The sequencing strategy will depend on available financial and bioinformatic resources. Whole exome sequencing would allow interrogation of the entire coding genome but is still relatively costly and requires bioinformatic skills to call, annotate and interpret variants. Sequencing a panel of genes known to be involved in breast cancer susceptibility, such as the BROCA panel (Shirts *et al.*, 2016), would reduce costs and simplify the bioinformatic analysis, but only investigate genes already implicated in other populations. The probability of detecting pathogenic variants in such genes would be increased by enriching the breast cancer samples for younger patients and those with a family history of breast cancer, as these are groups in which such variants are most likely to be identified. Finally, a substantial sample size will be needed to obtain a robust assessment of the contribution of these rare variants to breast cancer risk in local populations. An exemplar which includes these desirable features is the study by Zheng *et al* (2018) in the Nigerian population (Zheng *et al.*, 2018). Larger sample sizes would also allow tests of whether germline mutations were associated with clinical features such as hormone receptor status, tumour stage and survival. A large study of germline mutations in African American women, for example, showed a two-fold higher incidence of pathogenic mutations in ER-negative breast cancer compared to ER-positive breast cancer (Palmer *et al.*, 2020). Detection of germline mutations in genes such as *BRCA1* and *BRCA2* has important implications for clinical care such as the risk of further cancers in the patient, genetic counselling to other family members and potential access to new classes of drugs such as poly (ADP-ribose) polymerase inhibitors which can improve survival and reduce side effects in mutation carriers (Fong *et al.*, 2009).

Regarding the detection of common risk variants for breast cancer, small sample sizes such as those reported to date in Sub-Saharan African populations do not have the power to detect associations of common variants with small effects. A shortcoming of most candidate gene studies is that a limited number of variants are genotyped which do not capture the full genetic diversity of the gene, and genome (Duncan *et al.*, 2019b). Also, genotyping a small number of variants does not allow correction for population structure, which is particularly important in populations with a high degree of genetic diversity and substructure. This requires either a genome-wide study or at least a genotyping array containing thousands of common variants to

detect and control for population structure. The route to significant discoveries of common breast cancer risk variants in African populations is to carry out GWAS using SNP arrays which are designed to capture a substantial component of African genetic diversity and genotyping thousands of cases and controls to achieve the statistical power required to detect variants with modest effect sizes. We are currently carrying out a GWAS using the H3Africa array on over 2700 breast cancer cases and 1100 controls from the Johannesburg Cancer Study (Pacella-Norman *et al.*, 2002) in the black South African population as part of the ERICA-SA study (http://www.mrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa) which aims to identify evolving risk factors for cancer in African populations.

A major difficulty regarding the desirable study designs outlined above is the resource-intensive nature of molecular cancer research such as next-generation sequencing and genome-wide genotyping in terms of funding and appropriately skilled personnel, both of which are relatively scarce in Africa. Significant progress in improving our understanding of the genetics of breast cancer in Africa will require a substantial investment from governments or other sources to generate the grant income and training programs needed to achieve this goal.

In conclusion, while some studies on the genetics of breast cancer have been carried out in populations of African origin, few comprehensive large-scale studies have included populations resident in Sub-Saharan Africa. This scoping review has identified that more extensive studies of the contribution of both rare and common variants to breast cancer risk are needed in a broader range of African countries. The knowledge gap creates an opportunity for African researchers to gain new insights into the molecular epidemiology of breast cancer on the African continent.

2.5 Recent genetic studies

This section covers papers published on this topic from 02 April 2020 to 22 July 2022. The same search criteria were used as described in section 2.2. Ten additional relevant genetic studies were published on breast cancer in Sub-Saharan Africa during this period, seven of which are rare variant studies and three are common variant studies.

2.5.1 Rare variant studies

Biancolella *et al.*, (2021) used NGS to sequence the coding regions of *BRCA1* and *BRCA2* and identified six pathogenic mutations in a cohort of 51 women from Burkina Faso with early onset breast cancer; two of the variants in *BRCA1* occurred in 2 patients (Table 2.5), whereas the other 4 variants occurred in single patients. The second study in Burkina Faso (Ouedraogo *et al.*, 2022) used Sanger sequencing to identify mutations in mutation hotspots in *BRCA1* and *BRCA2* in 133 breast cancer patients who had previously had breast cancer. These patients were not selected based on the early onset of breast cancer. They found three pathogenic variants, each of which occurred in a single patient. A study in Mauritania sequenced *BRCA1* and *BRCA2* in 137 mainly female patients with relatively early onset and/or familial breast cancer and identified four pathogenic variants in *BRCA1* and *BRCA2*, all in two or more patients. These included the p.Thr276Alafs*14 and c.4986+6T>C variants, which are well-established African *BRCA1* founder mutations (Brahim *et al.*, 2022). Eygelaar *et al.*, (2022) included in their study 165 female breast cancer patients from South Africa who self-identified as women of African ancestry and were unselected for family history. Using the Illumina TruSight Cancer panel, they sequenced 94 genes. They identified 11 variants that were classified as pathogenic or likely pathogenic, but none were found in more than one participant. A Rwandan study (Uyisenga *et al.*, 2020) sequenced 26 genes in 40 early-onset breast cancer patients. They identified five pathogenic variants in five different patients, including three in *BRCA2* and one each in *BRCA1* and *TP53*. A large diagnostic study in South African populations (Van der Merwe *et al.*, 2022) screened 2974 breast cancer patients for mutations in the *BRCA1* and *BRCA2* genes and detected 481 clinically actionable variants. The variant list was not stratified by ethnicity, but the study did detect two previously reported African *BRCA2* pathogenic founder mutations (c.5771_5774del and

c.582G>A) in multiple patients of African ancestry. Finally, a study in Burkina Faso (Kiendrebeogo *et al.*, 2022) which included 50 breast cancer cases and 50 controls reported the detection of mutations in exon 11 of *BRCA1* in 4 patients, but the mutations were not characterised.

Table 0.5. Predicted Pathogenic Rare Variants Detected in two or more African breast cancer patients in recent studies

Gene	Variant	Variant effect	Mutation type	No. of cases	Country	Reference
BRCA1	rs398122680	p.Ser1363Ter	Nonsense	2	Burkina Faso	(Biancolella *et al.*, 2021)
	rs80358086	c.4986+6T>C	Splice site → truncated protein	2		
	rs80358086	c.4986+6T>C		7	Mauritania	(Brahim *et al.*, 2022)
	rs387906563	p.Thr276Alafs*14	Frameshift	13		
BRCA2	rs80359572	p.Tyr2094LeufsTer	Frameshift	2		
	rs397507906	p.Thr2412SerfsTer	Frameshift	4		

2.5.2 Common variant studies

Three common variant studies from sub-Saharan Africa were published in this period, using relatively small sample sizes in three different countries. These results are summarised in Table 2.6. Adolf *et al.*, (2022) investigated 10 variants in DNA repair genes in 263 breast cancer patients and 250 controls in Tanzania. They found that four variants had significant allelic differences between cases and controls (p-value<0.005): rs25487 in *XRCC1*, rs861539 in *XRCC3*, rs17655 in *XPG* and rs4987188 in *MSH2*. In Cameroon, a PCR-RFLP assay was used to explore the association between rs17506395 in *TP63* and breast cancer in 335 cases and 224 controls (Tiofack *et al.*, 2020) and was found to be associated with breast cancer in women under 40 years of age (p-value0.03, OR=0.5. 95%CI=0.26-0.94). A Malian study of 60 cases and controls found that a 16bp duplication in *TP53* intron 3 (PIN3) was associated with breast cancer (p-value0.03, OR=1.87, 95%CI=1.05-3.33) (Diakite *et al.*, 2020). This result was further supported by a meta-analysis which included 19 studies (p-value0.01, OR=1.11, 95%CI=1.02-1.19).

Table 0.6. New Association Studies of Common Variants in SSA breast cancer

Gene	Methodology	Variant	Country	Cases	Controls	Odds ratios (95% CI)	Reference
XRCC1	QRT-PCR	rs25487	Tanzania	263	250	2.18(1.51-3.16)	(Adolf *et al.*, 2022)
XRCC3	QRT-PCR	rs861539	Tanzania	263	250	2.91(2.06-4.11)	
XPG	QRT-PCR	rs17655	Tanzania	263	250	2.31(1.80 - 2.97)	
HMSH2	QRT-PCR	rs4987188	Tanzania	263	250	2.67(1.64-4.35)	
TP63	RFLP-PCR	rs15706395	Cameroon	335	224	0.5(0.26-0.94)	(Tiofack *et al.*, 2020)
TP53	PCR	PIN3 Ins16bp	Mali	60	60	1.87 (1.05-3.33)	(Diakite *et al.*, 2020)

95%CI 95% confidence interval
PCR polymerase chain reaction
QRT-PCR quantitative real-time PCR
RFLP-PCR restriction fragment length polymorphism PCR

2.5.3 Summary of the new studies

The new studies described above have the same limitations as those reviewed in sections 2.1 – 2.4. In the context of rare variants of high penetrance, there is a need for screening a broader range of genes in patients with defined clinical characteristics to assist in the prioritisation of future genetic screening. Future studies of common variants will require broader, preferably genome-wide studies, in much larger sample sizes to detect the association of variants with modest effect sizes.

Chapter 3: Methods and Materials

3.1. Overview

The aim of this study was to search for the association of genetic variants with breast cancer in a South African black population. To achieve this aim, a genome-wide association (GWA) study was performed in women with breast cancer and ethnically matched population controls from Soweto, South Africa. Following the identification of variants of interest, the potential functional significance of variants that achieved genome-wide significance (p-value<$5x10^{-08}$) was assessed for association with the breast cancer phenotype, as well as variants with suggestive significance (p-value<$5x10^{-06}$). During post-GWA analysis, I investigated whether the variants of interest in this study were previously associated with breast cancer. Then, it was investigated whether genetic risk variants identified in other populations (both African and non-African) could be replicated in this African population. A meta-analysis of this study was carried out with summary statistics from an African American breast cancer GWAS study. Finally, an existing polygenic risk score developed for breast cancer in European populations was tested in this African sample.

3.2. Study design

This study is a case-control GWA study. The breast cancer cases were enrolled by the Johannesburg Cancer Study (JCS) (Urban *et al.*, 2012; Chen *et al.*, 2020) and the ethnically matched non-cancer population controls were enrolled by the Africa Wits INDEPTH partnership for genomic studies (AWI-Gen) study (Ali *et al.*, 2018; Sengupta *et al.*, 2021). This study forms part of the larger Evolving Risk Factors for Cancer in African populations (ERICA-SA) study (https://www.samrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa) and it was funded by the South African Medical Research Council (SAMRC), Medical Research Council UK (UKMRC) and the Newton Fund to evaluate the risk factors associated with common African cancers (Ethics, Medical: M160807).

3.3. **Study sample**

The breast cancer cases were enrolled by the JCS. The JCS began with the aim to understand risk factors for cancer and cancer susceptibility in African populations. The enrolment was started in 1995 and concluded in 2016. Over the twenty-one-year period, the JCS enrolled 24,971 treatment-naïve individuals of African ancestry who were diagnosed with cancer (Chen *et al.*, 2020) and recruited from the Charlotte Maxeke Johannesburg Academic Hospital, the Radiation Oncology ward of Hillbrow Hospital (Johannesburg) and the Chris Hani Baragwanath Academic Hospital in Soweto (South-West Johannesburg) (Figure 3.1). The participants were surveyed with a detailed epidemiological questionnaire (Appendices), and they also donated bio-samples for future testing after providing informed consent. Hormone receptor status was not available at the time of this study. Of the 24,971 enrolled participants, 19,351 donated an EDTA blood sample for genomic DNA testing and 18,972 donated serum for serological testing. The JCS is a large biorepository with a wealth of information to enable future cancer research. A total of 2,573 female individuals (median age 46 years, 95% confidence interval (CI)(45-46)) of African ancestry with histologically confirmed breast cancer were present in the JCS and analysed in this study. Approximately 90% of cases were diagnosed with ductal carcinoma. Ethical approval for the JCS was obtained from the University of Witwatersrand Human Research Ethics Committee (HREC - Medical) (M1606103).

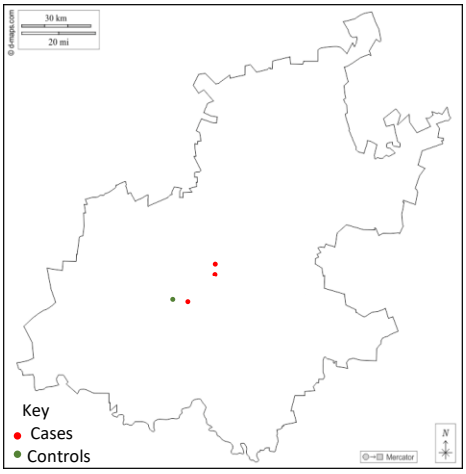


Figure 0.1: Map of Gauteng province showing case and control recruitment sites in Johannesburg, Gauteng
Cases from the JCS were recruited from the Chris Hani Baragwanath Hospital, Hillbrow Hospital and the Charlotte Maxeke Johannesburg Academic Hospital. Controls from the AWI-Gen study were recruited from Soweto.

Ethnically matched, non-cancer, population controls were selected from the AWI-Gen study (Ali *et al.*, 2018; Sengupta *et al.*, 2021) (Ethical approval: HREC M121029, M180535). The AWI-Gen study aims to understand risk factors, including genetic risk factors, for cardiovascular and metabolic disease in sub-Saharan Africa. The study recruited 11,723 participants from East, West and South Africa. The study sites were located in Soweto, South Africa; Agincourt, South Africa; Dikgale, South Africa; Nairobi, Kenya; Navrongo, Ghana and Nanoro, Burkina Faso. Extensive epidemiological data was collected through questionnaires and blood and urine samples were collected for laboratory assays. Initially, 2,128 controls were included in the initial analysis. However, 1,048 male controls from AWI-Gen and 269 female controls from the JCS were later removed to for reasons discussed in section 4.2. A total of 744 female (median age 49 years, 95%CI(48.21-49)), African, non-cancer individuals from Soweto, Johannesburg, were included in the final analysis of this study, as they were the most appropriate control group from the viewpoint of genetic similarity.

All participants in this study consented to the parent studies for the use of their DNA samples for genetic studies. The current study to understand the genetic aetiology of African breast cancer was approved by the University of the Witwatersrand's HREC (Medical), protocol number M180997.

3.4. **Sampling and DNA isolation**

Genomic DNA (gDNA) was isolated from peripheral whole blood samples from all study participants. gDNA was isolated using the salting out method (Miller *et al.*, 1988). Participants in the JCS donated 4 mL of whole blood. The blood samples were collected in vacutainer tubes and stored at -30°C until the blood samples were retrieved and processed for DNA extraction. Blood samples from JCS were stored for up to 20 years at -30°C before DNA extraction. The prolonged storage of the blood samples did not negatively affect the quality and quantity of the DNA. The mean 260/280 ratio was 1.79, suggesting that the DNA was of good quality (Chen *et al.*, 2018).

The quality and quantity of gDNA were assessed using the Nanodrop™. DNA concentration was normalized to 100ng/µL. The 260/280 ratio and the 260/230 ratio were calculated and the gDNA samples were stored at -80°C in the Sydney Brenner Institute for Molecular Bioscience (SBIMB) biobank until use (Soo *et al.*, 2017). The SBIMB biobank houses more than 50,000 bio-specimens for multidisciplinary research projects in South Africa.

3.5. **Genotyping and data quality control**

The genotyping in this study was performed at two high-throughput genomic centres.

Prior to shipping, the stock gDNA samples of the breast cancer cases were thawed and 50µl of the 100ng/µL stock DNA samples were aliquoted into 96 well semi-skirt PCR plates (Thermo-fisher) and the plates were heat sealed. The sealed PCR plates were then shipped on dry-ice to the IoPPN Genomics & Biomarker Core Facility, Department of Social Genetics & Developmental Psychiatry, King's College London for genotyping (https://www.kcl.ac.uk/research/the-ioppn-genomics-biomarker-core-facility). The IoPPN Genomics & Biomarker Core Facility is a CSPro-certified provider of Illumina's platforms and assays for genotyping and gene expression.

The control samples were genotyped by the AWI-Gen study using the Illumina® FastTrack™ Microarray services (Illumina, San Diego, USA). All samples at both centres were genotyped on the 2.3 million H3Africa array version 1.0. (https://www.h3abionet.org/h3africa-chip) (Mulder *et al.*, 2018; Choudhury *et al.*, 2022). This genotyping array was designed by H3Africa researchers after it was noted that most of the publicly available SNP genotyping arrays were based on genetic data from non-African populations. The H3Africa array considers a large amount of variation seen in African populations, as well as the smaller haplotype blocks (Mulder *et al.*, 2018). The array includes ~1.7 million SNPs that were already in the Illumina bead pools designed for other arrays, some of which were also common variants in African datasets. Approximately 700,000 additional SNPs were carefully selected to include common African variants, and SNPs of interest for some diseases (Mulder *et al.*, 2018). This array was chosen because the participants were all African.

3.5.1. Data analysis workflow

The data analysis workflow for this study is summarized in Figure 3.2 below:

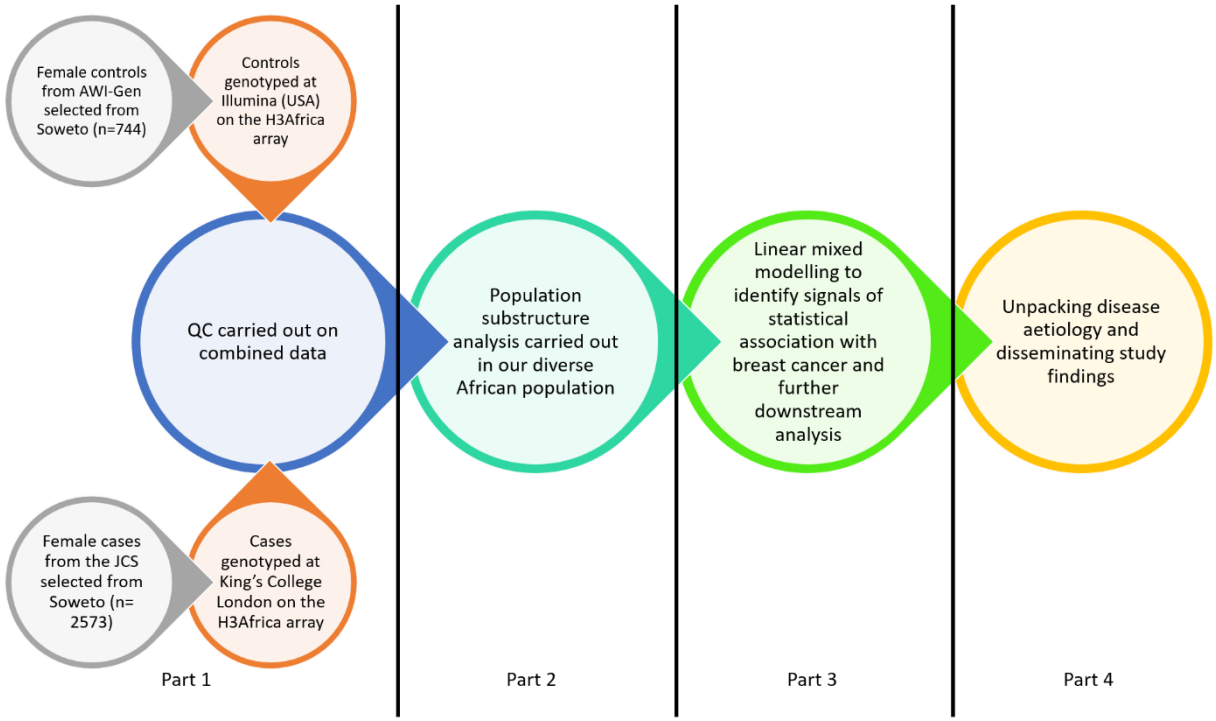


Figure 0.2: The GWAS workflow summary
The workflow consists of four main parts: 1. Genotype data collection, quality control and imputation. 2. Correction for population substructure. 3. Linear0mixed modelling of the genetic variants with covariates to identify associations with breast cancer. 4. Unpacking the findings to understand the underlying mechanisms for disease risk.

The GWAS workflow has four parts: 1. Genotype data collection, and quality control as well as genotype imputation. 2. Correction for complex population substructure that may be present within this African study population. 3. Linear-mixed modelling of the genetic variants, with covariates, to identify genetic variants that may be associated with the disease phenotype of interest. 4. Unpacking the genetic findings to understand the underlying mechanisms for disease risk, replication of previous findings and the computation of polygenic risk scores for breast cancer in an African population.

3.5.2. Genotype data and quality control

The raw genotype data was presented in the form of high-density images. There are two colour channels used for the Illumina microarray, red and green. The probes for each genetic marker are embedded into a specific location on the array. The intensity of the fluorescent probes (green or red) indicates the presence or absence of a specific genetic variant (Smith *et al.*, 2013). The raw data, in the *.iDat* format, was converted to a binary PLINK genotype file (*.bed*) using the Illumina Array Analysis Platform (IAAP) Genotyping Orchestrated Workflow (https://emea.support.illumina.com/downloads/iaap-genotyping-orchestrated-workflow.html). The IAAP-Genotyping Orchestrated Workflow takes the raw intensity data, a bead pool manifest, and a cluster file as input and converts the raw data into binary genotype calls. The bead pool manifest was developed during the production of the H3Africa microarray and was provided by Illumina for this study. The cluster file is a file that contains the pre-defined SNP clustering locations for all 2.3 million markers on the H3Africa microarray. The cluster files in the study were developed by the Illumina FastTrack Microarray service, based on the genotyping data from the AWI-Gen study. The IAAP-Genotyping Orchestrated Workflow implemented Illumina's GenCall algorithm to perform the genotype calling. The genotype calls were stored in the genotype call (GTC) format and then converted to PLINK binary genotype data file format. The IAAP workflow generated one output file per sample, thereafter all files were consolidated to generate one PLINK binary data file for downstream quality control workflows. PLINK 1.9 was used (Chang *et al.*, 2015) for data management.

The H3ABioNet's H3Agwas Pipeline version 3 (Brandenburg *et al.*, 2022) was used for the data quality control and association modelling. An overview of the H3Agwas pipeline is shown in Figure 3.3:

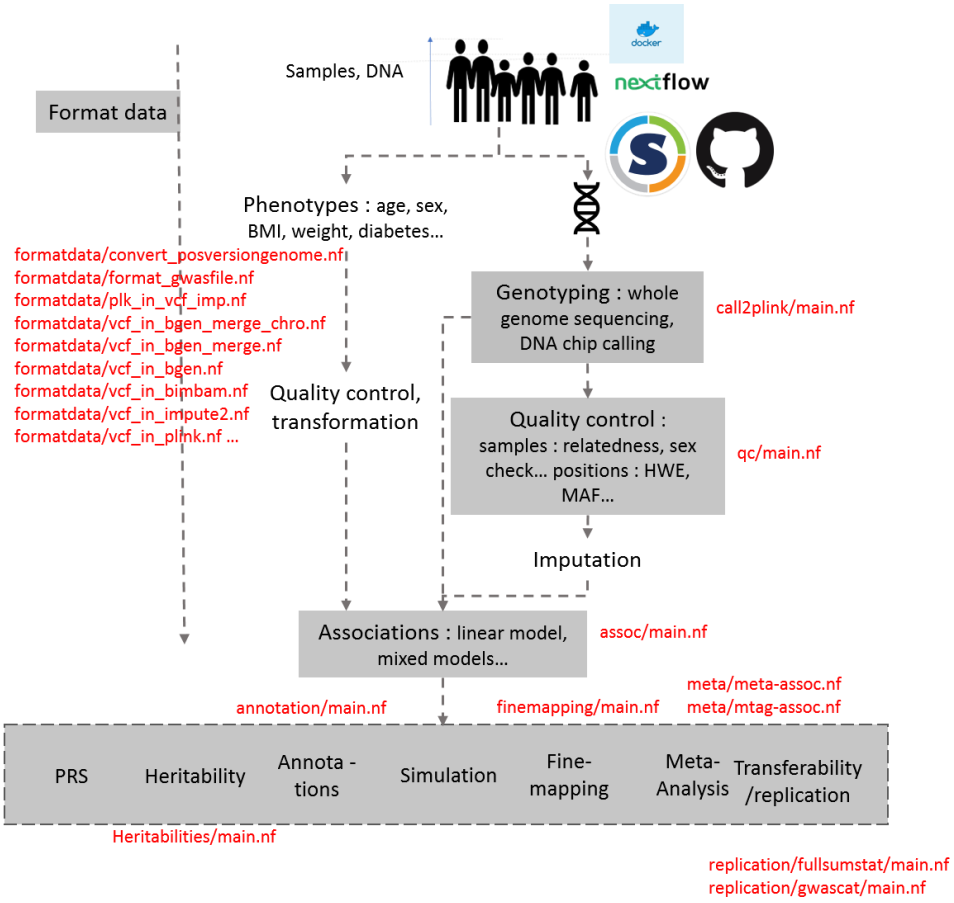


Figure 0.3: Schema of the H3Agwas pipeline (Brandenburg *et al.*, 2022), (https://github.com/h3abionet/h3agwas)

Genotype data quality control was an iterative process. The H3Agwas pipeline performed QC in several steps. First, the pipeline flags and removes duplicate markers. Then, the pipeline removes markers based on minor-allele frequency thresholds, marker missingness, and markers that deviated from Hardy-Weinberg equilibrium. The pipeline also removes individuals based on the percentage of missing-data within an individual, sample heterozygosity, identity-by-descent (to find duplicated samples or samples from related individuals), and genotype-phenotype sex mismatched individuals. Lastly, the QC workflow removed markers with differential missingness between cases and controls. This may indicate a genotyping bias and may be misleading and identified as a disease-associated marker. At the end of the QC pipeline, a PDF report is generated that highlights the number of markers and individuals excluded based on the pre-defined thresholds. Figure 3.4 summarizes the overview of the H3Agwas QC workflow.

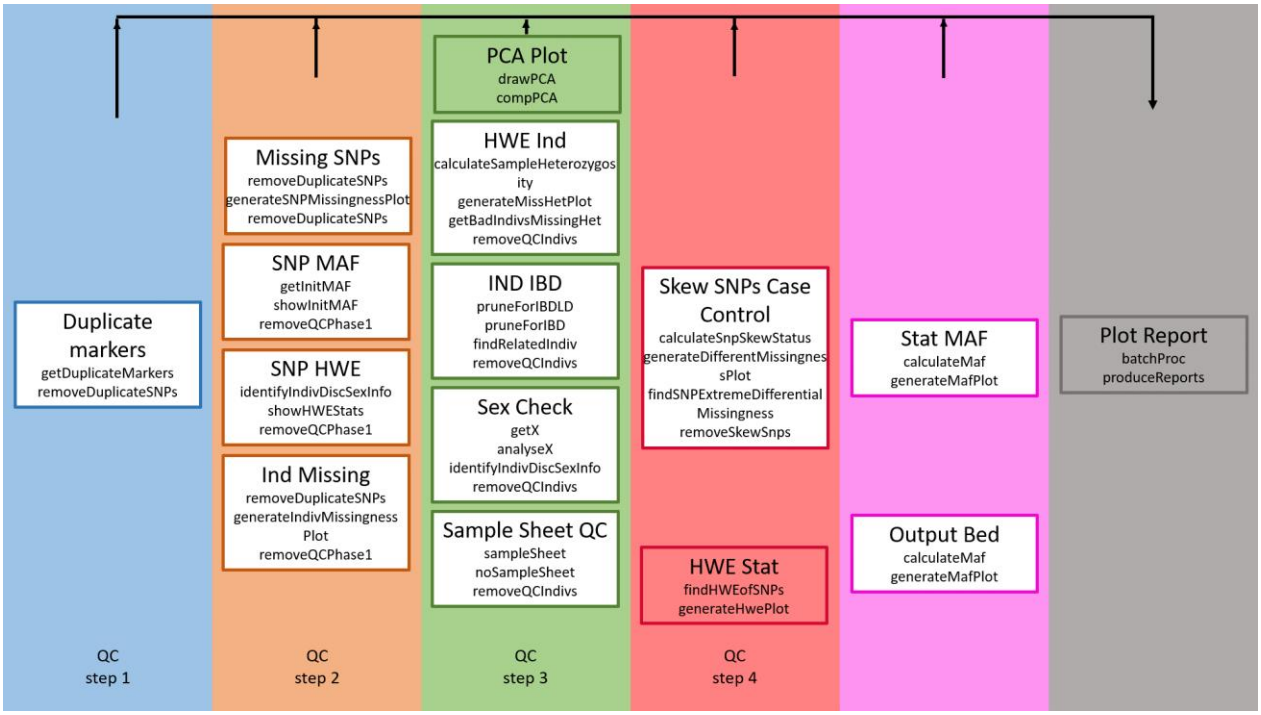


Figure 0.4: Overview of the H3Agwas QC workflow (adapted with permission from https://github.com/h3abionet/h3agwas.

For this analysis, only autosomal SNPs/markers were retained. The SNP QC thresholds used for this study for marker retention were:

- Missingness <= 0.01 (remove SNPs with low quality genotypes)
- Minor allele frequency (MAF) >= 0.01
- Hardy-Weinberg equilibrium (HWE) p-value >= 0.0005 (remove SNPs out of HWE due to migration/non-random mating)

The thresholds used for sample retention were:

- Missingness <= 0.02 (avoid poor DNA quality samples with low call rates)
- piHat <= 0.18 (unrelated individuals, i.e. more distant than third-degree relatives)
- Genotype-phenotype sex-matched (ensure self-reported and genotype sex match)
- Heterozygosity between the range of 0.15 and 0.343 (avoid sample contamination or inbreeding)

3.5.3. Principal component analysis and admixture analysis

After genotype and sample QC, the cleaned genotype dataset was used as the input dataset for the principal component and admixture analysis as part of the strategy for population substructure analysis. For ease of reference, the genotyped dataset will be referred to as chip data.

As a means to control for population substructure, batch effects or other technical study artefacts or biases, random subsampling and iterative testing of study samples was performed. This iterative approach consisted of comparing female individuals vs male individuals, as well as random subsets of samples across each genotyping batch and within batches. This testing strategy involved several iterations of random subsampling of 300 individuals as the "controls" and 300 individuals as the "cases", followed by principal component analysis and linear mixed models. Genomic inflations, potentially as a result of biases introduced by batch effects or population substructure, would be detectable by an increase in the average -log10(p-value).

A pairwise linkage disequilibrium of the chip data was performed in PLINK 1.9 (Chang *et al.*, 2015). Linkage disequilibrium (LD) was computed for markers within a 100KB sliding window and in a step size of 20 variants per shift. An r^2 threshold of 0.2 was used and one marker, from a pair, with r^2 greater than 0.2 was excluded. This allowed the generation of an LD-independent genotype dataset.

Principal component analysis (PCA) was performed by Eigen-decompositions of the genetic data. The CEU, CHB and YRI populations from the 1000Genomes project (1KGP) phase 3 (The 1000 Genomes Project Consortium, 2015) were used as reference populations. Ten principal components were computed and plotted. Generalized linear models (GLM), in R (Development Core Team, 2008), were used to determine which principal components were statistically associated with the phenotype. PCs 1 to 5 were selected and included as covariates in the final model. Eigen decomposition was performed using PLINK v.1.9 and visualized in R v4.0.3.

Admixture analysis was performed using ADMIXTURE 1.3.0 (Alexander *et al.*, 2009). Six folds of cross-validation were performed to determine the optimal *K* value. The analysis was done using CEU (n=99, European) and CHB (n=103, Han Chinese, East Asian) individuals from the 1KGP as the reference populations. The Admixture plots were visualized in Stata version 17 (StataCorp LLC, College Station, USA). Individuals with >5% CEU or CHB genetic contribution were excluded to ensure that the analysis was not confounded by admixed individuals.

3.5.4. Estimation of Heritability

SNP-based heritability (h^2_g) for breast cancer was estimated using the genomic restricted maximum likelihood (GREML) solver. The GREML estimated the h^2_g contributed by kinship matrices. The LDAK model was followed. This model proposed that the SNP's contribution to heritability is dependent on the MAF and the local LD (Speed *et al.*, 2020). The h^2_g estimation was performed using LDAK version 5.2 (Speed *et al.*, 2020). First, LDAK pruned and reduced the number of predictors by using a correlation squared threshold of 0.98. The predictors were then separated into sections. Thereafter, each section was weighted using a quadratic solver. The weights were then used to construct the genetic relationship matrix (GRM). GREML was used to estimate the h^2_g. An equal number of cases and controls were randomly subsampled from the total study population (744 cases against 744 controls) to prevent any over or under estimation of the heritability estimate due to unbalanced case-control ratio. The h^2_g was estimated on the liability scale, using the Globocan's 2020 breast cancer incidence rate in South Africa (52.6/100,000) as a proxy for disease prevalence (Sung *et al.*, 2021).

3.6. **Genetic imputation**

In microarray technology, only a select number of positions in the genome are targeted and genotyped. The genotyped markers are selected based on several factors (e.g. LD inference, functional importance, chromosomal coverage/backbone) to provide adequate coverage throughout the genome. It is often not feasible to genotype large numbers of genetic markers

due to the physical limitation of the array-marker layout and density limit of probes. Having too many probes on one array may increase the chances of probe failure.

In order to have more genetic data from microarray, genetic imputation was used. Genetic imputation is the process of inferring the genotypes of ungenotyped markers using various information provided by the genotyped markers, such as LD and haplotypes (Li *et al.*, 2009; Das *et al.*, 2018; Naj, 2019). Even unrelated individuals, provided they are of similar ancestry, share common genomic regions that they have inherited from a common ancestor. Therefore, imputation relies on the fact that the reference sample used is a recent ancestor of the sample that is being imputed. This ensures that the haplotype blocks of the reference sample will be similar to the imputed sample, resulting in a more accurate imputed dataset. The imputed dataset includes many more variants which can be included in studies, increasing the power of GWAS (Li *et al.*, 2009).

The Sanger Imputation Service (McCarthy *et al.*, 2016) was used to impute the genotypes. The African Genome Resource (AGR) panel was used as a reference which contains 4,956 samples. The reference panel includes all samples from the 1000 Genomes Project (1KGP) Phase 3 as well as samples from Uganda, Ethiopia, Egypt, Namibia and South Africa. Pre-phasing was performed with EAGLE2 (Loh *et al.*, 2016). Thereafter, post-imputation was carried out. The parameters for post-imputation QC were as follows. SNPs were retained if:

- SNP missingness<0.05
- MAF>0.01
- Impute2 Score>=0.8
- HWE p-value>=0.00001

In order to achieve the most robust imputation possible with this dataset, the H3A chip data was imputed from breast cancer case samples, all the samples from the AWI-Gen study, as well as two other sets of cancer cases from the JCS which had also been genotyped on the H3A array (oesophageal cancer and cervical cancer) as part of the ERICA-SA study.

3.7. **GWAS LMM modelling**

The GWAS was performed using linear-mixed modelling regression deployed in GEMMA (v.0.98.1) (Zhou and Stephens, 2012). The binary phenotype was residualised using a generalised linear model (GLM) in R, with 5 PCs as covariates. The residual values were used as a linear phenotype in the linear mixed model (LMM). The residualisation takes into account what the association is between the genotype and the phenotype and calculates a numerical value that best represents this association (Yang *et al.*, 2014). The residual values were then used in the GEMMA LMM which statistically analyses the relationship between the genotype and the phenotype, which was implemented through the H3Agwas pipeline. The LMM was further parallelised to improve computation speed by splitting the dataset into 22 chromosomes and regressing all 22 datasets in parallel. The summary statistics were processed in Stata 17 (StataCorp, College Station, USA) and R. The beta values were converted to odd ratios using the formula described by Cook *et al.* (2016). Even though GEMMA by default excludes all SNPs with a combined MAF of <0.01, all SNPs with MAF<0.01 were further excluded by case and control status respectively.

Given the very large number of SNPs being tested for association in GWAS, the probability of false positive association is corrected by using methods including the Benjamini-Hochberg or Bonferroni methods for multiple tests (Dunn, 1961; Benjamini and Hochberg, 1995). The generally accepted p-value in genome-wide association studies is $5x10^{-08}$, indicating that any SNP that falls below this threshold will be considered significantly associated with the phenotype in question.

Genotype cluster files for genotyped SNPs in the top hits were extracted from the GTC files and visualized using gtc2vcf (https://github.com/freeseek/gtc2vcf#convert-illumina-idat-files-to-gtc-files).

3.8. **Permutation testing for top hits**

Permutation testing was used to correct for the experiment error rate and to help ensure that the results from statistical models were robust (Che *et al.*, 2014). A permutation test is a non-parametric, randomising test in which the samples of a t-test are shuffled *n* number of times, and the test statistic is noted for each set (Phipson *et al.*, 2010; Che *et al.*, 2014). A p-value is determined from the distribution of the t-statistics. The dataset can be shuffled several times, and if the null hypothesis is true, the outcome from the shuffled data will resemble the original outcome, in this case, the original GWAS outcome (Che *et al.*, 2014). Adaptive permutation testing was done using PLINK v1.9's permutation algorithm deployed in Field-Programmable Gate Arrays (FPGA) designed to run on a cloud-based AWS EC2 FPGA instance (Swiel *et al.*, 2022). Seven hundred million adaptive permutations were performed on the imputed GWAS dataset using the linearised phenotype derived by residualising the binary phenotype and PCs 1-5 under an additive model. Seven hundred million adaptive permutations were the limit of the AWS EC2 FPGA deployment and sufficient for application.

3.9. **Fine mapping and functional assessment of associated variants**

Fine mapping was carried out using LocusZoom v1.4 (Pruim *et al.*, 2010), for all top hits with p-value<$5x10^{-08}$, with a 400kb flanking nucleotide window, using precomputed LD information from this study population. The FUMA (Watanabe *et al.*, 2017) online application was used to annotate and interpret associated GWAS variants with p-value<$1x10^{-05}$ as well as annotated co-localized eQTLs in the breast tissues of interest from GTEx version 8 (Lonsdale *et al.*, 2013). The SNPs (p-value<$1x10^{-05}$) were also submitted to SNP-nexus (snp-nexus.org) for additional functional annotation. All SNPs were analysed in FINEMAP v1.4 (Benner *et al.*, 2016) using the H3AGWAS pipeline. The programme uses the shotgun stochastic search (SSS) algorithm to quickly evaluate multiple configurations to identify those with the highest posterior probability.

3.10. **Replication**

SNPs from this study (p-value<$5x10^{-06}$) were test for replication in the Breast Cancer Association Consortium (BCAC) iCOGS (Michailidou *et al.*, 2013) data for European and African American datasets separately. The BCAC is an international consortium that aims to identify genes that are associated with the risk of breast cancer development (https://bcac.ccge.medschl.cam.ac.uk/). There are over 100 groups included in this consortium from all over the world, however, none of the studies includes resident African populations. The summary statistics are freely available on the BCAC website. In addition, top hits from the European (p-value<$5x10^{-08}$) and African American (p-value<$1x10^{-05}$) datasets were investigated in this African breast cancer GWAS.

3.11. **Meta-analysis**

A meta-analysis was performed using this African dataset (n=3,317), together with the BCAC iCOGS African American (n=2,048) summary statistics using METAL (v 2011-03-25) (Willer *et al.*, 2010). An overlap of 12,155,889 SNPs between both datasets was included in the meta-analysis. The sample size weighted model was used for the meta-analysis.

3.12. **Polygenic risk scores**

A polygenic risk score were generated using PRSice-2 (Choi and O'Reilly, 2019) in this dataset using the 313 SNP model by Mavaddat *et al.* (2019). However, from the 313-SNP model from the study, only 201 SNPs were imputed in the African dataset. Therefore, only 201 SNPs were used to generate the polygenic risk score in the African breast cancer dataset. A polygenic risk score was also generated using the 3,820 SNP model from the same study, but only 2,157 SNPs were available for use from the imputed dataset in the risk model.

Two further analyses were done using the African breast cancer dataset. A polygenic risk score was generated from the European and African American BCAC datasets and applied to the African breast cancer dataset. All the SNPs from the African American BCAC summary statistics that overlapped with the African breast cancer dataset (12,187,877 SNPs) were included in the input, but only 31,479 SNPs were used in generating a risk score to apply to the African data. This is because PRSice-2 calculated that using 31,479 SNPs resulted in the most informative risk score model. A polygenic risk score was also generated from the European BCAC summary statistics; 3,974,083 SNPs overlapped and were included in the input data, but only 10,043 SNP were used in the risk score analysis as it was the most informative model. All models included PCs 1-5 as covariates.

Chapter 4: Results

4.1. Study Samples and Genotyping

Histologically confirmed female breast cancer cases were selected from the JCS and ethnically matched population controls from the AWI-Gen study for genotyping and association analysis. All case participants were enrolled from the Soweto area, Johannesburg, South Africa. Controls were sampled from the JCS (n=269), as well as three sites from the AWI-Gen study: Soweto (n=1,723), Dikgale (n=1,034) and Agincourt (n=1,745). All study participants self-identified as of African ancestry. Admixture analysis on the cases and three AWI-Gen sites found substantial differences between the Agincourt and Dikgale groups and the JCS/Soweto cases and controls and were therefore, excluded from the analysis (Figure 4.1). K=4 was used because it had the lowest cross validation error.

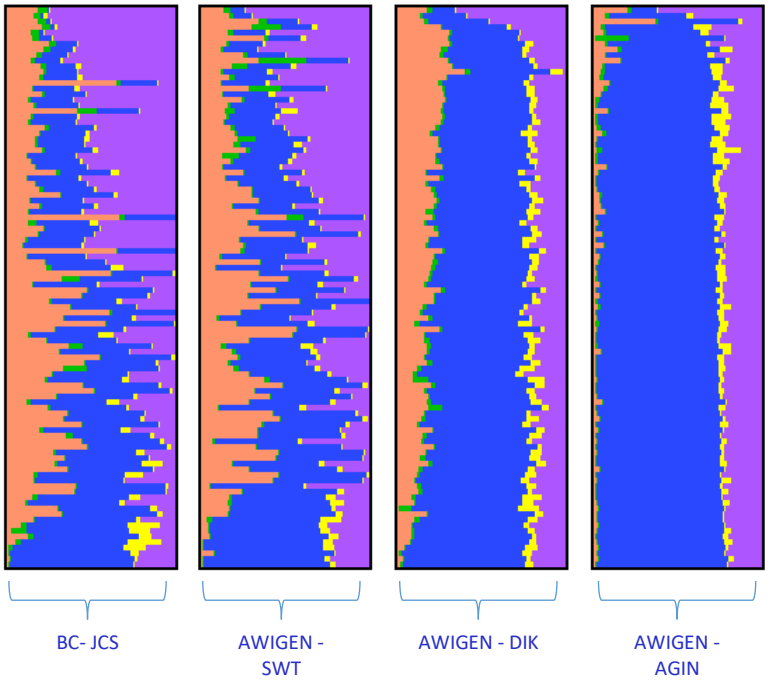


Figure 0.1: Admixture plot for cases and all control sites.
Admixture plot assesses genomic contribution of all breast cancer cases from Soweto and controls from three AWI-Gen sites (Soweto, Dikgale and Agincourt). This was done on a random subset of 100 individuals from each site. K=4. BC=breast cancer, JCS: Johannesburg Cancer Study, SWT: Soweto, DIK: Dikgale, AGIN: Agincourt.

After the removal of the Agincourt and Dikgale controls, 2,711 breast cancer cases, 1,080 female controls and 1,048 male controls all from Soweto were included in further analysis. They were genotyped on the H3A GWAS array and quality control analysis followed as described in chapter 3.

4.2. Adjustment for population structure

Three steps were taken to account for the population substructure. In combination, the three steps allowed us to identify population substructure inflation within the dataset. The first step was to perform iterative testing of male and female population controls within each genotyping batch and then across genotyping batches. This approach was performed to identify if the male and female controls were genetically homogeneous within the study sites and to test for possible genomic inflation as a result of genotyping batch effects. The second step was to perform admixture analysis using reference populations from the 1000 Genomes study. This was done to identify study participants with substantial non-African genetic contributions. The third step was to perform principal component (PC) analysis using Eigen decomposition. This was done to assess how well the cases matched with the control population and to correct for population stratification.

4.2.1. Initial GWAS analysis

Admixture analysis, PC analysis and imputation were performed on the dataset that included Soweto cases and male and female controls from the JCS and AWI-Gen. The initial GWAS was carried out on this dataset. The results showed some SNPs associated with breast cancer at a genome-wide significance level and suggestive signals with little or no support from surrounding SNPs (Figure 4.2). This raised concerns that some of these signals were likely artefacts and prompted us to further examine the dataset to identify the source of a potential false signals.

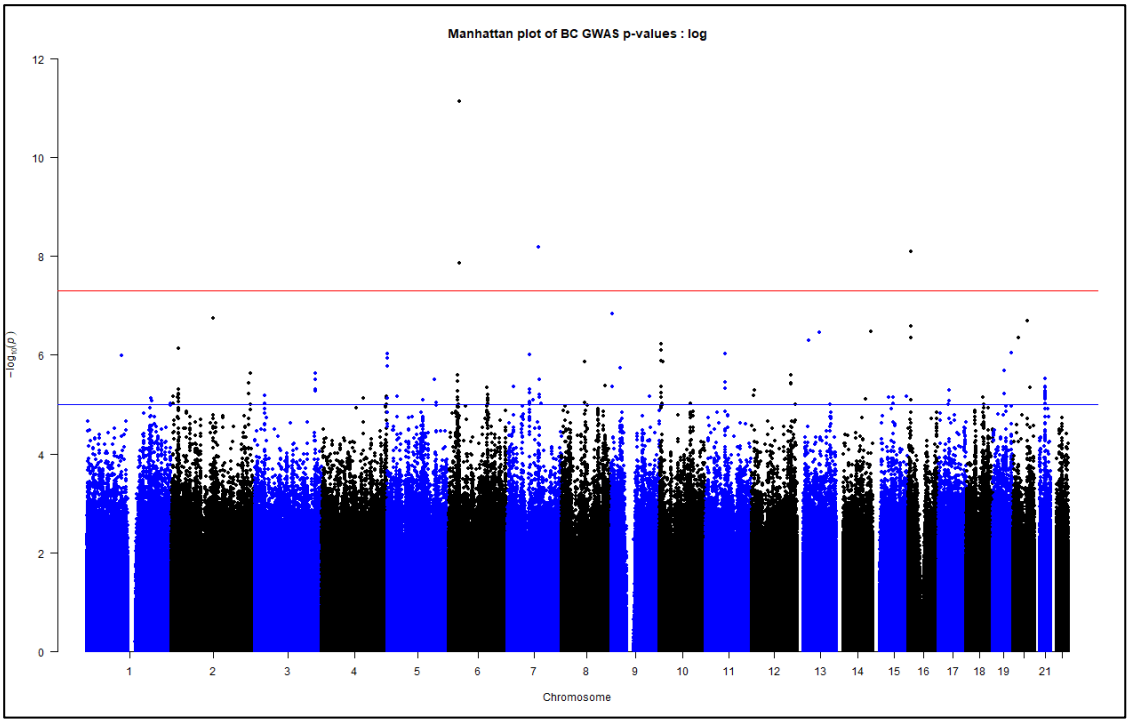


Figure 0.2 Manhattan plot including male and female control samples. 2,711 cases and 2,128 controls. The red line indicates a genome-wide significance p-value of $5x10^{-08}$. The blue line indicates a suggestive significance p-value of $5x10^{-06}$. Plotted against the –log p-value.

4.2.2. Analysis of genetic differences between male and female controls

A possible source of false signals in the initial GWAS was genetic differences between male and female controls from the AWI-Gen project. This idea was prompted by the fact that there are substantial differences in the proportions of the diverse different ethnolinguistic groups in Soweto male and female individuals in AWI-Gen. This is reflected in the population sub-substructure that aligns with the geography and linguistic phylogeny of South African Bantu speakers (Sengupta *et al.*, 2021).

Therefore, iterative testing was carried out between male and female controls randomly sampled both from within-batch and across-batch. Both within-batch and across-batch analysis demonstrated genetic differences between male (n=300) and female (n=300) controls, resulting in genomic inflation in the association analysis, with multiple SNPs showing p-values of $<5x10^{-08}$. This genomic inflation effect was independent of genotyping batches. Figure 4.3A shows the

general inflation of p-values when males are compared to females. Towers of SNPs are seen on the Manhattan plot in several chromosomes and some variants reached p-values<$5x10^{-08}$, the genome-wide significance threshold, indicating genetic differences between the two groups. Figure 4.3B shows the Manhattan plot of AWI-Gen females against other AWI-Gen female controls. Here, no over-inflation of p-values was observed and no variants reached the genome-wide significance threshold, which suggests that the female controls are relatively more homogeneous. As a result of this finding, all male controls (n=1,048) were removed from subsequent analyses.

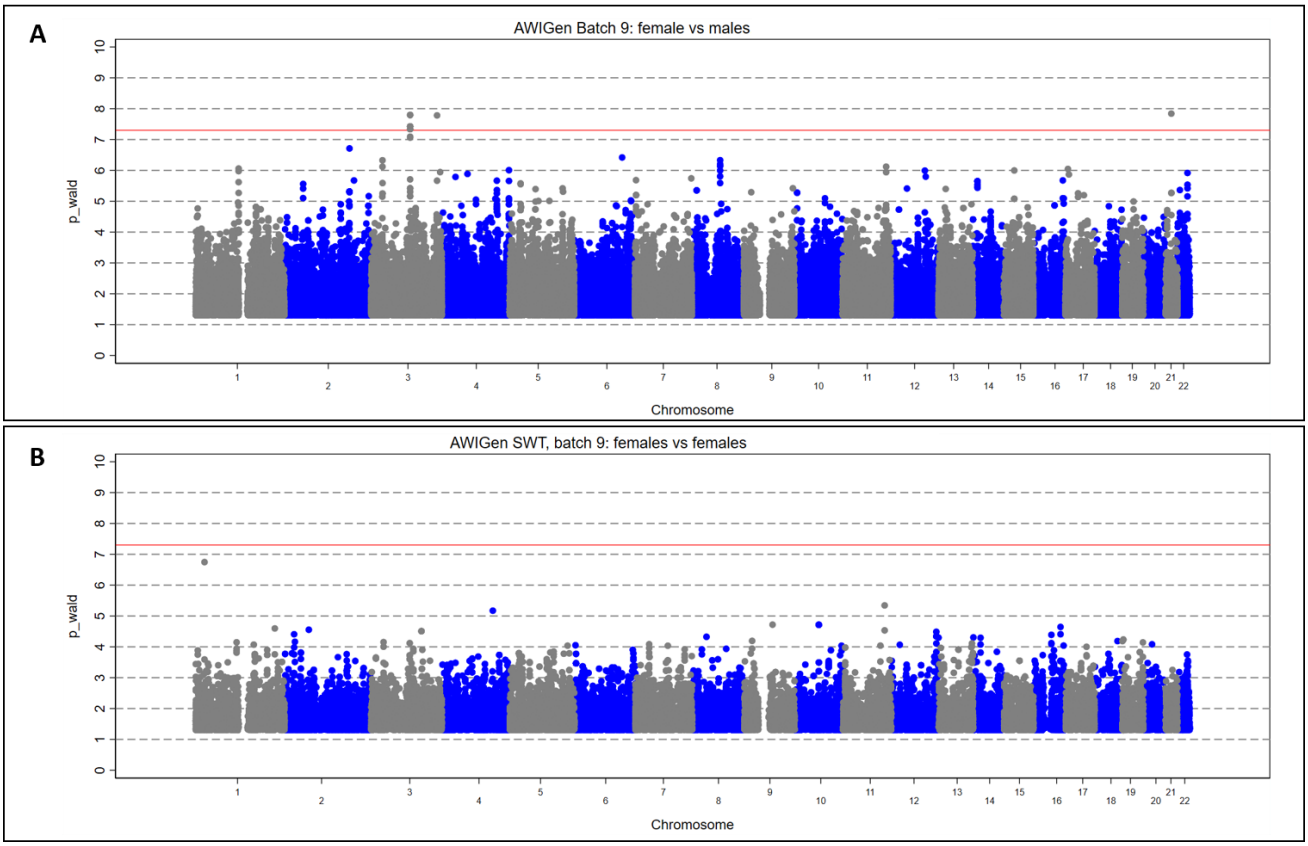


Figure 0.3: Manhattan plots showing differences between male and female controls.
A. Manhattan plot of AWI-Gen male controls against AWI-Gen female controls. B. Manhattan plot of AWI-Gen female controls against AWI-Gen female controls. The red line indicates a genome-wide significance p-value of $5x10^{-08}$. The blue line indicates a suggestive significance p-value of $5x10^{-06}$. Plotted against the –log p-value.

4.2.3. Analysis of batch effects

Our revised study included 1,080 female population controls from the AWI-Gen study and the JCS. However, despite removing male controls from the study, 17 different signals were observed at genome-wide significance, many of which were isolated SNPs with no support from surrounding SNPs that had extremely low p-values for association (Figure 4.4A). These results were then compared with those from a cervical cancer GWAS the ERICA group was conducting that was using the same study approach with cervical cancer cases from the JCS and the same female control dataset from AWI-Gen and the JCS. All SNPs showing suggestive or genome-wide association in the cervical cancer GWAS (p-value<$5x10^{-6}$) were then plotted on a Manhattan plot using the genotype data from the breast cancer cases and the AWI-Gen and JCS female controls (Figure 4.4B). This showed that many of the SNPs associated with cervical cancer also showed a strong association with breast cancer. This degree of overlap was unexpected as cervical cancer and breast cancer do not have known shared genetic aetiologies. Therefore, it prompted us to examine the dataset further to identify the problem.

Since this study used a female-only dataset, iterative testing was done of inter- and intra-genotype batches to determine if there were any differences between the JCS and AWI-Gen genotyping batches. Significantly inflated p-values were observed in random subsets of n=300 vs n=300 controls when the JCS female controls were compared to AWI-Gen female controls (Figure 4.4C). However, when the female controls were compared within each of the JCS and AWI-Gen female control batches, there were no significant differences between random subsets. The results indicated that there was a variation in the JCS genotyping batch that contained the female controls, which was the final batch of genotype data received from the KCL Genome Centre. The raw genotyping fluorescence intensities of the tenth and final batch were higher than the previous nine batches. This caused spurious clustering of the SNPs and produced false SNP calls when the Illumina-Cluster file was used to predefine clustering locations. As seen in Figure 4.4C, the false SNP calls introduced genomic inflation to the dataset. The female controls from the JCS (n=269) were subsequently removed from the dataset.

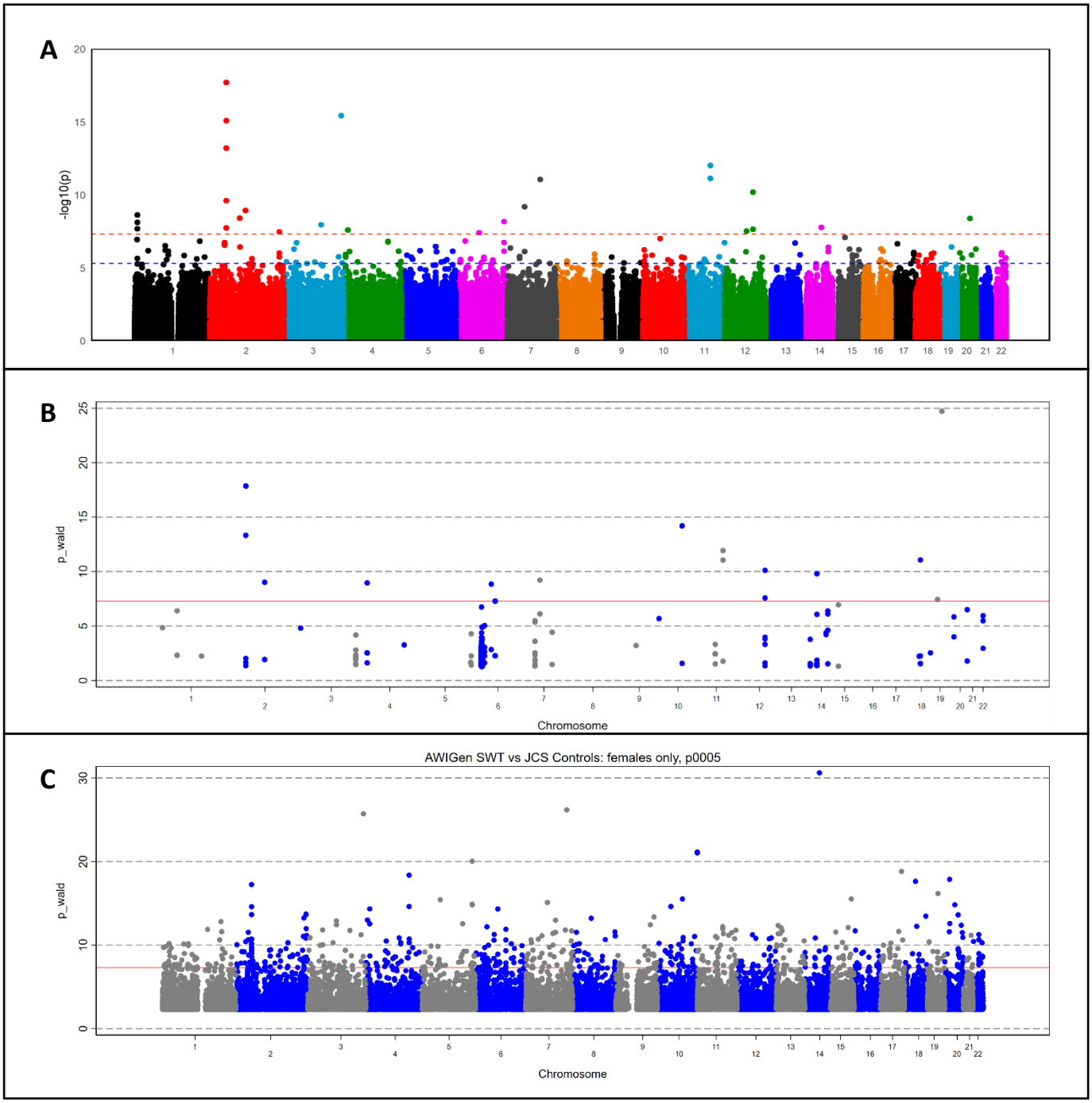


Figure 0.4: Manhattan plots showing inflation of p-values in the dataset after removing male controls.

A. Manhattan plot of 2,711 breast cancer cases against 1,080 AWI-Gen and JCS female controls from the initial modelling**.** B. Manhattan plot showing SNPs which showed association with cervical cancer are also associated with breast cancer. C. Manhattan plot between JCS female controls and AWI-Gen female Soweto controls. p0005: SNPs with p-value>0.005 were not included in the Manhattan plot. The red line indicates a genome-wide significance p-value of $5x10^{-08}$. The blue line indicates a suggestive significance p-value of $5x10^{-06}$. Plotted against the –log p-value.

4.2.4. Admixture analysis

After removing all male controls and the JCS female controls, an admixture analysis was performed of all breast cancer cases and AWI-Gen female controls with K=6 (Figure 4.5). I selected K=6 because cross-validation with the data at K=6 produced the lowest cross-validation error value (0.51286). This analysis revealed that some study participants had significant non-African genetic components. Using the 1000 Genomes project East Asian and European populations as a reference, a total of 205 participants with more than 5% European and East Asian genetic contributions was found. These participants were excluded from further analyses (Table 4.1).

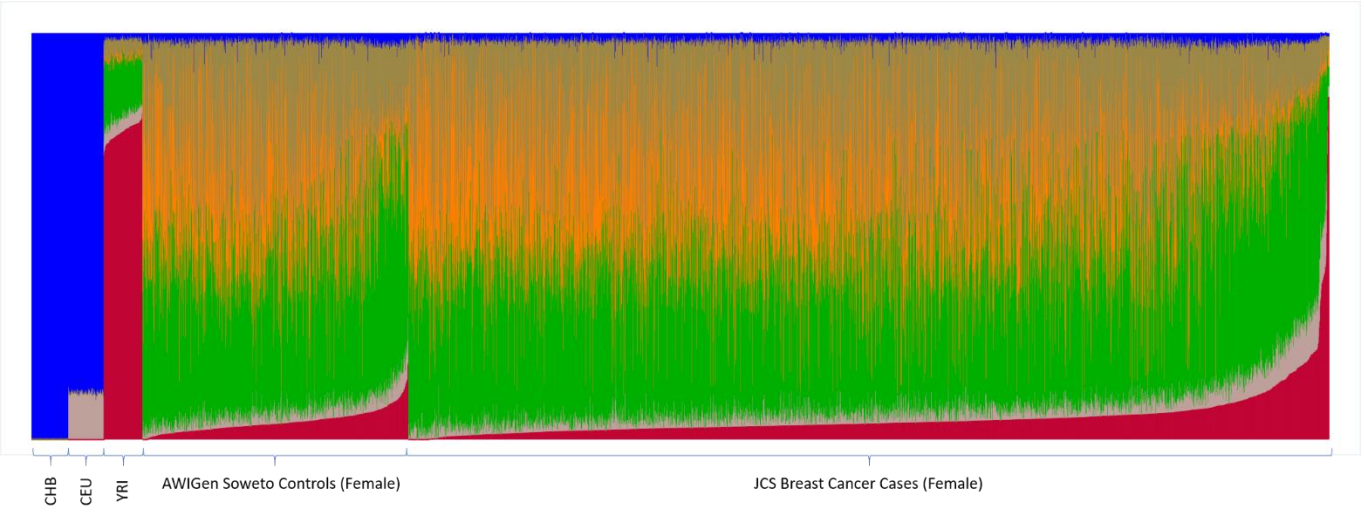


Figure 0.5: Admixture plot
Admixture analysis showing genetic components of European (CEU), East Asian (CHB), Yoruba (YRI), the AWI-Gen Soweto controls and the JCS breast cancer cases. K=6.

Table 0.1. Number of samples with 5% European and East Asian ancestry

	Cases	Controls	**All**
>5% Eurasian contribution removed	138	67	**205**
<5% Eurasian contribution retained	2,573	744	**3,317**

4.2.5. Principal Component Analysis

PC analysis was carried out using a subset of LD-independent genotyped SNPs (n=429,198). Ten PCs were constructed from Eigen-decomposition values. This PC analysis was performed on the dataset that excluded: individuals with >5% CEU or CHB ancestry, all the male controls and female JCS controls. The selection of PCs was done based on the eigenvalue curve (Figure 4.6) of which visual inspection of a point of inflection (arrow in Figure 4.6) as well as a generalised linear model regression of the PCs with the case-control phenotype. This revealed that PCs 1-5 accounted for the majority of variance in the dataset, and were used as covariates in the linear mixed model (LMM). All cases (n=2,573) and controls (n=744) were female, from Soweto, Johannesburg, and were well matched according to the PC analysis plots with overlapping clusters (Figure 4.7).

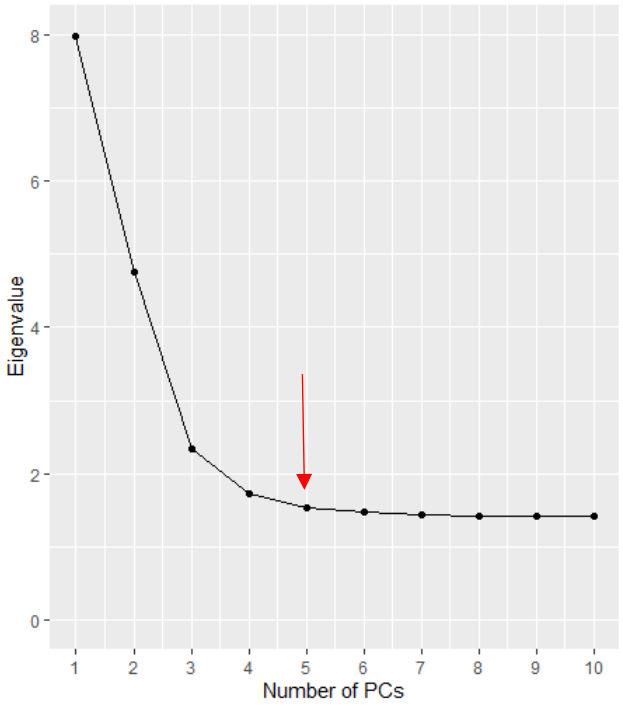


Figure 0.6 Eigenvalue curve of breast cancer cases and controls.
Arrow indicates point of inflection where the number of PCs would account for most of the variation in the dataset.

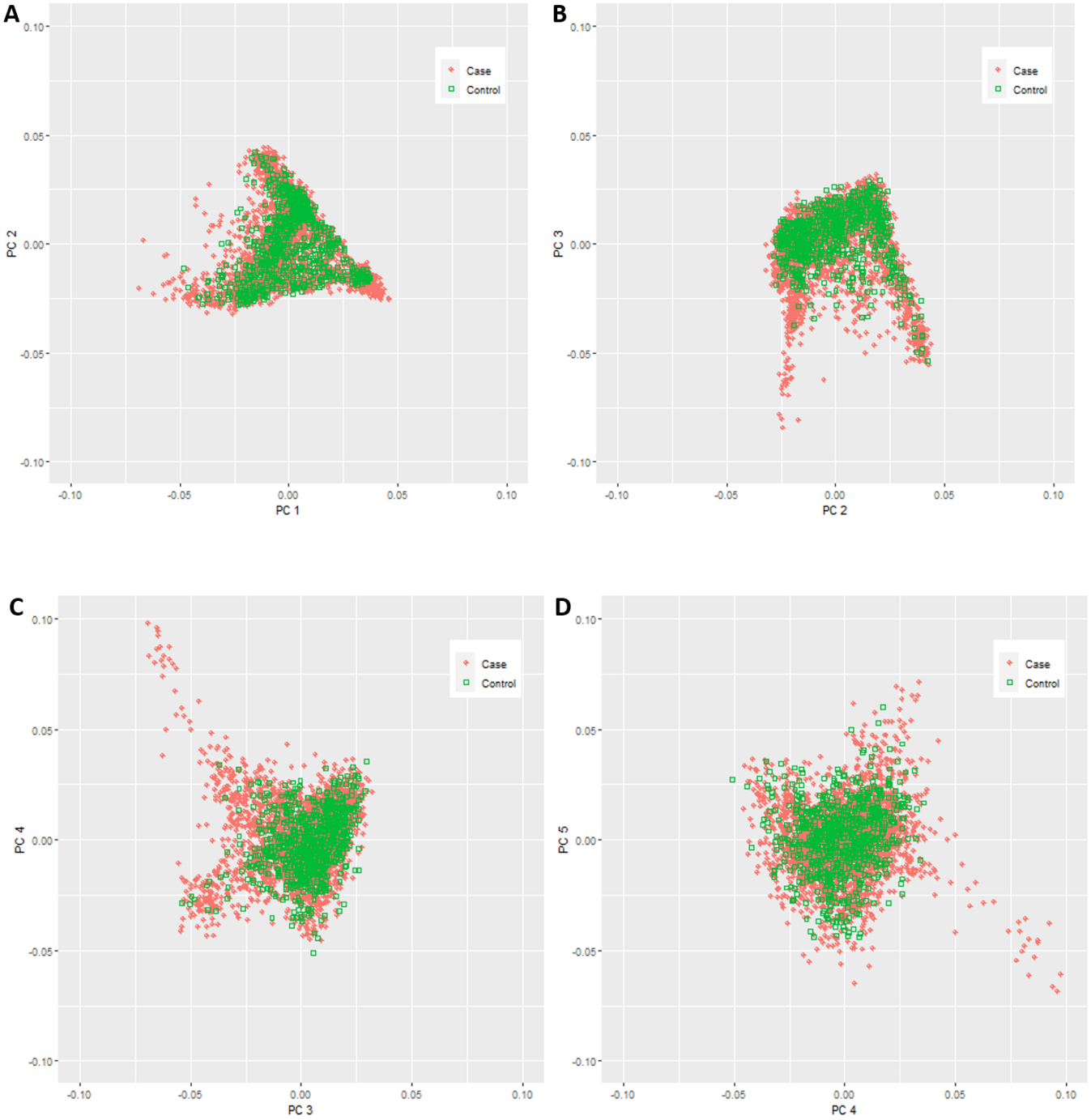


Figure 0.7: Principal Component Analysis plots of cases and controls.
A. PC1 and PC2. B. PC2 and PC3. C. PC3 and PC4. D. PC4 and PC5.

Once the presence of population substructure and the control matching issue arising from using a population control approach for the study had been addressed, the next step was to move onto modelling the association of the SNPs with the breast cancer phenotype. After all measures of QC and population structure correction were applied, a total of 2,573 cases and 744 population controls were retained for the analysis. There were 1,699,678 genotyped SNPs that were submitted for imputation, and post-imputation, QC 14,472,257 SNPs were tested for association with African breast cancer.

The Quantile-Quantile (QQ) plots illustrate the expected -log10(p-value) against the observed -log10(p-value). In Figure 4.8A, the QQ plot shows the effect of removing the males, participants with Eurasian admixture and the female JCS control samples. The expected and observed p-values do not deviate from the reference line until the top end of the plot. An early deviation from the reference line would suggest the presence of genomic inflation in the dataset. This was evident in the QQ plot which included the female JCS control samples (Figure 4.8B), where substantial genomic inflation of p-values was observed.

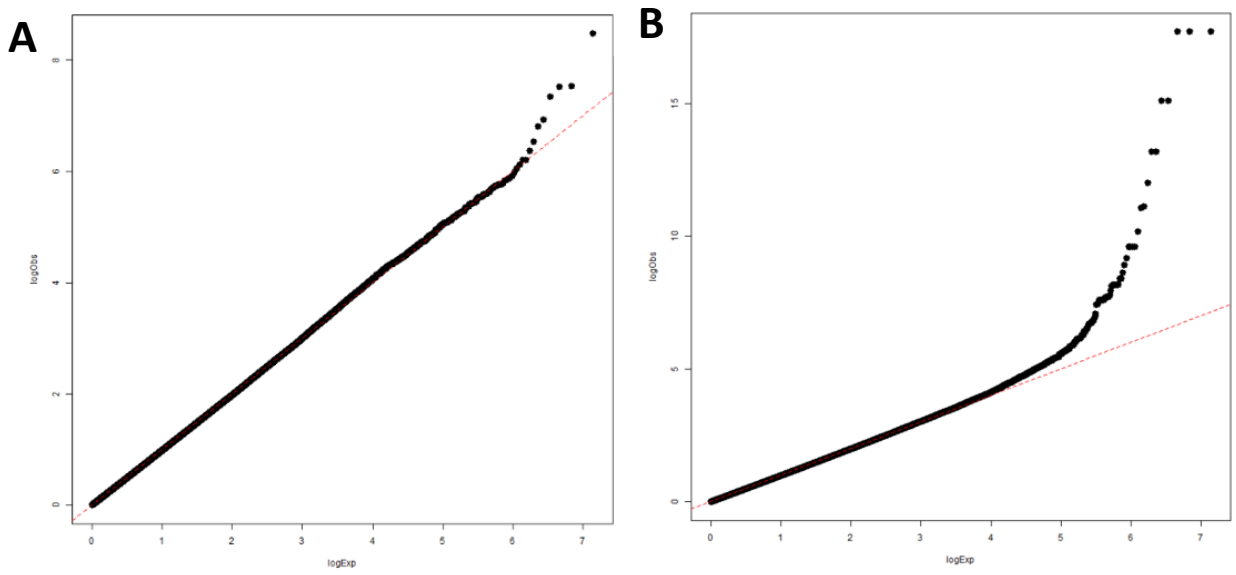


Figure 0.8: Quantile-quantile plots
A. QQ plot of 2,573 breast cancer cases against 744 AWI-Gen controls (genomic inflation factor λ=0.97). The observed p-value line fits the null line without lifting too early. B. QQ plot of 2,711 breast cancer cases against 1,080 AWI-Gen and JCS controls. The QQ plot shows very small p-values and observed p-value line lifting early (λ is 0.98).

4.3. Estimation of heritability

The heritability (h^2_g) for breast cancer using the African dataset was estimated using the generalised restricted maximum likelihood (GREML) estimation implemented in LDAK (Speed *et al.*, 2020). To calculate the h^2_g, 1,699,677 genotyped SNPs were used in a randomised dataset of 744 cases and 744 controls. An equal number of cases and controls were used to prevent biases that the imbalanced case-control ratio may introduce. The 1,699,677 SNPs were pruned and weighted using the LDAK-Thin model. Thereafter, a genetic relatedness matrix (GRM) was constructed and used to estimate SNP-based heritability. The estimated SNP-based heritability was then scaled to the liability scale using the Globocan 2020 South African breast cancer incidence rate of 52.6/100,000 as a proxy for the disease prevalence. The binary phenotype was residualised using a GLM with PCs 1-5 as the covariates to meet the assumptions of the GREML model. The estimated h^2_g for breast cancer was 0.36 (SD=0.13).

4.4. GWAS results

I performed the genome-wide association analysis using 2,573 breast cancer female cases and 744 female ethnically-matched population controls. An LMM method was used because of its effectiveness in correcting for genetic inflation through the use of a kinship matrix (Wen *et al.*, 2018). In the LMM, the binary phenotype was regressed with PCs 1-5 and the kinship relatedness matrix as the covariates. The genomic inflation factor lambda for the LMM model was 0.97 (Figure 4.5B), indicating adequate correction for the population substructure. The Impute2 scores for all of these hits were greater than 0.8 (scale of 0 to 1).

Three SNPs were identified at two independent loci associated with breast cancer at genome-wide significance (p-value<$5x10^{-8}$) which can be seen on the Manhattan plot (Figure 4.9). The two independent loci include two SNPs on chromosome 2 (rs77422433, p-value=$2.89x10^{-08}$, odds ratio (OR):0.46, 95% confidence interval (95%CI): 0.40-0.52 and rs112410019, p-value=$4.54x10^{-08}$, OR: 0.47, 95%CI: 0.41-0.53) and one SNP on chromosome 16 (rs3859109, p-

value=3.01x10^{-08}, OR=0.70, 95%CI: 0.68-0.73) (Table 4.2). An additional 63 SNPs with suggestive association with breast cancer at p-value<5x10^{-06} are shown in Table 4.2 along with the top three hits. The results for the top hits were further supported by the low values for the 700 million adaptive permutation tests that were carried out as described in chapter 3.8 (Table 4.2).

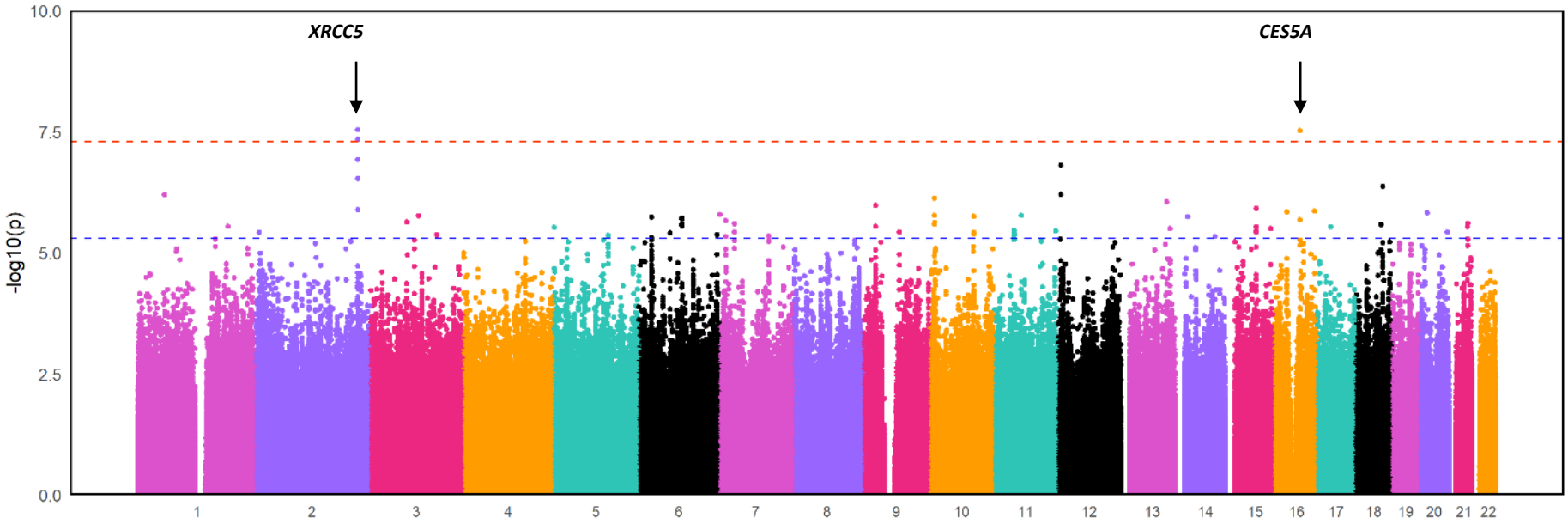


Figure 0.9: Manhattan plot of 2,573 breast cancer cases against 744 AWI-Gen controls.
Top hits on chromosomes 2 and 16 (p-value<$5x10^{-08}$). The red line indicates a genome-wide significance p-value of $5x10^{-08}$. The blue line indicates a suggestive significance p-value of $5x10^{-06}$. Plotted against the –log p-value.

Table 0.2. Top SNPs associated with African breast cancer (p-value<$5x10^{-06}$).

Chr	Genomic position (hg19)	rs ID	Alleles (Effect/non-effect)	Combined EAF	Case EAF	Control EAF	OR(95%CI)	Impute2 Score	P-value	Permutation p-value
2	216999015	rs77422433	T/A	0.05	0.04	0.08	0.46(0.40-0.52)	0.98	$2.89x10^{-08}$	$7.95x10^{-08}$
16	55910551	rs3859109	A/T	0.31	0.29	0.37	0.70(0.68-0.73)	*	$3.01x10^{-08}$	$2.05x10^{-08}$
2	217013506	rs112410019	T/G	0.05	0.04	0.07	0.47(0.41-0.53)	0.99	$4.54x10^{-08}$	$1.32x10^{-07}$
2	217045459	rs115541932	G/C	0.05	0.04	0.07	0.48(0.42-0.54)	0.99	$1.18x10^{-07}$	$2.79x10^{-07}$
12	7877133	rs116065203	T/C	0.02	0.01	0.03	0.30(0.19-0.40)	0.95	$1.56x10^{-07}$	$1.61x10^{-06}$
2	216901488	rs145526842	G/T	0.05	0.04	0.07	0.48(0.42-0.55)	0.98	$2.94x10^{-07}$	$7.32x10^{-07}$
18	59460619	rs865967246	T/A	0.02	0.02	0.04	0.36(0.27-0.45)	0.96	$4.26x10^{-07}$	$7.24x10^{-07}$
12	7945220	rs138749719	G/A	0.02	0.01	0.03	0.35(0.26-0.44)	0.88	$6.22x10^{-07}$	$2.29x10^{-06}$
1	56422262	rs111639778	C/T	0.05	0.04	0.07	0.50(0.44-0.56)	0.95	$6.32x10^{-07}$	$7.93x10^{-07}$
10	9407865	rs188272214	G/T	0.03	0.02	0.05	0.43(0.36-0.51)	0.89	$7.52x10^{-07}$	$1.57x10^{-06}$
13	97963979	rs9516904	C/T	0.26	0.24	0.31	0.72(0.69-0.75)	0.99	$8.83x10^{-07}$	$7.21x10^{-07}$
9	25746158	rs112921862	G/A	0.05	0.04	0.07	0.52(0.46-0.58)	0.94	$1.04x10^{-06}$	$1.70x10^{-06}$
15	65314785	rs371764873	T/C	0.03	0.03	0.05	0.46(0.39-0.53)	0.93	$1.20x10^{-06}$	$2.23x10^{-06}$
2	216880116	rs145253178	A/G	0.05	0.04	0.07	0.50(0.44-0.57)	0.98	$1.29x10^{-06}$	$2.35x10^{-06}$
16	85881838	rs76039744	C/A	0.03	0.02	0.05	0.42(0.34-0.50)	*	$1.35x10^{-06}$	$9.02x10^{-07}$
16	27169029	rs115002746	C/T	0.04	0.03	0.06	0.49(0.43-0.56)	*	$1.41x10^{-06}$	$2.92x10^{-06}$
20	16863532	rs4813253	G/C	0.17	0.18	0.13	1.45(1.42-1.49)	0.92	$1.48x10^{-06}$	$1.85x10^{-06}$
7	1382779	rs115505008	G/T	0.05	0.04	0.07	0.52(0.46-0.58)	0.85	$1.60x10^{-06}$	$2.99x10^{-06}$

Chr	Genomic position (hg19)	rs ID	Alleles (Effect/non-effect)	Combined EAF	Case EAF	Control EAF	OR(95%CI)	Impute2 Score	P-value	Permutation p-value
11	57990067	rs12272272	G/T	0.03	0.04	0.01	2.20(2.13-2.28)	0.97	$1.70x10^{-06}$	$4.60x10^{-06}$
10	9415189	rs1932182	C/A	0.21	0.19	0.25	0.71(0.68-0.74)	*	$1.70x10^{-06}$	$1.35x10^{-06}$
3	102528107	rs66470546	G/A	0.11	0.10	0.15	0.65(0.61-0.69)	*	$1.72x10^{-06}$	$3.48x10^{-06}$
10	92837157	rs75449186	T/G	0.05	0.04	0.07	0.52(0.45-0.58)	0.98	$1.75x10^{-06}$	$3.83x10^{-06}$
14	26911358	rs178232	T/C	0.18	0.16	0.22	0.70(0.66-0.73)	*	$1.81x10^{-06}$	$2.04x10^{-06}$
6	26839716	rs932626	G/T	0.30	0.29	0.35	0.74(0.71-0.77)	0.99	$1.84x10^{-06}$	$1.12x10^{-06}$
6	91626725	rs9451526	T/G	0.10	0.09	0.14	0.63(0.59-0.67)	0.98	$1.92x10^{-06}$	$1.45x10^{-06}$
6	91626257	rs9444802	T/C	0.10	0.09	0.14	0.63(0.59-0.67)	0.98	$1.98x10^{-06}$	$1.43x10^{-06}$
16	55921797	rs3848304	G/A	0.19	0.18	0.23	0.70(0.67-0.73)	*	$2.06x10^{-06}$	$3.14x10^{-06}$
7	12911532	rs139434414	G/A	0.06	0.05	0.08	0.56(0.50-0.61)	0.99	$2.17x10^{-06}$	$3.55x10^{-06}$
3	77643616	rs775722	C/T	0.29	0.31	0.25	1.35(1.32-1.38)	0.97	$2.32x10^{-06}$	$3.40x10^{-06}$
10	10019033	rs111486994	C/A	0.02	0.01	0.03	0.36(0.26-0.45)	0.90	$2.34x10^{-06}$	$2.63x10^{-06}$
21	40399795	rs1000388	G/C	0.04	0.04	0.06	0.50(0.43-0.56)	0.99	$2.47x10^{-06}$	$4.01x10^{-06}$
7	32203233	rs75961274	A/C	0.02	0.01	0.03	0.34(0.24-0.44)	0.98	$2.52x10^{-06}$	$1.17x10^{-05}$
6	91630849	rs12332852	G/A	0.10	0.09	0.14	0.63(0.59-0.68)	0.98	$2.56x10^{-06}$	$1.95x10^{-06}$
10	9416664	rs73577078	A/G	0.20	0.19	0.25	0.71(0.68-0.74)	*	$2.58x10^{-06}$	$2.02x10^{-06}$
18	56238912	rs6566983	C/T	0.39	0.37	0.44	0.75(0.73-0.78)	0.97	$2.62x10^{-06}$	$2.60x10^{-06}$
6	91631760	rs13437071	G/A	0.10	0.09	0.14	0.64(0.59-0.68)	0.98	$2.78x10^{-06}$	$2.16x10^{-06}$
9	25698354	rs4498644	G/A	0.04	0.03	0.05	0.47(0.40-0.54)	0.89	$2.82x10^{-06}$	$4.53x10^{-06}$

Chr	Genomic position (hg19)	rs ID	Alleles (Effect/non-effect)	Combined EAF	Case EAF	Control EAF	OR(95%CI)	Impute2 Score	P-value	Permutation p-value
1	191095808	rs375191664	A/G	0.02	0.01	0.03	0.36(0.27-0.46)	0.96	$2.86x10^{-06}$	$3.24x10^{-06}$
15	65291023	rs116466960	T/C	0.04	0.03	0.06	0.49(0.42-0.55)	0.95	$2.89x10^{-06}$	$3.91x10^{-06}$
21	39213412	rs2836017	C/T	0.05	0.04	0.08	0.55(0.49-0.60)	0.96	$2.89x10^{-06}$	$6.44x10^{-06}$
17	29867268	rs73988096	T/C	0.29	0.28	0.34	0.74(0.71-0.77)	0.99	$2.90x10^{-06}$	$3.08x10^{-06}$
5	1642876	rs192816812	A/G	0.02	0.02	0.04	0.39(0.30-0.48)	0.91	$2.97x10^{-06}$	$4.78x10^{-06}$
13	105416994	rs1652876	A/C	0.26	0.27	0.21	1.37(1.34-1.40)	0.99	$3.12x10^{-06}$	$3.08x10^{-06}$
15	95798419	rs74035605	C/T	0.05	0.05	0.08	0.55(0.49-0.60)	0.89	$3.14x10^{-06}$	$5.43x10^{-06}$
11	42697268	rs61900814	G/A	0.10	0.09	0.13	0.63(0.58-0.67)	*	$3.45x10^{-06}$	$6.24x10^{-06}$
11	131391245	rs11222683	G/A	0.03	0.03	0.05	0.46(0.39-0.53)	0.92	$3.48x10^{-06}$	$4.79x10^{-06}$
15	65289367	rs77552540	C/T	0.04	0.03	0.06	0.49(0.42-0.56)	0.95	$3.60x10^{-06}$	$4.81x10^{-06}$
20	59075447	rs138841903	A/G	0.02	0.02	0.04	0.42(0.33-0.50)	0.95	$3.71x10^{-06}$	$8.37x10^{-06}$
15	65307313	rs115188225	T/C	0.03	0.02	0.05	0.44(0.37-0.52)	0.91	$3.71x10^{-06}$	$5.19x10^{-06}$
7	32192891	rs78941744	A/G	0.02	0.01	0.03	0.35(0.25-0.45)	0.98	$3.72x10^{-06}$	$1.40x10^{-05}$
9	75672082	rs7038560	T/C	0.35	0.36	0.30	1.33(1.30-1.36)	*	$3.72x10^{-06}$	$2.79x10^{-06}$
2	7855796	rs144732806	C/T	0.09	0.08	0.11	0.61(0.57-0.66)	0.91	$3.77x10^{-06}$	$5.65x10^{-06}$
10	92837177	rs17104915	G/C	0.07	0.06	0.10	0.59(0.54-0.64)	0.99	$3.79x10^{-06}$	$7.44x10^{-06}$
10	92837261	rs17104921	A/G	0.07	0.06	0.10	0.59(0.54-0.64)	0.99	$3.79x10^{-06}$	$7.44x10^{-06}$
6	66214075	rs116248478	T/C	0.09	0.08	0.11	0.61(0.57-0.66)	*	$3.87x10^{-06}$	$4.26x10^{-06}$
11	42693837	rs4415751	T/G	0.10	0.09	0.13	0.63(0.58-0.67)	0.99	$3.93x10^{-06}$	$6.97x10^{-06}$

Chr	Genomic position (hg19)	rs ID	Alleles (Effect/non-effect)	Combined EAF	Case EAF	Control EAF	OR(95%CI)	Impute2 Score	P-value	Permutation p-value
10	92837380	rs75524400	G/A	0.07	0.06	0.10	0.59(0.54-0.64)	*	$4.08x10^{-06}$	$8.18x10^{-06}$
3	141339032	rs139299680	C/T	0.02	0.02	0.03	0.38(0.29-0.47)	0.85	$4.24x10^{-06}$	$1.76x10^{-05}$
6	166195063	rs9459506	A/G	0.42	0.40	0.47	0.77(0.74-0.79)	*	$4.25x10^{-06}$	$5.08x10^{-06}$
5	116132396	rs2964670	G/T	0.29	0.28	0.34	0.74(0.71-0.77)	0.98	$4.29x10^{-06}$	$4.55x10^{-06}$
11	42716362	rs61898781	T/G	0.10	0.09	0.13	0.63(0.58-0.67)	*	$4.34x10^{-06}$	$7.59x10^{-06}$
7	104587969	rs553507851	T/C	0.02	0.02	0.04	0.42(0.33-0.50)	0.87	$4.46x10^{-06}$	$1.67x10^{-05}$
7	104584892	rs567549498	A/T	0.02	0.02	0.04	0.42(0.33-0.50)	0.87	$4.46x10^{-06}$	$1.67x10^{-05}$
14	84955836	rs112339522	T/G	0.14	0.13	0.18	0.68(0.64-0.72)	0.99	$4.57x10^{-06}$	$8.01x10^{-06}$
7	12908470	rs79459095	A/G	0.06	0.05	0.08	0.57(0.51-0.62)	0.99	$4.60x10^{-06}$	$7.72x10^{-06}$
6	26801457	rs4236042	C/G	0.38	0.37	0.43	0.76(0.74-0.79)	0.96	$4.92x10^{-06}$	$3.33x10^{-06}$

* Genotyped SNP, EAF: Effect allele frequency, OR: odds ratio, 95%CI: 95% confidence interval, OR and 95% CI calculated in reference to effect allele

The lead signal on chromosome 2, in addition to being supported by two SNPs at genome-wide significance, is further supported by several other SNPs reaching the suggestive significance threshold (p-value<$5x10^{-06}$) (Figure 4.9 and Table 4.2). The LocusZoom plot of this locus shows that the top three SNPs all lie within an intronic region of the *XRCC5* gene (Figure 4.10A), and SNPs in high LD with rs77422433 (r^2>0.8) span across *PECR* (rs145526842, p-value=$2.94x10^{-07}$) and *MREG* (rs145253178, p-value=$1.29x10^{-06}$). The association at this locus is likely to be African-specific as the effect allele has very low frequencies in non-African populations. The genome-wide significant signal on chromosome 16 (rs3859109) is an intronic variant in the gene *CES5A* (Figure 4.10B). It is a genotyped SNP with three distinct genotype clusters (Figure 4.11). The lead signal at this locus is supported by several other SNPs in the region.

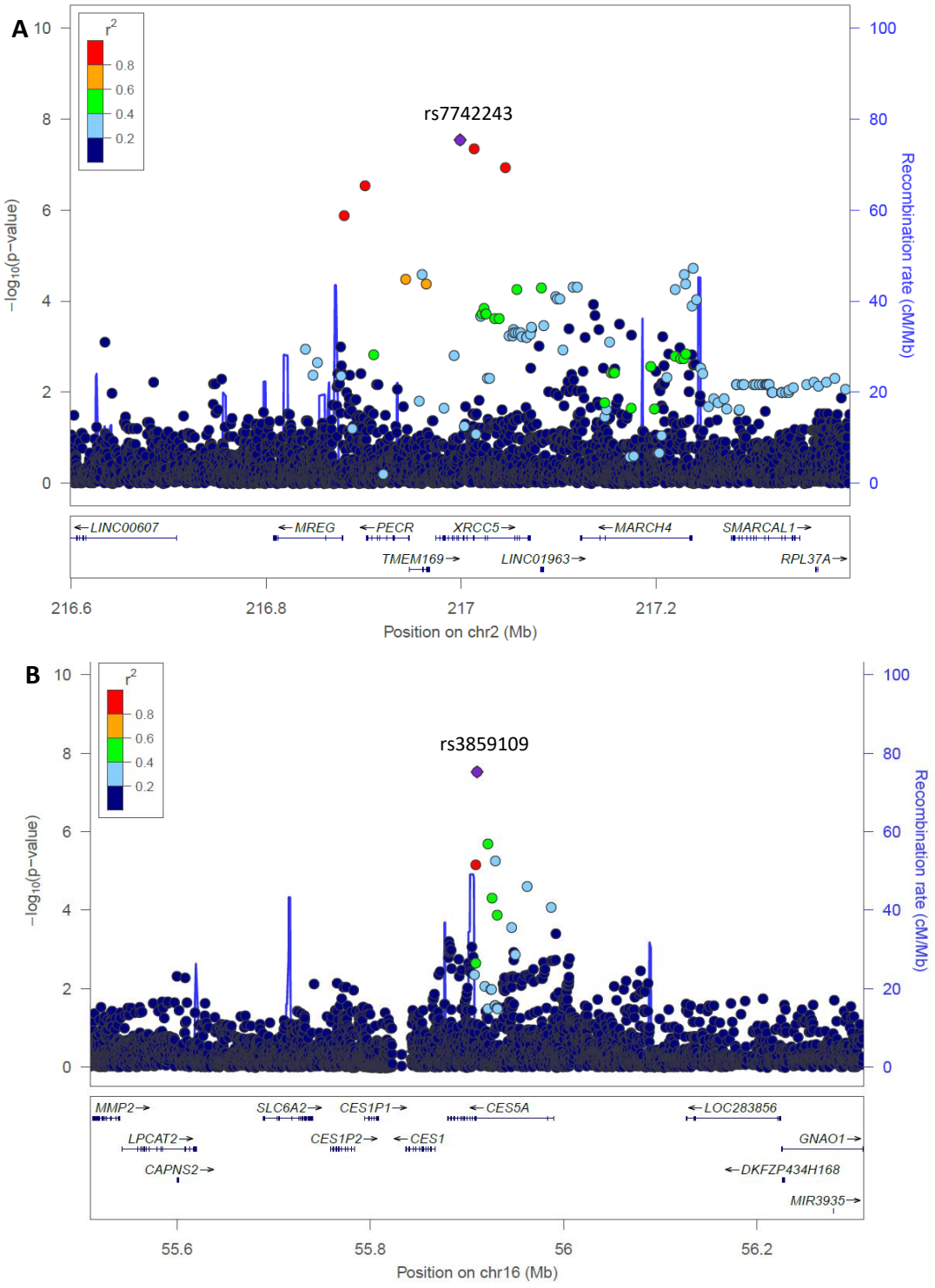


Figure 0.10: Breast cancer GWAS regional plots using LocusZoom.
A. Top hit on chromosome 2, rs77422433 in the *XRCC5* gene. B. Top hit on chromosome 16, rs3859109 in the *CES5A* gene. Plots show top hits with supporting SNPs surrounding it. Plots were generated using resident African GWAS LD as a reference.

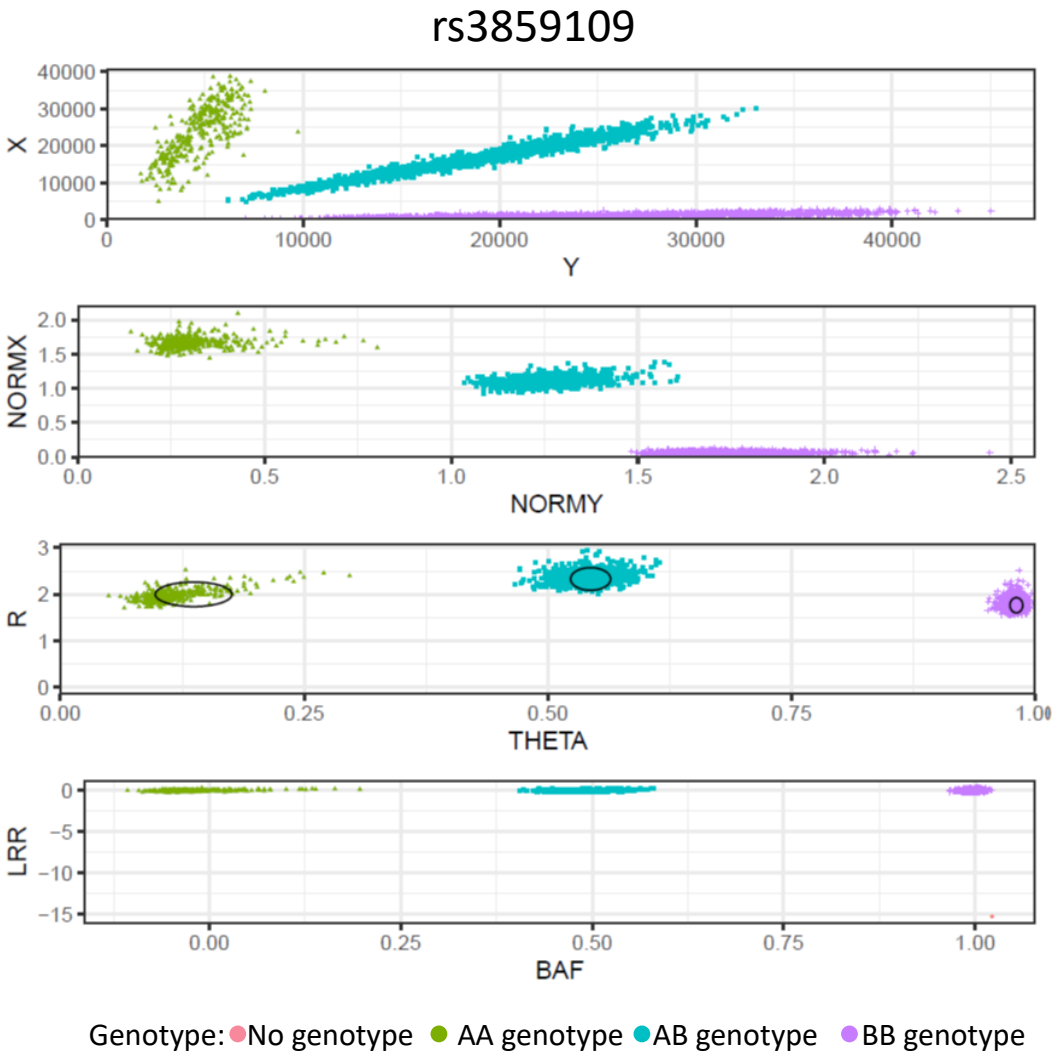


Figure 0.11: Genotype cluster plot of rs3859109.
The top hit on chromosome 16 (genomic location=55910551) was a genotyped SNP. X: raw intensity of A allele, Y: raw intensity of B allele, NORMX: normalised intensity of A allele, NORMY: normalised intensity of B allele, R: normalised R-value for the sample, THETA: normalised theta-value for the sample, LRR: base-2 log of the normalised R value over the expected R value for the theta value (interpolated from the R-values of clusters, BAF: B allele frequency for the sample as interpolated from known B allele frequencies of 3 canonical clusters: 0, 0.5 and 1 if it is equal to or greater than the theta mean of the BB cluster.

A SNP on chromosome 20 (rs2143559) also reached genome-wide significance for association with breast cancer, with a p-value<$5x10^{-08}$. However, this SNP had no evidence of support from surrounding SNPs. This was a genotyped SNP, so the cluster plot was investigated (Figure 4.12) and displayed poor genotype clusters (indicated by red arrows in Figure 4.12). I concluded that the genotype data for this SNP was unreliable I therefore excluded it from analyses.

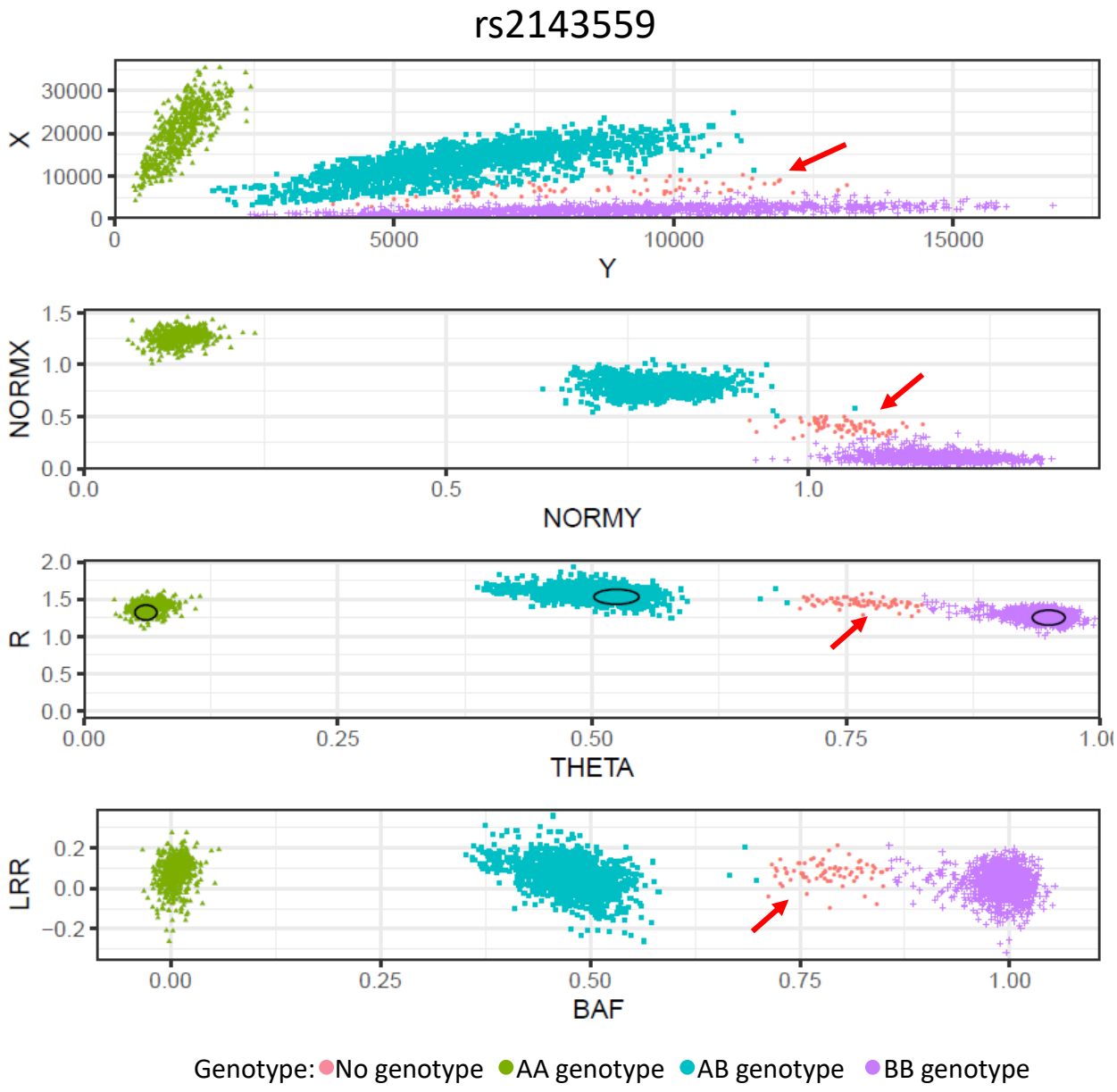


Figure 0.12: Genotype plot for rs2143559.
Genotype plot for rs2143559 on chromosome 20. This SNP was poorly genotyped as can be seen by the poor cluster formation (no genotype classified SNPs). Poor clusters are indicated by red arrows. Figure abbreviations as per Figure 4.11.

4.5. Fine mapping and functional analysis

4.5.1. FINEMAP: 95% credible set

A Bayesian-guided approach to identify the 95% credible sets for the top loci was performed using FINEMAP (v1.4) (Benner *et al.*, 2016). The results from FINEMAP (Table 4.3) showed that there were six SNPs in total that were listed as 95% credible SNPs that could be causal, four SNPs for the chromosome 2 locus and two SNPs for the chromosome 16 locus. The top SNPs from each of the two loci that were significantly associated with breast cancer on chromosomes 2 and 16 were most likely to be the causal SNPs for their respective loci. The SNP on chromosome 2, rs77422433, had a 0.47 posterior probability of being the causal SNP, and the SNP on chromosome 16 (rs3859109) had a 0.94 posterior probability of being the causal SNP. The remaining SNPs had posterior probabilities of less than 0.40. The first two variants listed for chromosome 2 were associated with breast cancer at genome-wide significance, and the latter two showed suggestive association: rs115541932 (p-value=1.18×10^{-07}) and rs145526842 (2.94×10^{-07}). In addition to the top hit on chromosome 16, there was one more SNP in that locus that was listed as a credible causal SNP: rs3848304 (p-value=2.06×10^{-06}). Thus at present rs77422433 on chromosome 2 and rs3859109 on chromosome 16 are the most likely causal variants at these loci.

Table 0.3. Credible set of causal variants in chromosomes 2 and 16

Chr	rs ID	Position	Allele (effect/non-effect)	Probability of being a causal SNP	log10bf*	h^2	h^2 95%CI	Lead SNP according to COJO#
2	rs77422433	216999015	T/A	0.47	2.94	0.0077	0.003-0.014	Yes
	rs112410019	217013506	T/G	0.32	2.66	0.0075	0.003-0.014	No
	rs115541932	217045459	G/C	0.14	2.2	0.0071	0.003-0.012	No
	rs145526842	216901488	G/T	0.06	1.82	0.0066	0.002-0.012	No
16	rs3859109	55910551	A/T	0.94	4.32	0.0077	0.003-0.014	Yes
	rs3848304	55921797	G/A	0.02	1.48	0.0057	0.002-0.011	No

* log10 Bayes factors which quantify the evidence that the SNP is causal. A log10 Bayes factors greater than 2 reports considerable evidence. #Conditional and Joint Analysis.

4.5.2. eQTL results from FUMA

The potential functional consequences of associated SNPs were assessed using FUMA (https://fuma.ctglab.nl/). The FUMA annotation showed that 95% of the variants with suggestive or genome-wide significant association with breast cancer were in intronic or intergenic regions. The genes from both top loci, *XRCC5* (https://www.gtexportal.org/home/gene/XRCC5), and *CES5A* (https://www.gtexportal.org/home/gene/CES5A) are expressed in breast tissue. However, none of the genome-wide significant SNPs in the GWAS had eQTLs associated with them. This is not surprising for the *XRCC5* locus since the associated SNPs are extremely rare or absent in non-African populations thus no eQTL data exists.

However, eight loci were associated with eQTLs ($p_{FDR}<0.05$) and are listed in Table 4.4. The p_{FDR} explains the significance of which the expression of the gene is linked to the SNP. Only one of the eight SNPs showed suggestive significance with breast cancer in the GWAS (rs932626). There were three loci on chromosome 6 that were strongly associated with eQTLs in breast tissue: rs10946859 with *BTN3A2*, rs932626 with *ZNF322* and rs12203512 with *HMGN4*. A locus on chromosome 5 was associated with an eQTL mapped to *C5orf22*. A fifth locus on chromosome 18 was associated with an eQTL mapped to *MALT1* (rs35591632). The sixth locus is on chromosome 19: rs2431807mapped to an eQTL in *SMIM7*. The seventh locus on chromosome 20 mapped to an eQTL in *TOX2*, and the final locus, rs17761473, on chromosome 22 mapped to an eQTL in *SFI1*.

Table 0.4. Results of eQTL analysis in breast tissue mapped to breast cancer from FUMA

			eQTL analysis in breast tissue			African breast cancer GWAS				
rsID	**Chr**	**Genomic Position (hg19)**	**Tested allele**	**P-value (FDR)**	**Gene variant is associated with**	**Alleles (EA/AA)**	**Case EAF**	**Control EAF**	**OR(95%CI)**	**P-value**
rs13190047	5	31565524	T	$9.16x10^{-08}$	*C5orf22*	T/C	0.04	0.06	0.53(0.46-0.59)	$5.88x10^{-06}$
rs932626	6	26839716	G	$4.49x10^{-28}$	*ZNF322*	G/T	0.29	0.35	0.74(0.71-0.77)	$1.84x10^{-06}$
rs10946859	6	26851785	G	$1.46x10^{-72}$	*BTN3A2*	G/A	0.37	0.42	0.81(0.78-0.84)	$5.82x10^{-04}$
rs12203512	6	26906017	T	$4.01x10^{-05}$	*HMGN4*	T/C	0.31	0.37	0.77(0.74-0.80)	$2.53x10^{-05}$
rs35591632	18	56241611	A	$1.30x10^{-06}$	*MALT1*	A/G	0.24	0.28	0.82(0.79-0.85)	$3.71x10^{-03}$
rs2431807	19	16798471	A	$2.50x10^{-05}$	*SMIM7*	A/C	0.24	0.29	0.77(0.74-0.80)	$8.33x10^{-05}$
rs244065	20	42430824	C	$1.60x10^{-27}$	*TOX2*	A/C	0.39	0.36	1.15(1.12-1.18)	$1.98x10^{-02}$
rs17761473	22	32217162	T	$1.38x10^{-14}$	*SFI1*	T/C	0.04	0.06	0.53(0.46-0.59)	$5.88x10^{-06}$

FDR false discovery rate, EA effect allele, AA alternate allele, EAF effect allele frequency, OR odds ratio, 95%CI 95% confidence interval

4.5.3. Functional annotation

SNPs associated with breast cancer at p-value<$5x10^{-04}$ were queried on SNP-nexus (snp-nexus.org). SNP-nexus is a web-based variant annotation tool for novel genomic variant functional annotation. Of all the SNPs that were submitted to SNP-nexus, only three missense variants were annotated as potentially deleterious by both SIFT and PolyPhen are listed in Table 4.5, but they did not reach genome-wide or suggestive significance for association with breast cancer.

rsID	Chr	Genomic Position (hg19)	Gene	SIFT	PolyPhen	Alleles (EA/AA)	Control AF	OR(95%CI)	P-value in African GWAS
rs3213746	2	37458668	*CEBPZ*	0.01 (Deleterious - Low Confidence)	0.755 (Possibly Damaging)	A/G	0.04	0.50(0.42-0.59)	$2.14x10^{-04}$
rs238829	17	39306004	*KRTAP4-5*	0.00 (Deleterious)	0.829 (Possibly Damaging)	G/A	0.21	0.73(0.69-0.76)	$5.62x10^{-05}$
rs2231987	19	19030120	*COPE*	0.01 (Deleterious)	0.518 (Possibly Damaging)	C/G	0.25	1.25(1.22-1.28)	$4.89x10^{-04}$

Table 0.5. SNPs predicted to be deleterious by PolyPhen and SIFT in SNP-nexus

EA effect allele, AA alternate allele, AF allele frequency, OR odds ratio, 95%CI 95% confidence interval

4.6. Replication

4.6.1. Replication of known breast cancer risk SNPs in the African breast cancer GWAS

Previously known breast cancer risk SNPs from the BCAC European (10,052 cases and 12,575 controls) and African American (1,116 cases and 932 controls) data were assessed to see whether they were replicated in the resident African dataset (Michailidou *et al.*, 2013). I assessed the direct replication of the known risk variants. No previously known breast cancer risk variants reached genome-wide significance in the resident African study. There were 17 potential risk SNPs (p-value<$5x10^{-6}$) from the BCAC African American GWAS present in the resident African breast cancer GWAS data, but none of these was replicated at p-value<0.01 (Table 4.6). The lowest p-value was $3.89x10^{-02}$ for rs28710983 on chromosome 1, but this SNP had different directions of effect between the resident African breast cancer GWAS and the BCAC African-American GWAS. Thus, I was not able to replicate any known risk markers previously described for breast cancer.

Table 0.6. Replicating top hits (p-value<$5x10^{-06}$) from BCAC African American hits in African breast cancer GWAS

			BCAC African American					African Breast Cancer GWAS				
Chr	**Position**	**rsID**	**EA**	**AA**	**EAF**	**OR**	**P-value**	**EA**	**AA**	**EAF**	**OR(95%CI)**	**P-value**
1	95027644	rs28710983	T	C	0.21	0.64	$4.50x10^{-06}$	T	C	0.16	1.18(1.14-1.21)	$3.89x10^{-02}$
18	29851354	rs142371098	T	C	0.03	0.25	$1.00x10^{-06}$	T	C	0.03	0.75(0.68-0.82)	$6.91x10^{-02}$
7	156447989	rs35694305	T	G	0.13	1.83	$4.10x10^{-06}$	T	G	0.13	0.87(0.83-0.90)	$1.01x10^{-01}$
18	29817381	rs141241593	T	C	0.03	0.26	$1.70x10^{-06}$	T	C	0.02	0.72(0.62-0.83)	$1.73x10^{-01}$
18	29823912	rs114304899	C	T	0.03	0.25	$5.40x10^{-07}$	C	T	0.02	0.76(0.67-0.85)	$1.76x10^{-01}$
6	40343771	rs9471329	C	T	0.16	0.6	$3.70x10^{-06}$	C	T	0.12	0.90(0.86-0.94)	$2.35x10^{-01}$
7	132017964	rs10263528	T	C	0.24	0.65	$2.10x10^{-06}$	T	C	0.32	0.95(0.92-0.97)	$3.83x10^{-01}$
6	92751723	rs2452973	T	A	0.46	0.73	$4.60x10^{-06}$	T	A	0.37	0.95(0.92-0.98)	$3.85x10^{-01}$
3	174407207	rs511304	A	C	0.25	1.56	$2.20x10^{-06}$	A	C	0.37	1.05(1.02-1.08)	$4.09x10^{-01}$
6	92766490	rs2452979	C	T	0.42	0.7	$2.00x10^{-06}$	C	T	0.36	0.96(0.94-0.99)	$5.30x10^{-01}$
6	92766491	rs2452980	T	A	0.42	0.7	$2.00x10^{-06}$	T	A	0.36	0.96(0.94-0.99)	$5.30x10^{-01}$
11	41175597	rs10837587	A	G	0.09	0.55	$3.50x10^{-06}$	A	G	0.07	0.94(0.89-0.99)	$5.57x10^{-01}$
5	125917125	rs78019446	T	C	0.03	3.1	$2.80x10^{-06}$	T	C	0.07	0.94(0.89-0.99)	$5.78x10^{-01}$
7	132037749	rs73434896	T	A	0.25	0.62	$2.90x10^{-07}$	T	A	0.21	0.96(0.93-0.99)	$5.98x10^{-01}$
3	171367959	rs73038091	A	G	0.24	0.69	$4.50x10^{-06}$	A	G	0.19	1.04(1.00-1.07)	$6.31x10^{-01}$
1	234451146	rs611676	G	A	0.28	0.66	$3.30x10^{-06}$	G	A	0.18	0.99(0.95-1.02)	$8.58x10^{-01}$
18	29140167	rs62095230	A	G	0.13	0.55	$2.00x10^{-06}$	A	G	0.08	1.02(0.97-1.06)	$8.72x10^{-01}$

EA: effect allele, AA: alternate allele, EAF: effect allele frequency, OR: odds ratio, 95%CI: 95% confidence interval

The European BCAC GWAS data described 8,563 SNPs with p-value<$5x10^{-08}$, of which 3,875 SNPs were present in the resident African breast cancer GWAS. Of the 3,875 SNPs, 33 SNPs showed modest associations (p-value<0.01) in the resident African dataset (Table 4.7). Two European loci showed some evidence of replication in the resident African dataset: chromosome 10 (rs719338, p-value=$1.02x10^{-04}$, OR=1.26, 95%CI=1.24-1.29) and chromosome 3 (rs2322830, p-value=$3.08x10^{-03}$, OR=0.83, 95%CI=0.81-0.86), but would not be significant after correction for the number of tests done.

Table 0.7. Test for replicating top hits (p-value<$5x10^{-08}$) from BCAC European hits in the African breast cancer GWAS.

			BCAC European					African Breast Cancer GWAS				
Chr	**Position**	**rs ID**	**EA**	**AA**	**Control EAF**	**OR**	**P-value**	**EA**	**AA**	**Control EAF**	**OR(95%CI)**	**P-value**
10	80851257	rs719338	T	G	0.61	0.92	$3.70x10^{-16}$	G	T	0.34	1.26(1.24-1.29)	$1.02x10^{-04}$
10	80852378	rs1268974	G	A	0.61	0.92	$6.00x10^{-16}$	A	G	0.29	1.26(1.23-1.29)	$2.13x10^{-04}$
10	80849153	rs1269867	T	C	0.51	0.94	$2.80x10^{-09}$	T	C	0.49	0.83(0.81-0.86)	$1.66x10^{-03}$
10	80853247	rs1250014	C	G	0.57	0.92	$1.40x10^{-15}$	G	C	0.06	1.41(1.36-1.46)	$1.95x10^{-03}$
3	4745230	rs2322830	C	T	0.56	1.06	$3.80x10^{-09}$	T	C	0.40	0.83(0.81-0.86)	$3.08x10^{-03}$
10	80846517	rs1250006	A	T	0.50	0.94	$4.20x10^{-09}$	T	A	0.45	1.18(1.16-1.21)	$4.23x10^{-03}$
3	4731468	rs9866837	T	C	0.56	1.06	$4.80x10^{-09}$	T	C	0.23	1.21(1.18-1.24)	$5.50x10^{-03}$
3	4731593	rs6803046	A	G	0.56	1.06	$5.30x10^{-09}$	A	G	0.23	1.21(1.18-1.24)	$5.50x10^{-03}$
3	4731442	rs28391200	A	G	0.56	1.06	$5.90x10^{-09}$	A	G	0.23	1.21(1.18-1.24)	$5.50x10^{-03}$
3	4731505	rs6802947	A	G	0.56	1.06	$5.90x10^{-09}$	A	G	0.23	1.21(1.18-1.24)	$5.50x10^{-03}$
3	4731533	rs9828130	A	G	0.56	1.06	$6.10x10^{-09}$	A	G	0.23	1.21(1.18-1.24)	$5.64x10^{-03}$
2	218302013	rs2618146	G	A	0.32	1.07	$4.10x10^{-12}$	G	A	0.42	1.18(1.15-1.20)	$5.66x10^{-03}$
10	80842827	rs704007	A	C	0.51	0.95	$6.40x10^{-09}$	C	A	0.47	1.18(1.15-1.20)	$5.72x10^{-03}$
3	4732040	rs6803471	A	G	0.56	1.06	$6.30x10^{-09}$	A	G	0.23	1.20(1.18-1.23)	$5.83x10^{-03}$
3	4730867	rs4685799	T	C	0.56	1.06	$4.50x10^{-09}$	C	T	0.42	0.85(0.82-0.87)	$6.16x10^{-03}$
16	52579906	rs8044585	A	C	0.44	1.11	$4.20x10^{-29}$	C	A	0.27	1.19(1.16-1.22)	$6.33x10^{-03}$
3	4731942	rs6792465	G	A	0.55	1.06	$4.80x10^{-09}$	G	A	0.23	1.20(1.17-1.23)	$6.41x10^{-03}$
3	4731977	rs6792539	G	A	0.56	1.06	$5.70x10^{-09}$	G	A	0.23	1.20(1.17-1.23)	$6.41x10^{-03}$
22	40819013	rs73167030	C	T	0.10	1.12	$8.30x10^{-14}$	C	T	0.28	1.19(1.16-1.22)	$6.56x10^{-03}$
2	217936148	rs12996888	T	G	0.23	0.91	$1.10x10^{-16}$	T	G	0.36	0.84(0.81-0.87)	$6.77x10^{-03}$
3	4732352	rs9833138	T	G	0.53	1.06	$1.10x10^{-08}$	T	G	0.24	1.20(1.17-1.23)	$7.31x10^{-03}$
16	80643675	rs2060861	C	T	0.23	1.08	$1.30x10^{-11}$	T	C	0.47	0.85(0.83-0.88)	$7.36x10^{-03}$
22	40819154	rs12159970	A	G	0.10	1.12	$1.60x10^{-13}$	A	G	0.28	1.19(1.16-1.22)	$7.64x10^{-03}$
16	52582816	rs1362550	G	C	0.46	1.13	$9.40x10^{-35}$	C	G	0.25	1.19(1.16-1.22)	$7.93x10^{-03}$
3	4742825	rs9846046	C	T	0.56	1.06	$4.80x10^{-09}$	T	C	0.38	0.85(0.82-0.88)	$8.04x10^{-03}$
3	4734181	rs11918796	C	T	0.56	1.06	$4.40x10^{-09}$	T	C	0.48	0.85(0.83-0.88)	$8.04x10^{-03}$
3	4731586	rs6803045	A	G	0.56	1.06	$3.50x10^{-09}$	G	A	0.40	0.85(0.82-0.88)	$8.11x10^{-03}$
10	80855423	rs10762844	A	G	0.39	1.07	$2.60x10^{-12}$	A	G	0.46	1.16(1.14-1.19)	$8.69x10^{-03}$
16	52583847	rs3095605	G	A	0.46	1.12	$1.10x10^{-34}$	A	G	0.25	1.19(1.16-1.22)	$9.01x10^{-03}$
5	56043426	rs7716644	T	C	0.57	0.91	$3.50x10^{-21}$	T	C	0.26	0.83(0.80-0.87)	$9.35x10^{-03}$
2	217927421	rs17778798	C	T	0.25	0.91	$4.80x10^{-18}$	C	T	0.22	0.83(0.79-0.86)	$9.54x10^{-03}$
2	218301119	rs4674176	C	G	0.71	0.93	$1.70x10^{-13}$	G	C	0.25	1.19(1.16-1.22)	$9.63x10^{-03}$
16	52580247	rs4784224	A	C	0.45	1.13	$1.30x10^{-34}$	C	A	0.33	1.17(1.14-1.20)	$9.69x10^{-03}$

EA: effect allele, AA: alternate allele, EAF: effect allele frequency, OR: odds ratio, 95%CI: 95% confidence interval

4.6.2. Replication of South African breast cancer SNPs in other populations

I assessed whether any of the 67 SNPs from the resident African breast cancer GWAS with p-value<$5x10^{-6}$ were replicated in the African-American GWAS or the European population GWAS from BCAC. Of the 67 SNPs, 60 SNPs were present in the African-American dataset, four SNPs in the dataset had p-$value_{AA}$<0.05 and only one SNP reached p-$value_{AA}$<0.01 (Table 4.8): chromosome 3, rs139299680 (p-$value_{AA}$=0.004).

Table 0.8. Replicating top hits from the African breast cancer GWAS in the BCAC African American GWAS (p-value<0.05)

			African Breast Cancer GWAS					BCAC African American				
Chr	**rs ID**	**Position**	**EA**	**AA**	**EAF**	**OR(95%CI)**	**P-value**	**EA**	**AA**	**EAF**	**OR**	**P-value**
3	rs139299680	141339032	C	T	0.02	0.38(0.29-0.47)	$4.24x10^{-06}$	C	T	0.03	0.50	$4.20x10^{-03}$
18	rs6566983	56238912	C	T	0.39	0.75(0.73-0.78)	$2.62x10^{-06}$	T	C	0.56	1.20	$3.10x10^{-02}$
6	rs116248478	66214075	T	C	0.09	0.61(0.57-0.66)	$3.87x10^{-06}$	T	C	0.10	1.29	$3.40x10^{-02}$
2	rs144732806	7855796	C	T	0.09	0.61(0.57-0.66)	$3.77x10^{-06}$	C	T	0.01	2.46	$3.60x10^{-02}$

EA: effect allele, AA: alternate allele, EAF: effect allele frequency, OR: odds ratio, 95%CI: 95% confidence interval

There were far fewer SNPs shared in the European population GWAS where only 32 SNPs of the 67 SNPs were present, with no evidence for replication (p-$value_{EUR}$<0.01) of the African breast cancer GWAS top hits (Table 4.9).

Table 0.9. Replicating top hits from the African breast cancer GWAS in the BCAC European GWAS

			African Breast Cancer GWAS					BCAC European				
Chr	**rs ID**	**Position**	**EA**	**AA**	**Control EAF**	**OR(95%CI)**	**P-value**	**EA**	**AA**	**Control EAF**	**OR**	**P-value**
11	rs11222683	131391245	G	A	0.05	0.46(0.39-0.53)	$3.48x10^{-06}$	G	A	0.08	1.06	$2.10x10^{-02}$
16	rs3848304	55921797	G	A	0.23	0.7(0.67-0.73)	$2.06x10^{-06}$	A	G	0.76	0.98	$1.20x10^{-01}$
3	rs775722	77643616	C	T	0.25	1.35(1.32-1.38)	$2.32x10^{-06}$	C	T	0.46	0.98	$1.60x10^{-01}$
5	rs2964670	116132396	G	T	0.34	0.74(0.71-0.77)	$4.29x10^{-06}$	G	T	0.42	0.99	$1.70x10^{-01}$
10	rs1932182	9415189	C	A	0.25	0.71(0.68-0.74)	$1.70x10^{-06}$	C	A	0.06	0.97	$1.90x10^{-01}$
21	rs1000388	40399795	G	C	0.06	0.5(0.43-0.56)	$2.47x10^{-06}$	G	C	0.17	0.98	$2.00x10^{-01}$
11	rs61900814	42697268	G	A	0.13	0.63(0.58-0.67)	$3.45x10^{-06}$	G	A	0.18	1.02	$2.20x10^{-01}$
10	rs73577078	9416664	A	G	0.25	0.71(0.68-0.74)	$2.58x10^{-06}$	A	G	0.06	0.97	$2.30x10^{-01}$
11	rs4415751	42693837	T	G	0.13	0.63(0.58-0.67)	$3.93x10^{-06}$	T	G	0.18	1.01	$2.40x10^{-01}$
6	rs4236042	26801457	C	G	0.43	0.76(0.74-0.79)	$4.92x10^{-06}$	C	G	0.66	1.01	$2.50x10^{-01}$
11	rs61898781	42716362	T	G	0.13	0.63(0.58-0.67)	$4.34x10^{-06}$	T	G	0.18	1.01	$2.60x10^{-01}$
10	rs17104921	92837261	A	G	0.10	0.59(0.54-0.64)	$3.79x10^{-06}$	A	G	0.08	0.98	$2.90x10^{-01}$
10	rs75524400	92837380	G	A	0.10	0.59(0.54-0.64)	$4.08x10^{-06}$	G	A	0.08	0.98	$3.00x10^{-01}$
10	rs75449186	92837157	T	G	0.07	0.52(0.45-0.58)	$1.75x10^{-06}$	T	G	0.08	0.98	$3.10x10^{-01}$
10	rs17104915	92837177	G	C	0.10	0.59(0.54-0.64)	$3.79x10^{-06}$	G	C	0.08	0.98	$3.10x10^{-01}$
6	rs9459506	166195063	A	G	0.47	0.77(0.74-0.79)	$4.25x10^{-06}$	A	G	0.33	0.99	$3.40x10^{-01}$
1	rs111639778	56422262	C	T	0.07	0.5(0.44-0.56)	$6.32x10^{-07}$	C	T	0.00	1.16	$3.60x10^{-01}$
13	rs1652876	105416994	A	C	0.21	1.37(1.34-1.4)	$3.12x10^{-06}$	A	C	0.55	0.99	$3.70x10^{-01}$
14	rs178232	26911358	T	C	0.22	0.7(0.66-0.73)	$1.81x10^{-06}$	C	T	0.86	1.02	$4.20x10^{-01}$
9	rs7038560	75672082	T	C	0.30	1.33(1.3-1.36)	$3.72x10^{-06}$	T	C	0.01	0.93	$4.40x10^{-01}$
18	rs6566983	56238912	C	T	0.44	0.75(0.73-0.78)	$2.62x10^{-06}$	T	C	0.49	0.99	$4.50x10^{-01}$
2	rs144732806	7855796	C	T	0.11	0.61(0.57-0.66)	$3.77x10^{-06}$	C	T	0.00	0.91	$4.50x10^{-01}$
20	rs4813253	16863532	G	C	0.13	1.45(1.42-1.49)	$1.48x10^{-06}$	G	C	0.57	1.01	$4.60x10^{-01}$
16	rs3859109	55910551	A	T	0.37	0.7(0.68-0.73)	$3.01x10^{-08}$	T	A	0.54	0.99	$5.30x10^{-01}$
17	rs73988096	29867268	T	C	0.34	0.74(0.71-0.77)	$2.90x10^{-06}$	T	C	0.02	0.98	$5.40x10^{-01}$
21	rs2836017	39213412	C	T	0.08	0.55(0.49-0.6)	$2.89x10^{-06}$	T	C	0.74	0.99	$6.20x10^{-01}$
5	rs192816812	1642876	A	G	0.04	0.39(0.3-0.48)	$2.97x10^{-06}$	A	G	0.00	0.97	$7.90x10^{-01}$
6	rs932626	26839716	G	T	0.35	0.74(0.71-0.77)	$1.84x10^{-06}$	G	T	0.53	1.00	$8.60x10^{-01}$
13	rs9516904	97963979	C	T	0.31	0.72(0.69-0.75)	$8.83x10^{-07}$	C	T	0.50	1.00	$8.90x10^{-01}$
6	rs116248478	66214075	T	C	0.11	0.61(0.57-0.66)	$3.87x10^{-06}$	T	C	0.00	1.01	$9.30x10^{-01}$
16	rs76039744	85881838	C	A	0.05	0.42(0.34-0.5)	$1.35x10^{-06}$	C	A	0.03	1.00	$9.40x10^{-01}$
3	rs66470546	102528107	G	A	0.15	0.65(0.61-0.69)	$1.72x10^{-06}$	A	G	0.80	1.00	$9.70x10^{-01}$

EA: effect allele, AA: alternate allele, EAF: effect allele frequency, OR: odds ratio95%CI: 95% confidence interval

4.7. Meta-analysis

A meta-analysis was carried out using the African American GWAS data from BCAC and the resident African breast cancer GWAS to assess whether an increase in power derived from two populations of African ancestry might provide further insight into a potential shared risk variants. The meta-analysis was performed under fixed-effect and weighted by the sample-size of each study. There were 2,573 cases and 744 controls from the African breast cancer GWAS, and 1,116 cases and 932 controls in the African American GWAS data. The Manhattan plot of the meta-analysis showed some of the loci of interest on chromosomes 2, 3, 10 and 18 (Figure 4.13). While no SNPs reached genome-wide significance in the meta-analysis, one SNP on chromosome 3, rs139299680, reached a p-value of 7.25×10^{-08} with no support from surrounding SNPs (Figure 4.13), and a further 22 SNPs showed suggestive evidence of association, including nine SNPs at the same locus on chromosome 10 (Table 4.10). Although this locus does not show a tower on the Manhattan plot, it has a lot of support at p-value$<2.5 \times 10^{-06}$.

Table 0.10. Meta-analysis with African breast cancer data and BCAC African American data

Chr	Position	rsID	Allele 1	Allele 2	Meta-analysis p-value	Direction of effect	African GWAS p-value	BCAC African American GWAS p-value
3	141339032	rs139299680	T	C	7.25×10^{-08}	++	4.24×10^{-06}	4.20×10^{-03}
18	56238912	rs6566983	T	C	4.98×10^{-07}	++	2.62×10^{-06}	3.10×10^{-02}
15	61230296	rs4775326	A	G	9.84×10^{-07}	++	1.54×10^{-04}	1.90×10^{-03}
15	91586210	rs4932386	C	G	1.21×10^{-06}	--	1.90×10^{-04}	1.90×10^{-03}
9	33132788	rs34813509	A	G	1.25×10^{-06}	--	2.42×10^{-04}	0.15
2	217013506	rs112410019	T	G	1.29×10^{-06}	--	4.54×10^{-08}	0.38
18	56241254	rs7232632	A	G	1.49×10^{-06}	--	8.75×10^{-06}	3.30×10^{-02}
2	216999015	rs77422433	A	T	1.52×10^{-06}	++	2.89×10^{-08}	0.47
8	801068	rs73673021	A	T	1.66×10^{-06}	--	2.29×10^{-05}	1.80×10^{-02}
10	92304027	rs74148447	A	G	2.04×10^{-06}	++	7.50×10^{-06}	4.70×10^{-02}
10	92304037	rs74148759	A	G	2.04×10^{-06}	--	7.50×10^{-06}	4.70×10^{-02}
10	92304190	rs149171056	A	G	2.04×10^{-06}	++	7.50×10^{-06}	4.70×10^{-02}
13	100567716	rs1924101	C	G	2.23×10^{-06}	++	1.10×10^{-04}	6.20×10^{-03}
8	136442797	rs56204805	A	G	2.28×10^{-06}	--	7.33×10^{-05}	0.6
10	92306530	rs74148762	A	C	2.45×10^{-06}	--	9.31×10^{-06}	4.70×10^{-02}
10	92308153	rs114231835	T	G	2.45×10^{-06}	--	9.31×10^{-06}	4.70×10^{-02}
10	92308248	rs77366015	T	G	2.45×10^{-06}	++	9.31×10^{-06}	4.70×10^{-02}
10	92309037	rs79846022	A	G	2.45×10^{-06}	++	9.31×10^{-06}	4.70×10^{-02}
10	92309526	rs7084013	T	C	2.45×10^{-06}	--	9.31×10^{-06}	4.70×10^{-02}
10	92312110	rs7087035	T	C	2.51×10^{-06}	++	9.31×10^{-06}	4.80×10^{-02}
20	59665974	rs145559620	A	G	2.80×10^{-06}	--	2.71×10^{-04}	3.20×10^{-03}
16	55881326	rs111922842	T	C	2.88×10^{-06}	--	6.17×10^{-04}	1.30×10^{-03}
11	44053415	rs74957006	C	G	2.96×10^{-06}	--	1.20×10^{-04}	7.60×10^{-03}

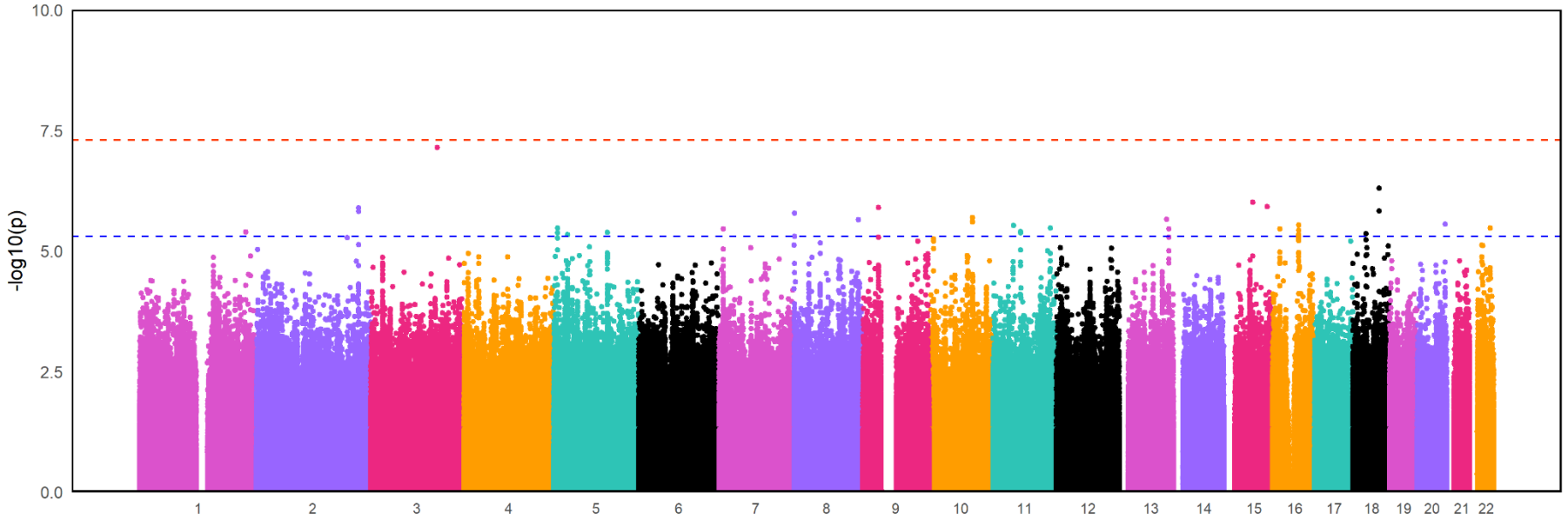


Figure 0.13: Manhattan plot of meta-analysis.
Manhattan plot shows no significantly associated SNPs with breast cancer in the meta-analysis of the African breast cancer data and the BCAC African American data. Red dotted line: p-value5x10^{-08}. Blue dotted line: p-value5x10^{-06}

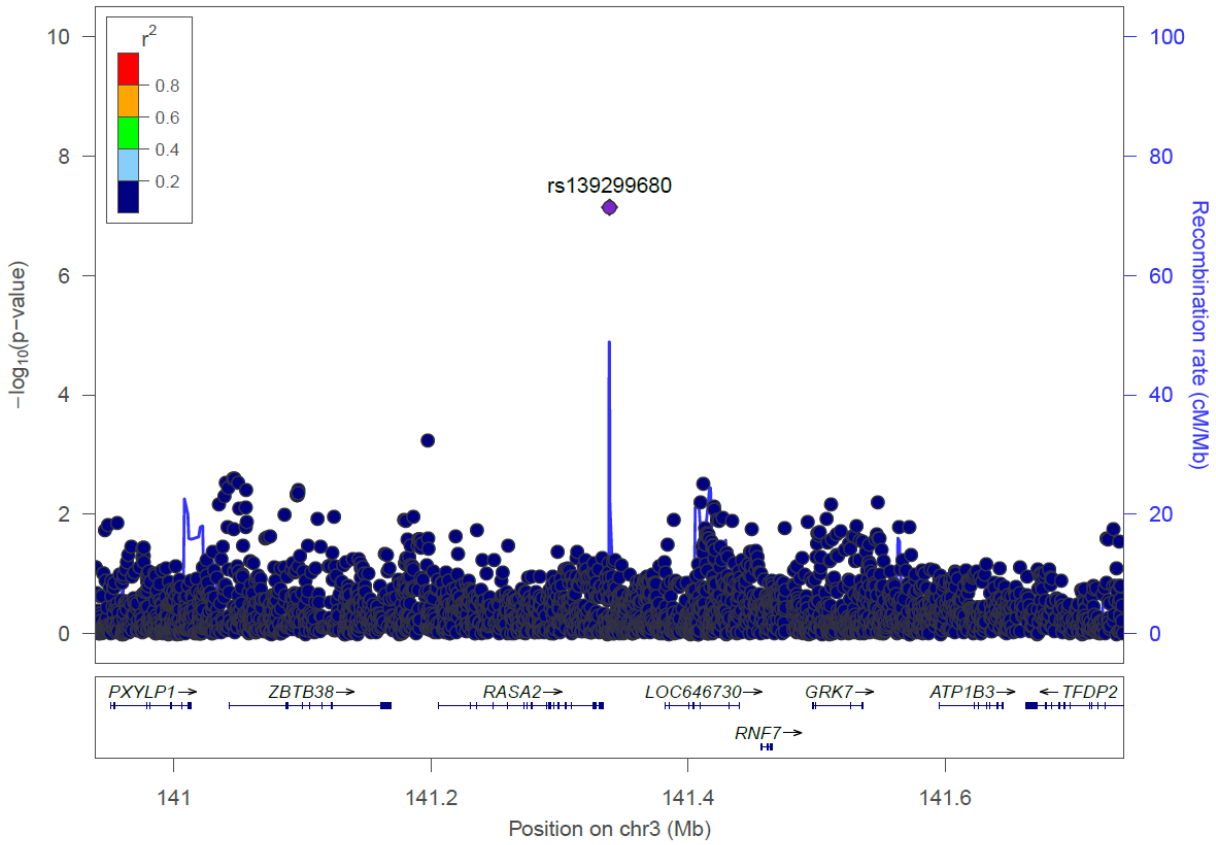


Figure 0.14: Regional plot using LocusZoom of the top marker in the meta-analysis of the resident African data and BCAC African American data.

Although evidence of trans-ethnic replication was limited, the possible reasons for this will be discussed in chapter 5.

4.8. Polygenic Risk Score

The performance of a polygenic risk score for breast cancer previously described by Mavaddat *et al.*, (2019) for European populations was assessed in the resident African GWAS dataset. Of the 313 SNPs used for the breast cancer PRS by Mavaddat *et al*., (2019), only 201 SNPs were present in the resident African dataset (genotyped and/or imputed). This PRS model constructed from the 201 SNPs only explained ~1% of the variance observed in the resident

African dataset. Limited discrimination was seen between cases and controls in Figure 4.15A. The Area Under the Curve (AUC) for this model is 0.56 (Figure 4.15B).

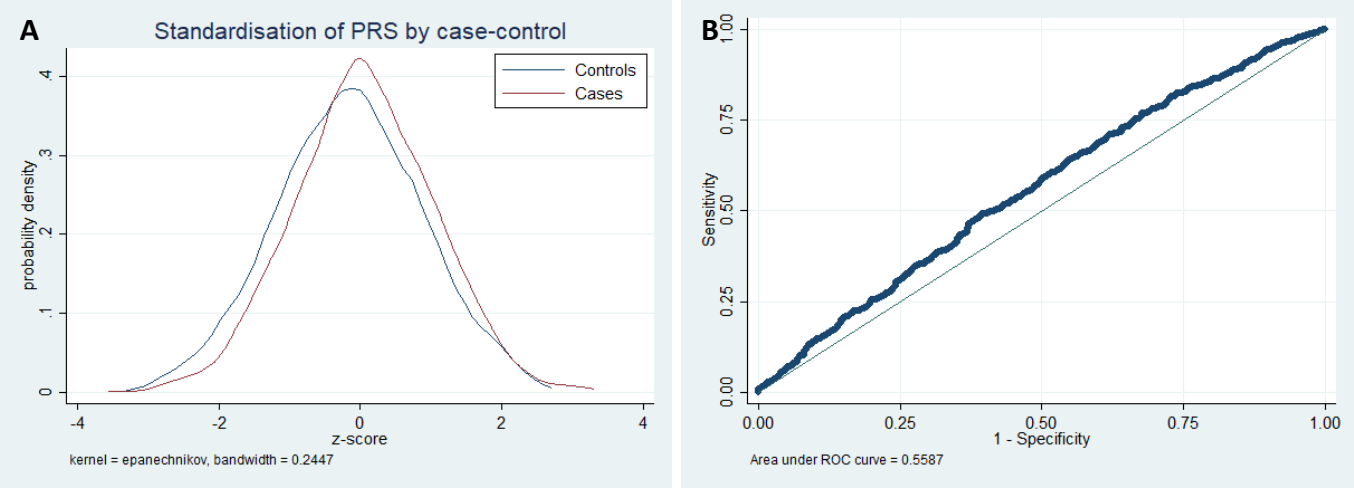


Figure 0.15: Polygenic Risk Score of 313 SNP model applied to African breast cancer dataset.
A. Standardisation of polygenic risk score curve of cases and controls. B. AUC of 313 SNP model applied to the African breast cancer dataset.

I also attempted to assess the performance of another polygenic risk score model described by Mavaddat *et al*., (2019) using 3,820 SNPs. This model was reported to have the best predictability for breast cancer. Only 2,157 SNPs (of the 3,820 SNPs) were present (imputed and/or genotyped) in the African dataset, which was used to assess this polygenic risk score model in the African GWAS. The 3,820 SNP model also showed limited discrimination between cases and controls (Figure 4.16A) and only explained 0.9% of the variance in the African dataset with an AUC of 0.55 (Figure 4.16B).

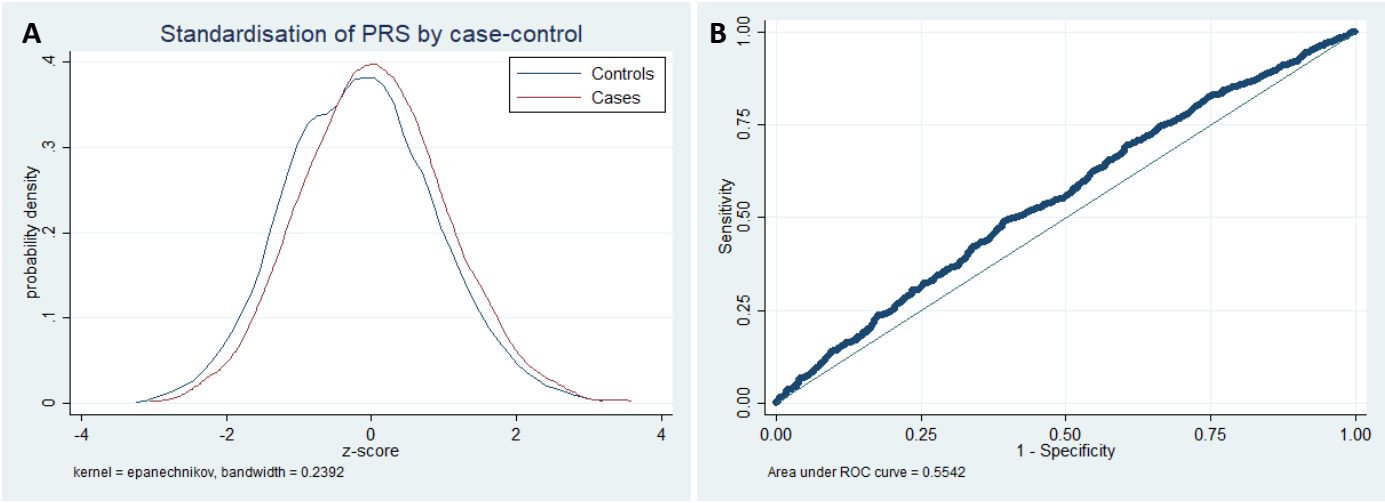


Figure 0.16: Polygenic Risk Score of 3,820 SNP model applied to African breast cancer dataset.
A. Standardisation of polygenic risk score curve of cases and controls. B. AUC of 0.55 from the 3,820 SNP model applied to African breast cancer dataset.

I also attempted to construct PRS models using the summary statistics from the BCAC African American GWAS and European GWAS, as described in the Methods (chapter 3.10). The African American GWAS-derived polygenic risk score (Figure 4.17) was less predictive than the European GWAS-derived polygenic risk score (Figure 4.18), while both performed poorly in case-control stratification. The African American GWAS and European GWAS-derived PRS explained 0.2% and 0.4% of the variance, respectively. The AUC for the African American polygenic risk score was 0.527, and the European polygenic risk score was slightly more predictive with an AUC of 0.533.

These analyses illustrate that PRS for breast cancer derived from other populations have limited application in a continental African population.

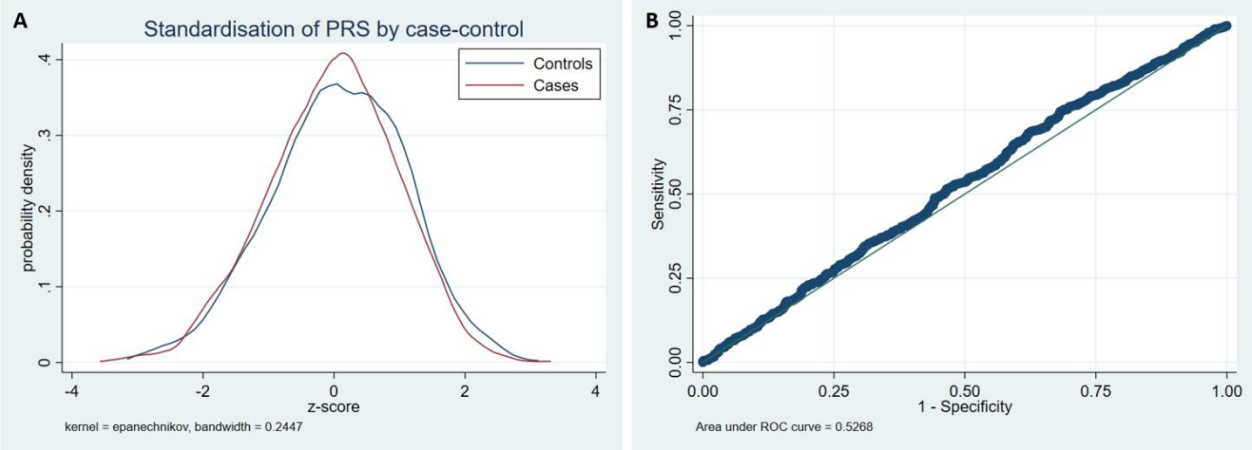


Figure 0.17: Polygenic risk score from the African American BCAC summary statistics applied to the African breast cancer dataset.
There were 12,187,877 SNPs included in the input, but only 31,479 SNPs were used in the risk score generation. A. Standardisation of polygenic risk score curve of cases and controls. B. AUC of 0.527 from the African American BCAC polygenic risk score was applied to the African breast cancer dataset.

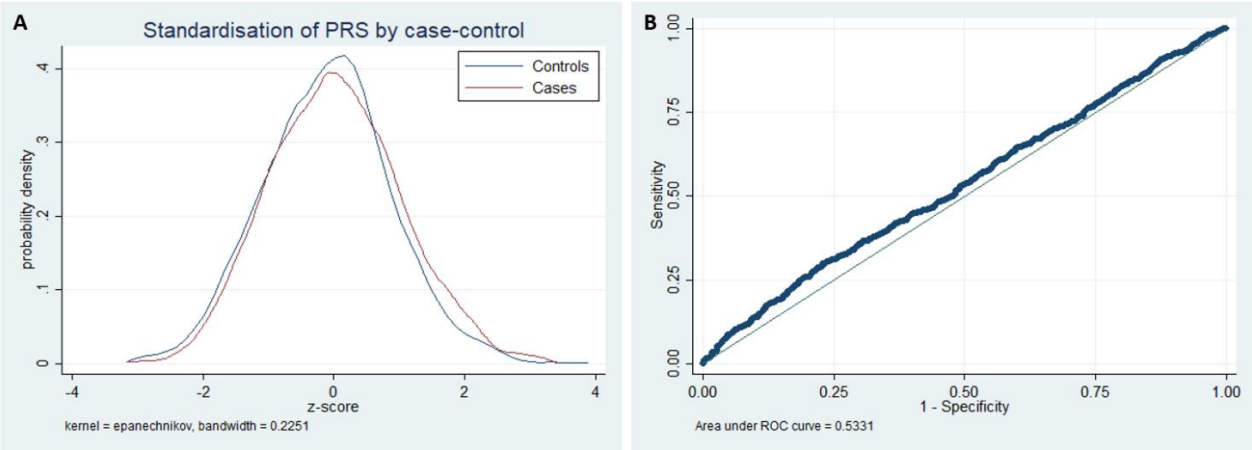


Figure 0.18: Polygenic risk score generated from the European BCAC summary statistics applied to the African breast cancer dataset.
There were 3,974,083 SNPs included in the input, but only 10,043 SNPs were used in the risk score generation. A. Standardisation of polygenic risk score curve of cases and controls. B. AUC of 0.533 from the European BCAC polygenic risk score applied to the African breast cancer dataset.

Chapter 5: Discussion

Genetic risk loci for breast cancer are well described for European populations in which most GWAS have been carried out (Martin *et al.*, 2019). However, as current literature stands, no GWAS has been done in a resident sub-Saharan African population for breast cancer (Hayat *et al.*, 2021; Ezeome *et al.*, 2022). Thus the genetic aetiology of breast cancer in resident African populations, particularly for South African black women, is restricted largely to the role of rare variants of high penetrance. This study reports the findings of the first breast cancer GWAS in a resident sub-Saharan African population. The GWAS included 2,573 breast cancer cases and 744 population controls, enrolled from the Soweto area in Johannesburg, South Africa.

5.1. Study design

This study used a population control approach by sampling controls from existing genomic resources. There are certain caveats with the use of population controls. The selection of the controls to match the origins and ethnicity of the cases as closely as possible is important to minimise the effects of population structure. In addition, genotyping will likely be performed in different centres and different batches, so assessing the data for possible batch effects is also important. The benefit of this approach is that it required substantially fewer resources to carry out a GWAS, without the need to enrol and genotype controls. The ERICA-SA study leveraged this approach by genotyping approximately 6,000 cancer cases for three common African cancers, namely, breast cancer, cervical cancer and oesophageal cancer, and using participants from the AWI-Gen study of cardiometabolic traits as population controls. This was made possible by the fact that one of the major recruitment sites for the AWI-Gen study was Soweto in greater Johannesburg, which is the same region where cases were recruited for the Johannesburg Cancer Study. Also, the AWI-Gen participants were Black South Africans and were genotyped on the same H3A GWAS array as the cases in this study. Therefore, this study was able to leverage limited resources to maximize study power in a resource-constrained funding environment.

The H3Africa GWAS array that was used in this study was specifically enriched for common variants in resident African populations. Even though LD blocks are smaller in African populations and more tag SNPs are needed to cover the genome completely, this approach is an improvement over using arrays designed for non-African populations. It would be useful to focus future efforts to developing improved GWAS arrays for African studies as well as developing sets of SNP markers that specific for different African regions due to regional variation (Choudhury *et al.*, 2020).

In this study, principal component analysis through Eigen-decomposition and admixture analysis to identify and correct the complex population substructures and admixtures present between and within the cases and controls was used. The batch effect was also assessed by iteratively testing randomly sampled subsets of samples from each batch, both within batches and across batches, against each other and identified sex differences in the population controls that were enrolled from the same geographic area, Soweto, in this case. Once I was able to adequately select the controls, as well as correct for the presence of complex population substructures, a linear mixed model regression was used which included a kinship relatedness matrix to further account for any cryptic relatedness within this study population. The combination of these approaches allowed for the removal of potentially confounding issues which arose during the conduct of this breast cancer GWAS.

Given the complexity of these populations, and the lack of GWAS data therefrom, imputation is carried out with reference panels that are not optimal for these populations. Although populations with Niger-Congo (East Africa) ancestry perform well and non-Niger-Congo populations do not, the AGR reference panel objectively performs best (Sengupta *et al.*, 2023). While the study by Sengupta *et al.,* (2023) was not published at the time of the analysis of this study, it does show and confirm that the reference panel used in this study was the best from what is currently available.

5.2. Admixture analysis, population substructure and batch effects

The South African black populations harbour complex genetic architecture and population substructure, particularly in Soweto (Choudhury *et al.*, 2020; Sengupta *et al.*, 2021). In the last century in South Africa, substantial migrations of major ethnolinguistic groups have increased the degree of admixture in different population groups (Sengupta *et al.*, 2021). We took three steps to first identify the complexity and thereafter correct for the presence of complex population substructures within this study population.

5.2.1. Removal of male controls from the analysis

Male controls have been included in GWAS when analysing female-only diseases since they can be regarded as un-phenotyped controls and provide additional data and power to determine the frequency of variants in the population being studied. However, as described in chapter 4, the inclusion of male controls from the Soweto AWI-Gen study in this breast cancer GWAS produced some highly significant associations of isolated SNPs with breast cancer.

Further analysis led to the discovery of general inflation of p-values when comparing male and female controls from Soweto (Figure 4.3). It is possible that these sex differences may be attributed, at least in part, to economic-driven migration patterns over the past 100 years in South Africa. During the gold mining era, black men migrated to Johannesburg from all over South Africa to find employment (Pirie and da Silva., 1986; May *et al.*, 2013). Populations within South Africa harbour a large degree of genetic diversity due to admixture with the San and Khoe populations, and further migration of men to Soweto from other regions of South Africa added to the already complex genetic architecture of the residents of Soweto (Sliwa *et al.*, 2008; May *et al.*, 2013). Neither admixture analysis nor PCs detected major differences between the males and females because they are in general genetically similar. It seems probable that a relatively small number of SNPs varied in frequency between males and females because of different degrees of admixture such as Khoe-San admixture from male migrations, and these resulted in false positive signals in the association analysis.

Since the use of admixture analysis and principal components steps did not adequately correct for these differences, the male controls were removed from the dataset entirely, leaving only female population controls in the analysis.

5.2.2. Removal of the JCS female controls from the analysis

After removing males from the study, the association analysis was remodelled but multiple isolated SNPs showing association with breast cancer at genome-wide significance were still observed. This suggested the presence of additional factors that were causing genomic inflation. We carried out iterative testing of randomly selected samples both across and within genotype batches. We also tested female controls between the AWI-Gen and the JCS batches. We found that one batch of the JCS female controls performed substantially differently from the remaining nine genotyping batches (Figure 4.4). This JCS batch had a log fold higher fluorescence intensity, resulting in the genotypes for this batch being incorrectly called when using the predefined cluster positions described by the Illumina genotype cluster file. The genotypes were called incorrectly because the cluster file used to call the JCS genotypes was not generated from the JCS data, and appeared to show genetic differences between cases and controls. The cause of the different fluorescent intensities in this batch is unclear, but batch effects can arise through different batches of SNP chips and reagents (Luo *et al.*, 2010).

As a matter of interest and thoroughness, these differences were confirmed by comparing the cervical cancer association analysis data to the breast cancer association analysis data. The cervical cancer GWAS used the same set of female controls as the breast cancer data. A substantial overlap was observed of top hits across both GWAS, which is not plausible based on the GWAS of these cancers in other populations and differences in their biology and epidemiology.

Therefore, the JCS control samples were removed from the analysis and remodelled the association analysis. This resulted in a more credible association result as discussed in section 5.4 below and justified the substantial loss of power resulting from the removal of the male and JCS female controls. A post-hoc power analysis using Quanto (Gauderman, 2002) showed that the study was 34% and 52% powered to detect the top hits on chromosomes 2 and 16, respectively.

5.2.3. Removing individuals with East Asian and European ancestry

As described in chapter 4 (4.2.4), admixture analysis of the remaining cases and controls showed that a small but significant minority of individuals had >5% East Asian or European ancestry. We removed 205 participants (5.8%), leaving a final dataset consisting of 2,573 breast cancer female cases from the JCS and 744 female population controls from the AWI-Gen study.

As an extra measure of quality control post-GWAS, the top SNPs were evaluated after the association analysis and assessed cluster plots of genotyped SNPs to confirm that three distinct genotype clusters were present. If an SNP was imputed, I assessed the genotype cluster plot of the closest genotyped SNP or the genotyped SNP in LD with the imputed SNP. During this process, a lone SNP on chromosome 20 was identified with poorly formed clusters on the genotype cluster plot (Figure 4.12), which also did not have support from surrounding SNPs. This SNP was therefore removed from this analyses and classified as a genotyping artefact.

5.3. Heritability of African breast cancer

Heritability is the proportion of variation that can be explained by genetic factors. SNP-based heritability (h^2_g) measures the proportion of a phenotype that is explained by only the additive effects of a set of SNPs, be they genotyped, imputed or sequenced (Yang *et al.*, 2017; Zhu and Zhou, 2020). Broad-sense heritability (H^2), on the other hand, is the proportion of phenotypic variance explained by total genetic variance (Visscher *et al.*, 2008). Although h^2_g has usually

been more appropriate for continuous traits, h^2_g can adequately be reported for binary traits on a liability scale, particularly by tools like LDAK (Speed *et al.*, 2017; Barry *et al.*, 2023).

In this study, the genomic relatedness matrix (GRM) restricted maximum likelihood (GREML) approach was implemented (Yang *et al.*, 2010) in tools such as GCTA and LDAK (Yang *et al.*, 2011; Speed *et al.*, 2020) to estimate h^2_g. The GREML approach works best for linear traits but it can be adapted for binary traits, provided certain conditions have been met. To estimate h^2_g for a linear trait, instead of taking only SNPs that reach a certain significance level, all SNPs, called predictors, are taken into account. Individuals are excluded based on relatedness described by the GRM and SNPs are excluded based on linkage-disequilibrium. Thereafter, a linear model is fitted to the linear phenotype data. The variance explained by the SNPs is then estimated (Yang *et al.*, 2010).

While the GREML approach works well for linear traits, a method was required that would estimate h^2_g in a case-control dataset. In case-control studies, cases could be overrepresented which would inflate the h^2_g value due to ascertainment bias. We therefore subsetted the cases to equal the number of controls in the h^2_g calculation. The h^2_g for binary traits is reported on the liability scale (Yang *et al.*, 2017). This model assumes that there exists a normal liability to disease, but those who are affected have liabilities above the threshold, thereby essentially accounting for the inflationary effects that oversampling of cases would have on the h^2_g estimation. SNPs in partial or complete linkage-disequilibrium could carry different weights or contribute different heritability to the phenotype when compared to LD-independent SNPs, thus weighting of the SNPs or predictors is required to obtain a more accurate h^2_g. The h^2_g was calculated using the GREML model implemented in LDAK and scaled the estimate onto a liability scale using the breast cancer incidence in South Africa from Globocan 2020 (Sung *et al.*, 2021) as a proxy for disease prevalence. This GREML implementation in LDAK takes into account the minor allele frequency and the LD of SNPs for the weight calculation to select a suitable set of SNPs for the GRM construction and h^2 estimation (Speed *et al.*, 2020).

The estimated SNP-based heritability for breast cancer in this dataset was 36% (SD=13%). The number of predictors or SNPs used for the construction of the GRM may impact the percentage of variance explained. The greater the number of SNPs, the greater the percentage of variance explained. Therefore, h^2_g comparisons across studies are not as straightforward. A Nordic twin study reported a heritability estimate of 31% for breast cancer (Möller *et al.*, 2016), although, the heritability reported here is H^2 and is usually higher than the reported h^2_g. However, estimates of SNP-based heritability for breast cancer are much lower in other studies. Sampson *et al.*, (2015), using a large GWAS dataset of multiple cancers from the US National Cancer Institute, estimated h^2_g for ER-negative breast cancer as 0.096 or 9.6%. Similarly, Jiang *et al.*, (2019) used a large GWAS dataset from populations of European ancestry to estimate h^2_g for breast cancer as 0.13 or 13%. While heritability may vary among different populations, the initial estimate requires confirmation in a much larger sample size and other African populations.

5.4. Two loci associated with African breast cancer

This breast cancer GWAS identified three signals from two loci that were significantly associated at genome-wide significance (p-value<$5x10^{-08}$) with breast cancer risk in a South African black population, including two SNPs on chromosome 2 (rs77422433 and rs112410019) and one SNP on chromosome 16 (rs3859109). These three variants also passed 700 million adaptive permutation tests. None of these variants has been previously reported to be associated with the risk of breast cancer. There were an additional 63 SNPs (Table 4.2) from 47 independent loci associated with breast cancer risk with suggestive significant p-values (p-value<$5x10^{-06}$).

The top hit, SNP rs77422433 located on chromosome 2q35, lies within the *XRCC5* gene. This gene encodes the 80kDa subunit of ATP-Dependent DNA Helicase II (KU80), which has a key role in maintaining the stability of chromosomes, by repairing double-strand DNA breaks by non-homologous end-joining (Thacker and Zdzienicka, 2004).

The two statistically significant variants associated with breast cancer have very low frequencies in non-African populations (allele frequency<0.000), and therefore they could be African-specific risk alleles. The ORs for the two top variants (rs77422433 OR(95%CI)=0.46(0.40-0.52), rs112410019 OR(95%CI)=0.47(0.41-0.53) are below one for the effect allele indicating a protective effect against breast cancer. The SNP (rs77422433) is monomorphic in European populations, but has a frequency of about 5% even in this dataset, so it will be very interesting to see if it replicates in future resident African breast cancer studies.

As *XRCC5* is involved in DNA repair, some studies have attempted to link it to breast cancer development. There are several reports that SNPs within or close to *XRCC5* are associated with breast cancer, although none are at genome-wide significance. The SNP rs828907, located 2kb upstream of *XRCC5*, was first identified in a study of 1,272 cases and controls in Taiwan as associated with breast cancer with a p-value of $5.05x10^{-07}$ (Wang *et al.*, 2009). A meta-analysis (Zhou *et al.*, 2012) which evaluated seven case-control studies found only the same SNP, rs828907, showing association with an increased risk of breast cancer (p-value<$1x10^{-05}$, OR(95%CI)=1.59(1.34-1.89)). An East Asian study found another SNP in *XRCC5*, rs16855458, that was associated with an increased risk of breast cancer with p-value=0.003 (Yu *et al.*, 2021) in 1,039 cases and 1,040 controls in China. In a study of 172 European women with breast cancer and 106 controls, another SNP in *XRCC5*, rs3835 (OR(95%CI)=1.32(0.29-5.96) was also found to be associated with an increased risk of breast cancer, albeit with non-significant statistical support (Willems *et al.*, 2008). Taken together, these and the findings suggest that further investigation of the potential role of this gene in breast cancer is warranted.

The second lead SNP, rs3859109, is on chromosome 16 and is located in the *CES5A* gene. This gene encodes a member of the carboxylesterase family, which is involved in the detoxification and tolerance of xenobiotics and insecticides (Uno *et al.*, 2014). An SNP in this gene, rs3859104, has been associated with toxicity in Platinum-based chemotherapy in lung cancer patients (Cui *et al.*, 2017). However, this gene has no previous reports of association with cancer risk. The Reactome database does not have any information on the pathway of this gene in humans. The

effect allele at this SNP has an OR of less than 1, suggesting that this allele has a protective effect against breast cancer.

5.5. Suggestive significant hits

An additional 63 SNPs had a suggestive significant association with breast cancer (p-value<$5x10^{-06}$). Some of these SNPs have a MAF<5% and are associated with relatively large effect sizes and a reduced risk of breast cancer. This phenomenon has been reported in, for example, a study of multiple traits in African populations which led to the discovery of relatively rare variants (MAF<5%) with ORs as high as 1.3 (Gurdasani *et al.*, 2019). None of the associated variants in our study is predicted to be pathogenic and most are non-coding, but they may affect gene function if they are located in regulatory regions of genes that promote tumorigenesis and affect functional elements (Khurana *et al.*, 2016; Wells *et al.*, 2019) . A significantly larger dataset from African populations would be needed to detect association of rare alleles with smaller effect sizes.

Additional SNPs that did not reach genome-wide significance, but were in strong LD with the lead SNPs on chromosome 2 occurred in *MREG* (rs145253178, p-value=$1.29x10^{-06}$, OR(95%CI)=0.50(0.44-0.57)) and *PECR* (rs145526842, p-value=$2.94x10^{-07}$), OR(95%CI)=0.48(0.42-0.55)). *MREG* is a gene with limited functional information, but it appears that it is involved in lysosome and phagosome activation and clearing, respectively (Damek-Poprawa *et al.*, 2009). While it has no connection to breast cancer, loss of gene function has been reported to drive the development of thyroid cancer (Meng *et al.*, 2017). According to Reactome (reactome.org), *PECR* is involved in the catalysis of NADPH. A recent study has found that excess NADPH can drive proliferation in cancer cells (Rather *et al.*, 2021). Although *MREG* and *PECR* do not have a strong link to breast cancer, some downstream processes could support a weak association with cancer risk in general.

There are other loci of potential interest, with p-values<10^{-6} and which are supported by additional SNPs in the region. A larger dataset of SSA ancestry may serve to strengthen some of these signals. In particular, SNPs on chromosomes 12, 18 and 10 show promise. The SNP with a suggestive p-value on chromosome 12, rs116065203 (p-value=1.56x10^{-07}, OR(95%CI)=0.30(0.19-0.40) lies between two genes, *DPPA3* and *CLEC4C*. Despite having no connection to breast cancer, the gene *DPPA3* has been associated with germ cell tumours (Bowles *et al.*, 2003; Killian *et al.*, 2016). The gene *CLEC4C* translates to a cell surface protein that is usually expressed on plasmacytoid dendritic cells (Dzionek *et al.*, 2001), and *CLEC4C* expression is used as a therapeutic target in liver cancer (Zhang *et al.*, 2021). Another locus of interest is on chromosome 18; the SNP rs865967246 (p-value=4.26x10^{-07}, OR(95%CI)= 0.36(0.27-0.45)) is intergenic, located close to *RNF152*. Mutations in *RNF152* are not known to be associated with breast cancer, but some tumour studies link somatic mutations and altered expression of *RNF152* to colon cancer (Xinhua *et al.*, 2021) and colorectal cancer (Cui *et al.*, 2018). Finally, an intergenic SNP on chromosome 10 (rs188272214, p-value=7.52x10^{-07}, OR(95%CI)= 0.43(0.36-0.51)) is located between an uncharacterized locus, LOC101928272, and long non-protein coding RNA, LINC02663. This locus has not been reported to be associated with breast cancer, or any other cancer.

5.6. Functional analysis

We submitted all SNPs with p-value<1x10^{-05} to the FUMA tool for eQTL analysis. eQTL mapping allows the identification of SNPs that affect gene expression. A non-coding variant associated with a disease could seemingly have no functional effect, but could be associated with altered expression of a gene which in turn could have a functional effect. Through this analysis, eQTLs that are as far as 1 Mb away from a gene can be identified to affect the phenotype. In addition, it can identify SNPs and loci distinctly different to those identified previously through GWAS (Gao *et al.*, 2023). The genes closest to the genome-wide significant SNPs, *XRCC5* and *CES5A*, are expressed in breast tissue. However, none of the top SNPs was associated with known eQTLs. This was not unexpected for the *XRCC5* SNPs since they are very rare or absent in the non-African population and are therefore absent from eQTL databases. This underlines an important

problem for bioinformatic analysis of African-specific SNPs. A new initiative is needed for genotyping and expression analysis of cell and tissue samples from African populations to address this issue.

While most of the variants in the GWAS were intronic or intergenic, eQTL analysis in FUMA revealed strong eQTLs in breast tissue for several genes, including *BTN3A2, MALT1* and *ZNF322*. The expression of *BTN3A2* was reported to be a prognostic marker for triple-negative breast cancer (Cai *et al.*, 2020), and showed a modest association with breast cancer in this GWAS (p-value=5.82×10^{-4}). *MALT1* promotes metastatic behaviour in breast cancer cells and an epithelial-to-mesenchymal transition program that is characteristic of triple-negative breast cancer (McAuley *et al.*, 2019; Lee *et al.*, 2022). An SNP on chromosome 6 (rs932626), which showed a suggestive association with breast cancer in this GWAS (p-value=1.84×10^{-6}) is a very strong eQTL for the *ZNF322* gene. This gene is a transcription factor, and although it has no association with breast cancer, it promotes gastric cancer cell proliferation (Hu *et al.*, 2018) and tumourigenesis in lung cancer (Liao *et al.*, 2017; Cheung *et al.*, 2020).

The SNPs of interest were also analysed in SNPnexus for functional annotation, and this revealed that none of the genome-wide significant or suggestive significant SNPs had predicted deleterious effects or was previously associated with breast cancer. A missense SNP in the *CEBPZ* gene is possibly damaging and showed a modest association with breast cancer in this GWAS (p-value=2.14×10^{-4}). This gene encodes a transcription factor and is amplified and overexpressed in oesophageal squamous cell carcinoma (Huang *et al.*, 2020).

5.7. Replication in the African breast cancer GWAS

A two-way replication test was done in which I assessed whether known breast cancer risk loci identified in the BCAC European and African American studies were replicated in this GWAS. In addition, the top hits from this African breast cancer study were investigated in the BCAC European and African American summary statistics to test for replication in other populations.

No SNPs with p-value<$5x10^{-6}$ in the African American BCAC GWAS dataset replicated in this African breast cancer GWAS at p-value<0.01. When assessing the European GWAS loci from BCAC in the African breast cancer data, there was modest replication of two loci on chromosome 10 (rs719338) and chromosome 3 (rs2322830), at p-value<0.005. The variant on chromosome 10 falls within the *ZMIZ1* gene, which is a locus that has been associated with breast cancer previously (Turnbull *et al.*, 2010). The gene regulates the activity of transcription factors, including p53 (Lee *et al.*, 2007), and may be a therapeutic target (Godicelj *et al.*, 2019). The second replicated variant on chromosome 3 falls within the *ITPR1* gene and has several links to breast cancer particularly in autophagy of cells (Soghoian *et al.*, 2005; Gu *et al.*, 2016; Han *et al.*, 2020; Parys *et al.*, 2021). It may also be a possible drug target for breast cancer therapy (Han *et al.*, 2022). The effect sizes for these two SNPs are relatively small in the European BCAC data, so it is possible that these SNPs may show a stronger association in a larger-sized African GWAS.

While some hits from the European BCAC GWAS were replicated in this study, none of the hits from the African breast cancer GWAS was replicated in the European BCAC GWAS. This is not surprising in that the majority of the top hits had relatively low frequencies in this resident African population and were extremely rare or absent in Europeans. However, there were a number of relatively common SNPs with a suggestive association in the African dataset that did not show an association in the very large, well-powered BCAC dataset. This could be explained by them being chance findings, or that they have much smaller effect sizes in European populations and the absence of relevant gene x environmental interactions. There was, however, modest replication for one African breast cancer GWAS hit in the African American BCAC GWAS on chromosome 3. This SNP is extremely rare or absent in non-African populations.

In summary, there are several reasons for the very limited extent of replication that was observed in these studies. An important limitation was the modest sample sizes of both the African and the BCAC African American studies in the context of their ability to detect the small effect sizes characteristic of common risk variants in European breast cancer. There is no

obvious reason to expect that effect sizes in African populations will be larger, so a major expansion of sample sizes in African studies is needed to detect both replication and novel associations. Other contributing factors include differences in allele frequencies in different populations, and also the lower degree of linkage disequilibrium in African genomes which will result in SNPs tagging the causal variants less effectively in African studies.

5.8. Meta-analysis

A meta-analysis combines data from different studies to increase the power to detect associations with a particular phenotype (Michailidou, 2018). The meta-analysis of the African and BCAC African American GWAS datasets included a total of 3,689 cases and 1,676 controls but did not find any association signals at genome-wide significance. However, three SNPs located on chromosomes 3, 15 and 18 reached p-values of $<1x10^{-6}$. The top SNP in the meta-analysis lies on chromosome 3: rs139299680, p-value$_{meta}$=$7.25x10^{-08}$. The SNP is intergenic and is located 4.9kb upstream of the closest gene, *RASA2*. This encodes a RAS p21 protein activator and its functions include control of cellular proliferation. There is one published study that found an SNP (rs16851483) in *RASA2* associated with increased breast cancer risk in 818 cases and 935 controls in China (Zhu *et al.*, 2017). *RASA2* has been identified as a target to treat cancer, as a study has found that cells deficient in RASA2 improve a cell's sensitivity to antigens, and as a result, cell death (Carnevale *et al.*, 2022). This SNP is not supported by surrounding SNPs, as seen on the regional plot (Figure 4.14), but is located at a recombination hotspot.

The second top SNP in the meta-analysis is rs6566983, p-value$_{meta}$=$4.98x10^{-07}$ and it lies in chromosome 18 within the intronic region of the *ALPK2* gene. Loss of function of this gene has been shown to prevent cancer progression in ovarian and renal cancer (Jiang *et al.*, 2020; Zhu *et al.*, 2020). The third SNP from the meta-analysis is on chromosome 15. The SNP rs4775326 had p-value$_{meta}$=$9.84x10^{-07}$ and is found in or near *RORA*. This gene is a member of the NR1 subfamily of nuclear hormone receptors. Mutations in *RORA* are implicated in at least two studies about the disruption of the circadian rhythm and the development of breast cancer (Truong *et al.*,

2014; Mocellin *et al.*, 2018). Furthermore, a small Iranian study with 122 breast cancer cases and 200 controls found two variants, rs11639084 and rs4774388, associated with breast cancer.

Although a meta-analysis may increase the power by combining studies, the sample sizes of both the African American GWAS dataset and the African breast cancer GWAS dataset were relatively small, and therefore not well-powered to detect risk loci with small effects. In addition, African Americans are largely descended from West African populations (Tishkoff *et al.*, 2009), and genetic differences between West African populations and South African populations are substantial, therefore some common risk variants may not be shared between the two populations (Kamiza *et al.*, 2022). This further supports the argument for a large expansion in the sample sizes of these two groups.

5.9. Polygenic Risk Scores

Several described PRS developed for breast cancers in European populations were tested in the resident African population data from this study, with discrimination between cases and controls measured with Receiver Operating Characteristic Curves (ROC). An area under the curve (AUC) ranges from 0.5 to 1, with an AUC of 0.5 indicating that the PRS does not provide any discrimination between cases and controls (Hanley and McNeil, 1982; Song *et al.*, 2019). The AUC values in the PRSs applied to the African GWAS from this study ranged from 0.527 – 0.56, depending on which data sets were used for the test. Thus all the polygenic risk score models tested demonstrated poor transferability to the African breast cancer dataset and very low predictive accuracy for breast cancer, whether derived from small or large numbers of SNPs. This is likely because the polygenic risk scores were generated in non-African or a small African American dataset, and further supports the evidence that risk scores generated in non-African populations are not predictive in resident African populations because of the vast diversity present or the presence of different genetic aetiologies (Martin *et al.*, 2017, 2019; Duncan *et al.*, 2019a). We could not access any resident African population genotype datasets to develop and validate an African-specific PRS score. While including African American data in

PRS training sets may improve risk prediction in resident African populations, it does not encompass all the variation present in sub-Saharan Africa.

A recent study of genetic risk scores (GRS) for lipid traits found that prediction of GRS derived from African American data varied greatly within sub-Saharan Africa between South African Zulu and Ugandan cohorts (Kamiza *et al.*, 2022). Although a study by (Chikowore *et al.*, (2022) found that an African American PRS enhanced prediction of type 2 diabetes in continental Africans, this may be related to the very large sample size in the study and that the performance of PRSs can differ substantially across different disease phenotypes. It is important to not make the generalisation that PRSs derived from an African American population will transfer well to all resident African populations and different phenotypes. Several studies show that PRSs need to be generated using SNPs from diverse or resident African populations are predictive in non-European populations (Martin *et al.*, 2019; Cavazos and Witte, 2021; Ruan *et al.*, 2022; Adam *et al.*, 2023).

In order to develop risk scores that will be appropriate for populations in sub-Saharan Africa, more GWAS need to be carried out with larger sample sizes in the different areas of sub-Saharan Africa. Larger sample sizes will allow appropriate-sized training sets to apply to test in relevant populations. Including more participants from all regions in sub-Saharan Africa will encompass the vast genomic diversity in sub-Saharan Africa and will result in the generation of more predictive PRS, as including diverse populations in polygenic risk score generation can improve the transferability of risk loci and PRSs across different populations (Cavazos and Witte, 2021).

6. Conclusion

This study used a GWAS to identify two loci significantly associated with the risk of breast cancer in a South African black population. The variants on chromosome 2 are located in the

XRCC5 gene and are likely to be African-specific breast cancer risk variants because they have extremely low frequencies in non-African populations. Preliminary evidence for the association of SNPs in *XRCC5* has been published in two previous reports, but this study places this finding on a much firmer footing, and the African variants have not been previously associated with breast cancer in African populations. The finding of the association of multiple SNPs in the *CES5A* gene appears to be novel, but both the genome-wide significant findings need to be replicated in additional case/control samples from South Africa and elsewhere. The other limitation of this study was the relatively small sample size, and thus modest power to detect risk variants with small effects. The small sample size was exacerbated by the need to remove subsets of controls, but this was done to retain the integrity and quality of the results of the GWAS. Finally, breast cancer hormone receptor status was not available from these historical breast cancer cases and could therefore not be included in the analysis. It is acknowledged that including this information can be insightful and reveal receptor specific variants, as is seen in studies by Huo *et al.*, (2016) and Milne *et al.*, (2017). Strengths of this study were the use of the H3A GWAS array, designed by African scientists for genetic studies in African populations, and the recruitment of black South African cases and controls from the same city in South Africa.

Doing a GWAS in an African population was even more challenging than I had anticipated. South African black populations in Soweto, specifically, harbour great genetic variation, with the admixture of multiple ethnolinguistic groups. Not only did this genetic diversity have to be accounted for, but we also encountered the challenges associated with the effect of historical immigration within South Africa which appears to be reflected in the genetic differences between males and females in Soweto. An improvement for the design of future studies would be to have all the samples from both cases and controls genotyped in one place, however this is not possible when data from different studies are combined. This would reduce the risk of technical variations occurring between different centres. Further improvements could also be made in the imputation process going forward. Even though imputation servers have African reference populations available, they do not include enough individuals in those reference populations, and the individuals are not sampled from enough sites across sub-Saharan Africa. While the H3A array was used in this study due to its incorporation of a large number of African

specific SNPs, it is not sufficient in capturing the vast genomic diversity across all of sub-Saharan Africa. Studies such as ours would also benefit from the design of more population-specific genotyping arrays at an affordable price which reflected the diversity present in the different populations of Africa.

The science of polygenic risk scores is advancing rapidly, with the prospect of introducing analysis of common risk variants for breast cancer and other disorders into population screening programs. This study has added to the emerging data that polygenic risk scores generated in non-African populations do not offer effective predictability of disease risk if applied to an African population. This highlights the need to generate sufficient genome-wide genetic data in the populations of sub-Saharan Africa to enable the design of useful PRS in multiple different African populations.

This is the first breast cancer GWAS to be conducted in a resident black African population and has laid the groundwork for more promising work to come. Future work in populations of African ancestry includes the Breast Cancer Genetic Study in African-Ancestry Populations which includes investigators from the African-American Breast Cancer Consortium, the African-American Breast Cancer Epidemiology and Risk (AMBER) Consortium, and the NCI Cohort Consortium. In addition, our research group has joined the very large global project on the genetics of breast cancer called Confluence, which aims to do a GWAS on 300,000 breast cancer cases and controls from different ethnicities to identify risk factors in diverse populations (https://dceg.cancer.gov/research/cancer-types/breast-cancer/confluence-project). This study has the potential to uncover breast cancer risk variants in understudied populations around the world. As larger studies in Africa will be carried out, the potential to discover novel risk variants with smaller effect sizes will increase. Also, fine-mapping of risk loci that are common between African and non-African populations might be possible. I look forward with excitement to seeing the genomic work in breast cancer that will come from African scientists in the future.

References

Abe, O., Abe, R., Enomoto, K., *et al.* (2011) 'Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials', *Lancet*, 378, pp. 771–784. doi: 10.1016/S0140-6736(11)60993-8.

Adam, Y., Sadeeq, S., Kumuthini, J., *et al.* (2023) 'Polygenic Risk Score in African populations: progress and challenges', *F1000Research 2023 11:175*, 11, p. 175. doi: 10.12688/f1000research.76218.2.

Adedokun, B., Zheng, Y., Ndom, P., *et al.* (2020) 'Prevalence of inherited mutations in breast cancer predisposition genes among women in Uganda and Cameroon', *Cancer Epidemiology Biomarkers and Prevention*, 29(2), pp. 359–367. doi: 10.1158/1055-9965.EPI-19-0506.

Adedokun, B., Du, Z., Gao, G., *et al.* (2021) 'Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women', *Nature Communications*, 12(4198), pp. 1–8. doi: 10.1038/s41467-021-24327-x.

Adolf, I. C., Rweyemamu, L. P., Akan, G., *et al.* (2022) 'The interplay between XPG-Asp1104His polymorphism and reproductive risk factors elevates risk of breast cancer in Tanzanian women: A multiple interaction analysis', *Cancer Medicine*, 2022, pp. 1–16. doi: 10.1002/CAM4.4914.

Agyepong, I. A., Sewankambo, N., Binagwaho, A., *et al.* (2017) 'The path to longer and healthier lives for all Africans by 2030: the Lancet Commission on the future of health in sub-Saharan Africa', *The Lancet*, 17, pp. 1–57. doi: 10.1016/S0140-6736(17)31509-X.

Ahmed, J. H., Makonnen, E., Fotoohi, A., *et al.* (2019) 'Vitamin D status and association of VDR genetic polymorphism to risk of breast cancer in Ethiopia', *Nutrients*, 11(2), pp. 1–14. doi: 10.3390/nu11020289.

Ahmed, M. and Rahman, N. (2006) 'ATM and breast cancer susceptibility', *Oncogene*, pp. 5906–5911. doi: 10.1038/sj.onc.1209873.

Ahmed, S., Thomas, G., Ghoussaini, M., *et al.* (2009) 'Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2.', *Nature Genetics*, 41(5), pp. 585–90. doi: 10.1038/ng.354.

Alexander, D. H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research*, 19, pp. 1655–1664. doi: 10.1101/GR.094052.109.

Ali, S. A., Soo, C., Agongo, G., *et al.* (2018) 'Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study', *Global Health Action*, 11(Suppl 2). doi: 10.1080/16549716.2018.1507133.

Amos, C. I., Dennis, J., Wang, Z., *et al.* (2017) 'The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers', *Cancer Epidemiology Biomarkers & Prevention*, 26(1), pp. 126–135. doi: 10.1158/1055-9965.EPI-16-0106.

Antoniou, A. C., Spurdle, A. B., Sinilnikova, O. M., *et al.* (2008) 'Common Breast Cancer-Predisposition Alleles Are Associated with Breast Cancer Risk in BRCA1 and BRCA2 Mutation Carriers', *American Journal of Human Genetics*, 82, pp. 937–948. doi: 10.1016/J.AJHG.2008.02.008.

Anyigba, C. A., Awandare, G. A. and Paemka, L. (2021) 'Breast cancer in sub-Saharan Africa: The current state and uncertain future', *Experimental Biology and Medicine*, 246(12), pp. 1377–1387. doi: 10.1177/15353702211006047.

Arksey, H. and O'Malley, L. (2005) 'Scoping studies: Towards a methodological framework', *International Journal of Social Research Methodology: Theory and Practice*, 8(1), pp. 19–32. doi: 10.1080/1364557032000119616.

Audrey, A., de Haan, L., van Vugt, M. A. T. M., *et al.* (2022) 'Processing DNA lesions during mitosis to prevent genomic instability', *Biochemical Society Transactions*, 50(4), pp. 1105–1118. doi: 10.1042/BST20220049.

Barr, A. L., Young, E. H., Smeeth, L., *et al.* (2016) 'The need for an integrated approach for chronic disease research and care in Africa', *Global Health, Epidemiology and Genomics*, 1(19), pp. 1–6. doi: 10.1017/gheg.2016.16.

Barry, C. J. S., Walker, V. M., Cheesman, R., *et al.* (2023) 'How to estimate heritability: a guide for genetic epidemiologists', *International Journal of Epidemiology*, 52(2), pp. 624–632. doi: 10.1093/IJE/DYAC224.

Barzaman, K., Moradi-Kalbolandi, S., Hosseinzadeh, A., *et al.* (2021) 'Breast cancer immunotherapy: Current and novel approaches', *International Immunopharmacology*, 98. doi: 10.1016/J.INTIMP.2021.107886.

Bedrosian, I., Hu, C. Y. and Chang, G. J. (2010) 'Population-Based Study of Contralateral Prophylactic Mastectomy and Survival Outcomes of Breast Cancer Patients', *JNCI Journal of the National Cancer Institute*, 102, pp. 401–409. doi: 10.1093/JNCI/DJQ018.

Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289–300.

Benner, C., Spencer, C. C. A., Havulinna, A. S., *et al.* (2016) 'FINEMAP: efficient variable selection using summary data from genome-wide association studies', *Bioinformatics*, 32(10), pp. 1493–1501. doi: 10.1093/BIOINFORMATICS/BTW018.

Biancolella, M., Ouédraogo, N. L. M., Zongo, N., *et al.* (2021) 'Breast cancer in West Africa: molecular analysis of BRCA genes in early-onset breast cancer patients in Burkina Faso', *Human Genomics*, 15(1), pp. 1–11. doi: 10.1186/S40246-021-00365-W/TABLES/3.

Boua, P. R., Brandenburg, J. T., Choudhury, A., *et al.* (2022) 'Genetic associations with carotid intima-media thickness link to atherosclerosis with sex-specific effects in sub-Saharan Africans', *Nature Communications*, 13(1), pp. 1–11. doi: 10.1038/s41467-022-28276-x.

Bowles, J., Teasdale, R. P., James, K., *et al.* (2003) 'Dppa3 is a marker of pluripotency and has a human homologue that is expressed in germ cell tumours', *Cytogenetic and Genome Research*, 101, pp. 261–265. doi: 10.1159/000074346.

Brahim, S. M., Zein, E. E., Bonnet, C., *et al.* (2022) 'Screening of BRCA1/2 variants in Mauritanian breast cancer patients', *BMC Cancer*, 22(1). doi: 10.1186/S12885-022-09903-8.

Brandenburg, J.-T., Clark, L., Botha, G., *et al.* (2022) 'H3AGWAS: a portable workflow for genome wide association studies', *BMC Bioinformatics 2022 23:1*, 23(1), pp. 1–15. doi: 10.1186/S12859-022-05034-W.

Bray, F., Ferlay, J., Soerjomataram, I., *et al.* (2018) 'Global cancer statistics 2018: GLOBOCAN

estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*, 68(6), pp. 394–424. doi: 10.3322/caac.21492.

Burton, P. R., Clayton, D. G., Cardon, L. R., *et al.* (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661–678. doi: 10.1038/nature05911.

Bush, W. S. and Moore, J. H. (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*. Edited by F. Lewitter and M. Kann, 8(12), p. e1002822. doi: 10.1371/journal.pcbi.1002822.

Cai, P., Lu, Z., Wu, J., *et al.* (2020) 'BTN3A2 serves as a prognostic marker and favors immune infiltration in triple-negative breast cancer', *Journal of Cellular Biochemistry*, 121(3), pp. 2643–2654. doi: 10.1002/JCB.29485.

Cameron, D., Piccart-Gebhart, M. J., Gelber, R. D., *et al.* (2017) '11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial', *Lancet*, 389(10075), pp. 1195–1205. doi: 10.1016/S0140-6736(16)32616-2.

Carnevale, J., Shifrut, E., Kale, N., *et al.* (2022) 'RASA2 ablation in T cells boosts antigen sensitivity and long-term function', *Nature 2022 609:7925*, 609(7925), pp. 174–182. doi: 10.1038/s41586-022-05126-w.

Carver, T., Hartley, S., Lee, A., *et al.* (2021) 'CanRisk Tool-A Web Interface for the Prediction of Breast and Ovarian Cancer Risk and the Likelihood of Carrying Genetic Pathogenic Variants', *Cancer Epidemiology Biomarkers & Prevention*, 30(3), pp. 469–473. doi: 10.1158/1055-9965.EPI-20-1319.

Cavazos, T. B. and Witte, J. S. (2021) 'Inclusion of variants discovered from diverse populations improves polygenic risk score transferability', *Human Genetics and Genomics Advances*, 2(1), p. 100017. doi: 10.1016/j.xhgg.2020.100017.

Chang, B. L., Spangler, E., Gallagher, S., *et al.* (2011) 'Validation of genome-wide prostate cancer associations in men of african descent', *Cancer Epidemiology Biomarkers and Prevention*, 20(1),

pp. 23–32. doi: 10.1158/1055-9965.EPI-10-0698.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., *et al.* (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience*, 4(7), pp. 1–16. doi: 10.1186/s13742-015-0047-8.

Che, R., Jack, J. R., Motsinger-Reif, A. A., *et al.* (2014) 'An adaptive permutation approach for genome-wide association study: Evaluation and recommendations for use', *BioData Mining*, 7(1), pp. 1–13. doi: 10.1186/1756-0381-7-9/FIGURES/3.

Chen, F., Chen, G. K., Millikan, R. C., *et al.* (2011) 'Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans', *Human Molecular Genetics*, 20(22), pp. 4491–4503. doi: 10.1093/hmg/ddr367.

Chen, F., Chen, G. K., Stram, D. O., *et al.* (2013) 'A genome-wide association study of breast cancer in women of African ancestry', *Human Genetics*, 132(1), pp. 39–48. doi: 10.1007/S00439-012-1214-Y.

Chen, W. C., Kerr, R., May, A., *et al.* (2018) 'The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term Storage at −30°C', *Biopreservation and Biobanking*, 16(2), pp. 106–113. doi: 10.1089/bio.2017.0050.

Chen, W. C., Bye, H., Matejcic, M., *et al.* (2019) 'Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population', *Carcinogenesis*, 40(4), pp. 513–520. doi: 10.1093/carcin/bgz026.

Chen, W. C., Singh, E., Muchengeti, M., *et al.* (2020) 'Johannesburg Cancer Study (JCS): contribution to knowledge and opportunities arising from 20 years of data collection in an African setting', *Cancer Epidemiology*, 65. doi: 10.1016/j.canep.2020.101701.

Cheung, C. H. Y., Hsu, C. L., Lin, T. Y., *et al.* (2020) 'ZNF322A-mediated protein phosphorylation induces autophagosome formation through modulation of IRS1-AKT glucose uptake and HSP-elicited UPR in lung cancer', *Journal of Biomedical Science*, 27(1). doi: 10.1186/S12929-020-00668-5.

Chikowore, T., Ekoru, K., Vujkovi, M., *et al.* (2022) 'Polygenic Prediction of Type 2 Diabetes in

Africa', *Diabetes Care*, 45(3), pp. 717–723. doi: 10.2337/DC21-0365.

Chlebowski, R. T., Anderson, G. L., Aragaki, A. K., *et al.* (2020) 'Association of Menopausal Hormone Therapy With Breast Cancer Incidence and Mortality During Long-term Follow-up of the Women's Health Initiative Randomized Clinical Trials', *JAMA*, 324(4), pp. 369–380. doi: 10.1001/JAMA.2020.9482.

Choi, S. W., Mak, T. S. H. and O'Reilly, P. F. (2020) 'A guide to performing Polygenic Risk Score analyses', *Nature Protocols*, 15(9), p. 2759. doi: 10.1038/S41596-020-0353-1.

Choi, S. W. and O'Reilly, P. F. (2019) 'PRSice-2: Polygenic Risk Score software for biobank-scale data', *GigaScience*, 8(7). doi: 10.1093/GIGASCIENCE/GIZ082.

Choudhury, A., Aron, S., Botigué, L. R., *et al.* (2020) 'High-depth African genomes inform human migration and health', *Nature*, 586(7831), pp. 741–748. doi: 10.1038/s41586-020-2859-7.

Choudhury, A., Brandenburg, J. T., Chikowore, T., *et al.* (2022) 'Meta-analysis of sub-Saharan African studies provides insights into genetic architecture of lipid traits', *Nature Communications*, 13(1), pp. 1–13. doi: 10.1038/s41467-022-30098-w.

Claus, E. B., Risch, N. J. and Thompson, W. D. (1990) 'Age at onset as an indicator of familial risk of breast cancer', *American Journal of Epidemiology*, 131(6), pp. 961–972. doi: 10.1093/OXFORDJOURNALS.AJE.A115616.

Cook, J. P., Mahajan, A. and Morris, A. P. (2016) 'Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes', *European Journal of Human Genetics 2017 25:2*, 25(2), pp. 240–245. doi: 10.1038/ejhg.2016.150.

Cui, J. J., Wang, L. Y., Zhu, T., *et al.* (2017) 'Gene-gene and gene-environment interactions influence platinum-based chemotherapy response and toxicity in non-small cell lung cancer patients', *Scientific Reports 2017 7:1*, 7(1), pp. 1–8. doi: 10.1038/s41598-017-05246-8.

Cui, X., Shen, W., Wang, G., *et al.* (2018) 'Ring finger protein 152 inhibits colorectal cancer cell growth and is a novel prognostic biomarker', *American Journal of Translational Research*, 10(11), pp. 3701–3712.

Cumber, S. N., Nchanji, K. N. and Tsoka-Gwegweni, J. M. (2017) 'Breast cancer among women in

sub-Saharan Africa: prevalence and a situational analysis', *Southern African Journal of Gynaecological Oncology*, 9(2), pp. 35–37. doi: 10.1080/20742835.2017.1391467.

Curtin, N. J. and Szabo, C. (2020) 'Poly(ADP-ribose) polymerase inhibition: past, present and future', *Nature Reviews Drug Discovery 2020 19:10*, 19(10), pp. 711–736. doi: 10.1038/s41573-020-0076-6.

Dalivandan, S. T., Plummer, J. and Gayther, S. A. (2021) 'Risks and Function of Breast Cancer Susceptibility Alleles', *Cancers*, 13(16). doi: 10.3390/CANCERS13163953.

Damek-Poprawa, M., Diemer, T., Lopes, V. S., *et al.* (2009) 'Melanoregulin (MREG) modulates lysosome function in pigment epithelial cells', *The Journal of Biological Chemistry*, 284(16), pp. 10877–10889. doi: 10.1074/JBC.M808857200.

Das, S., Abecasis, G. R. and Browning, B. L. (2018) 'Genotype Imputation from Large Reference Panels', *Annual Reviews*, 19, pp. 73–96. doi: 10.1146/annurev-genom-083117-021602.

Deng, C. and Brodie, S. (2000) 'Roles of BRCA1 and its interacting proteins', *Bioessays*, 22(8), pp. 728–737. doi: 10.1002/1521-1878(200008)22:8<728::AID-BIES6>3.0.CO;2-B.

Development Core Team (2008) 'R: A Language and Environment for Statistical Computing', *Foundation for Statistical Computing*.

Dewan, A., Liu, M., Hartman, S., *et al.* (2006) 'HTRA1 promoter polymorphism in wet age-related macular degeneration.', *Science*, 314(5801), pp. 989–992. doi: 10.1126/science.1133807.

Diakite, B., Kassogue, Y., Dolo, G., *et al.* (2020) 'Association of PIN3 16-bp duplication polymorphism of TP53 with breast cancer risk in Mali and a meta-analysis', *BMC Medical Genetics*, 21(1). doi: 10.1186/S12881-020-01072-4.

Diez, O., Pelegrí, A., Gadea, N., *et al.* (2011) 'Novel BRCA1 deleterious mutation (c.1949_1950delTA) in a woman of Senegalese descent with triple-negative early-onset breast cancer', *Oncology Letters*, 2(6), pp. 1287–1289. doi: 10.3892/ol.2011.390.

Diop, J. P. D., Diallo, R. N., Bourdon-Huguenin, V., *et al.* (2019) 'Novel BRCA2 pathogenic variant c.5219 T > G; P.(Leu1740Ter) in a consanguineous Senegalese family with hereditary breast cancer', *BMC Medical Genetics*, 20(73), pp. 1–7. doi: 10.1186/s12881-019-0814-y.

Dorling, L., Carvalho, S., Allen, J., *et al.* (2021) 'Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women', *New England Journal of Medicine*, 384(5), pp. 428–439. doi: 10.1056/NEJMoa1913948.

Dudbridge, F. (2013) 'Power and Predictive Accuracy of Polygenic Risk Scores', *PLoS Genetics*, 9(3), p. e1003348. doi: 10.1371/journal.pgen.1003348.

Duncan, L., Shen, H., Gelaye, B., *et al.* (2019a) 'Analysis of polygenic risk score usage and performance in diverse human populations', *Nature Communications*, 10(1), pp. 1–9. doi: 10.1038/s41467-019-11112-0.

Duncan, LE, Ostacher, M. and Ballon, J. (2019b) 'How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete', *Neuropsychopharmacology*, 44(9), pp. 1518–1523. doi: 10.1038/s41386-019-0389-5.

Dunn, O. J. (1961) 'Multiple Comparisons among Means', *Journal of the American Statistical Association*, 56(293), pp. 52–64. doi: 10.1080/01621459.1961.10482090.

Dzionek, A., Sohma, Y., Nagafune, J., *et al.* (2001) 'BDCA-2, a Novel Plasmacytoid Dendritic Cell–specific Type II C-type Lectin, Mediates Antigen Capture and Is a Potent Inhibitor of Interferon α/β Induction', *Journal of Experimental Medicine*, 194(12), pp. 1823–1834. doi: 10.1084/JEM.194.12.1823.

Easton, D. F., Bishop, D. T., Ford, D., *et al.* (1993) 'Genetic Linkage Analysis in Familial Breast and Ovarian Cancer: Results from 214 Families', *American Journal of Human Genetics*, 52, pp. 678–701.

Easton, D. F., Pooley, K. A., Dunning, A. M., *et al.* (2007) 'Genome-wide association study identifies novel breast cancer susceptibility loci', *Nature*, 447(7148), pp. 1087–1093. doi: 10.1038/nature05887.

Easton, D. F., Pharoah, P. D. P., Antoniou, A. C., *et al.* (2015) 'Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk', *New England Journal of Medicine*, 372(23), pp. 2243–2257. doi: 10.3390/cancers13164154.

Engin, A. (2017) 'Obesity-associated Breast Cancer: Analysis of risk factors', *Advances in*

Experimental Medicine and Biology, 960, pp. 571–606. doi: 10.1007/978-3-319-48382-5_25.

Eygelaar, D., van Rensburg, E. J. and Joubert, F. (2022) 'Germline sequence variants contributing to cancer susceptibility in South African breast cancer patients of African ancestry', *Scientific Reports*, 12(1). doi: 10.1038/S41598-022-04791-1.

Ezeome, E. R., Yawe, K. D. T., Ayandipo, O., *et al.* (2022) 'The African Female Breast Cancer Epidemiology Study Protocol', *Frontiers in Oncology*, 12. doi: 10.3389/FONC.2022.856182/PDF.

Fachal, L., Aschard, H., Beesley, J., *et al.* (2020) 'Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes', *Nature Genetics 2020 52:1*, 52(1), pp. 56–73. doi: 10.1038/s41588-019-0537-1.

Fackenthal, J. D., Zhang, J., Zhang, B., *et al.* (2012) 'High prevalence of BRCA1 and BRCA2 mutations in unselected Nigerian breast cancer patients', *International Journal of Cancer*, 131(5), pp. 1114–1123. doi: 10.1002/ijc.27326.

Feng, Y., Rhie, S. K., Huo, D., *et al.* (2017) 'Characterizing genetic susceptibility to breast cancer in women of african ancestry', *Cancer Epidemiology Biomarkers and Prevention*, 26(7), pp. 1016–1026. doi: 10.1158/1055-9965.EPI-16-0567.

Ferlay, J., Colombet, M., Soerjomataram, I., *et al.* (2019) 'Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods', *International Journal of Cancer*, 144(8), pp. 1941–1953. doi: 10.1002/IJC.31937.

Fletcher, O., Johnson, N., Orr, N., *et al.* (2011) 'Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study', *JNCI: Journal of the National Cancer Institute*, 103(5), pp. 425–435. doi: 10.1093/jnci/djq563.

Fong, P. C., Boss, D. S., Yap, T. A., *et al.* (2009) 'Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers', *New England Journal of Medicine*, 361(2), pp. 123–134. doi: 10.1056/NEJMOA0900212/SUPPL_FILE/NEJM_FONG_123SA1.PDF.

Foulkes, W. D., Smith, I. E. and Reis-Filho, J. S. (2010) 'Triple-Negative Breast Cancer', *New England Journal of Medicine*, 363, pp. 1938–1948. doi: 10.1056/NEJMRA1001389.

Francies, F. Z., Wainstein, T., De Leeneer, K., *et al.* (2015) 'BRCA1, BRCA2 and PALB2 mutations

and CHEK2 c.1100delC in different South African ethnic groups diagnosed with premenopausal and/or triple negative breast cancer', *BMC Cancer*, 15(912), pp. 1–10. doi: 10.1186/s12885-015-1913-6.

Freudenheim, J. L. (2020) 'Alcohol's Effects on Breast Cancer in Women', *Alcohol Research : Current Reviews*, 40(2), pp. 1–12. doi: 10.35946/ARCR.V40.2.11.

Gao, G., Fiorica, P. N., McClellan, J., *et al.* (2023) 'A joint transcriptome-wide association study across multiple tissues identifies candidate breast cancer susceptibility genes', *American Journal of Human Genetics*, 110(6), pp. 950–962. doi: 10.1016/J.AJHG.2023.04.005.

Gao, Q., Adebamowo, C. A., Fackenthal, J., *et al.* (2000) 'Protein truncating BRCA1 and BRCA2 mutations in African women with pre-menopausal breast cancer', *Human Genetics*, 107(2), pp. 192–194. doi: 10.1007/s004390000342.

Garcia-Closas, M., Hall, P., Nevanlinna, H., *et al.* (2008) 'Heterogeneity of Breast Cancer Associations with Five Susceptibility Loci by Clinical and Pathological Characteristics', *PLoS Genetics*, 4(4), pp. 1–10. doi: 10.1371/JOURNAL.PGEN.1000054.

Gauderman, W. J. (2002) 'Sample size requirements for association studies of gene-gene interaction', *American Journal of Epidemiology*, 155(5), pp. 478–484. doi: 10.1093/AJE/155.5.478.

Ghoussaini, M., Fletcher, O., Michailidou, K., *et al.* (2013) 'Genome-wide association analysis identifies three new breast cancer susceptibility loci', *Nature Genetics*, 44(3), pp. 312–318. doi: 10.1038/ng.1049.Genome-wide.

Godicelj, A., Blake, R., Giorgi, F. M., *et al.* (2019) 'ZMIZ1 — A novel Estrogen Receptor co-activator that enhances the growth of ER+ breast cancer', *bioRxiv*, p. 789610. doi: 10.1101/789610.

Gouda, H. N., Charlson, F., Sorsdahl, K., *et al.* (2019) 'Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: results from the Global Burden of Disease Study 2017', *The Lancet Global Health*, 7(10), pp. e1375–e1387. doi: 10.1016/S2214-109X(19)30374-2.

Gravena, A. A. F., Lopes, T. C. R., Demitto, M. de O., *et al.* (2018) 'The Obesity and the Risk of

Breast Cancer among Pre and Postmenopausal Women', *Asian Pacific Journal of Cancer Prevention*, 19(9), pp. 2429–2436. doi: 10.22034/APJCP.2018.19.9.2429.

Group on Hormonal Factors in Breast Cancer, C. (2021) 'Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence', *The Lancet*, 394(10204), pp. 1159–1168. doi: 10.1016/S0140-6736(19)31709-X.

Gu, Y., Li, P., Peng, F., *et al.* (2016) 'Autophagy-related prognostic signature for breast cancer', *Molecular Carcinogenesis*, 55(3), pp. 292–299. doi: 10.1002/MC.22278.

Gurdasani, D., Barroso, I., Zeggini, E., *et al.* (2019) 'Genomics of disease risk in globally diverse populations', *Nature Reviews Genetics 2019 20:9*, 20(9), pp. 520–535. doi: 10.1038/s41576-019-0144-0.

Habyarimana, T., Bakri, Y., Mugenzi, P., *et al.* (2018) 'Association between glutathione peroxidase 1 codon 198 variant and the occurrence of breast cancer in Rwanda', *Molecular Genetics and Genomic Medicine*, 6(2), pp. 268–275. doi: 10.1002/mgg3.367.

Habyarimana, T., Attaleb, M., Mugenzi, P., *et al.* (2018a) 'Association of p53 Codon 72 Polymorphism with Breast Cancer in a Rwandese Population', *Pathobiology*, 85(3), pp. 186–191. doi: 10.1159/000481664.

Habyarimana, T., Attaleb, M., Mugenzi, P., *et al.* (2018b) 'CHEK2 germ line mutations are lacking among familial and sporadic breast cancer patients in Rwanda', *Asian Pacific Journal of Cancer Prevention*, 19(2), pp. 375–379. doi: 10.22034/APJCP.2018.19.2.375.

Haiman, C. A., Chen, G. K., Vachon, C. M., *et al.* (2011) 'A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer', *Nature Genetics*, 43(12), pp. 1210–1214. doi: 10.1038/ng.985.

Han, B., Zhang, H., Zhu, Y., *et al.* (2020) 'Subtype-specific risk models for accurately predicting the prognosis of breast cancer using differentially expressed autophagy-related genes', *Aging*, 12(13), pp. 13318–13337. doi: 10.18632/AGING.103437.

Han, B., Zhen, F., Zheng, X. S., *et al.* (2022) 'Systematic analysis of the expression and prognostic

value of ITPR1 and correlation with tumor infiltrating immune cells in breast cancer', *BMC Cancer*, 22(1). doi: 10.1186/S12885-022-09410-W.

Hanley, J. A. and McNeil, B. J. (1982) 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, 143(1), pp. 29–36. doi: 10.1148/RADIOLOGY.143.1.7063747.

Hayat, M., Chen, W. C., Brandenburg, J.-T., *et al.* (2021) 'Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations', *JCO Global Oncology*, 7(7), pp. 1462–1471. doi: 10.1200/GO.21.00089.

Hindorff, L. A., Bonham, V. L., Brody, L. C., *et al.* (2018) 'Prioritizing diversity in human genomics research', *Nature Reviews Genetics*, 19(3), pp. 175–185. doi: 10.1038/nrg.2017.89.

Hong, Z., Xu, Q., Yan, X., *et al.* (2022) 'Analysis of Signs and Effects of Surgical Breast Cancer Patients Based on Big Data Technology', *Computational Intelligence and Neuroscience*, 2022. doi: 10.1155/2022/3373553.

Hu, X., Zhang, M., Miao, J., *et al.* (2018) 'miRNA-4317 suppresses human gastric cancer cell proliferation by targeting ZNF322', *Cell Biology International*, 42(8), pp. 923–930. doi: 10.1002/CBIN.10870.

Huang, Y., Lin, L., Shen, Z., *et al.* (2020) 'CEBPG promotes esophageal squamous cell carcinoma progression by enhancing PI3K-AKT signaling', *American Journal of Cancer Research*, 10(10), p. 3328.

Hunter, D. J., Kraft, P., Jacobs, K. B., *et al.* (2007) 'A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer', *Nature Genetics*, 39(7), pp. 870–874. doi: 10.1038/ng2075.

Huo, D., Zheng, Y., Ogundiran, T. O., *et al.* (2012) 'Evaluation of 19 susceptibility loci of breast cancer in women of African ancestry', *Carcinogenesis*, 33(4), pp. 835–840. doi: 10.1093/carcin/bgs093.

Huo, D., Feng, Y., Haddad, S., *et al.* (2016) 'Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor

negative breast cancer', *Human Molecular Genetics*, 25(21), pp. 4835–4846. doi: 10.1093/hmg/ddw305.

Jiang, J., Han, P., Qian, J., *et al.* (2020) 'Knockdown of ALPK2 blocks development and progression of renal cell carcinoma', *Experimental Cell Research*, 392(2). doi: 10.1016/J.YEXCR.2020.112029.

Jiang, X., Finucane, H. K., Schumacher, F. R., *et al.* (2019) 'Shared heritability and functional enrichment across six solid cancers', *Nature Communications*, 10(1), pp. 1–23. doi: 10.1038/s41467-018-08054-4.

Justia US Supreme Court (2013) *Assoc. for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576*.

Kamiza, A. B., Toure, S. M., Vujkovic, M., *et al.* (2022) 'Transferability of genetic risk scores in African populations', *Nature Medicine*, 28, pp. 1163–1166. doi: 10.1038/s41591-022-01835-x.

Khurana, E., Fu, Y., Chakravarty, D., *et al.* (2016) 'Role of non-coding sequence variants in cancer', *Nature Reviews Genetics 2016 17:2*, 17(2), pp. 93–108. doi: 10.1038/nrg.2015.17.

Kiendrebeogo, I. T., Zoure, A. A., Sorgho, P. A., *et al.* (2019) 'Glutathione S-transferase M1 (GSTM1) and T1 (GSTT1) variants and breast cancer risk in Burkina Faso', *Biomolecular Concepts*, 10(1), pp. 175–183. doi: 10.1515/bmc-2019-0020.

Kiendrebeogo, I. T., Zoure, A. A., Zongo, F. I., *et al.* (2022) 'Mutations in exon 11 (11.1 and 11.2) of the BRCA1 gene and risk factors for breast cancer in Burkina Faso', *Journal of Public Health in Africa*, 13(1). doi: 10.4081/JPHIA.2022.1921.

Killian, J. K., Dorssers, L. C. J., Trabert, B., *et al.* (2016) 'Imprints and DPPA3 are bypassed during pluripotency -and differentiation-coupled methylation reprogramming in testicular germ cell tumors', *Genome Research*, 26(11), pp. 1490–1504. doi: 10.1101/GR.201293.115/-/DC1.

Kuchenbaecker, K., Telkar, N., Reiker, T., *et al.* (2019) 'The transferability of lipid loci across African, Asian and European cohorts', *Nature Communications*, 10(1). doi: 10.1038/s41467-019-12026-7.

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., *et al.* (2017) 'Risks of breast, ovarian, and

contralateral breast cancer for BRCA1 and BRCA2 mutation carriers', *JAMA - Journal of the American Medical Association*, 317(23), pp. 2402–2416. doi: 10.1001/jama.2017.7112.

Lee, J., Beliakoff, J. and Sun, Z. (2007) 'The novel PIAS-like protein hZimp10 is a transcriptional co-activator of the p53 tumor suppressor', *Nucleic Acids Research*, 35(13), pp. 4523–4534. doi: 10.1093/NAR/GKM476.

Lee, J. Y. L., Ekambaram, P., Carleton, N. M., *et al.* (2022) 'MALT1 Is a Targetable Driver of Epithelial-to-Mesenchymal Transition in Claudin-Low, Triple-Negative Breast Cancer', *Molecular Cancer Research*, 20(3), pp. 373–386. doi: 10.1158/1541-7786.MCR-21-0208.

Lee, K., Kruper, L., Dieli-Conwright, C. M., *et al.* (2019) 'The Impact of Obesity on Breast Cancer Diagnosis and Treatment', *Current Oncology Reports*, 21(5). doi: 10.1007/S11912-019-0787-1.

Li, Y., Willer, C., Sanna, S., *et al.* (2009) 'Genotype Imputation', *Annual Review of Genomics and Human Genetics*, 10, pp. 387–406. doi: 10.1146/annurev.genom.9.081307.164242.

Liao, S. Y., Chiang, C. W., Hsu, C. H., *et al.* (2017) 'CK1δ/GSK3β/FBXW7α axis promotes degradation of the ZNF322A oncoprotein to suppress lung cancer progression', *Oncogene*, 36(41), pp. 5722–5733. doi: 10.1038/ONC.2017.168.

Lilyquist, J., Ruddy, K. J., Vachon, C. M., *et al.* (2018) 'Common Genetic Variation and Breast Cancer Risk - Past, present, and future', *Cancer Epidemiology Biomarkers & Prevention*, 27(4), pp. 380–394. doi: 10.1158/1055-9965.EPI-17-1144.

Loh, P. R., Danecek, P., Palamara, P. F., *et al.* (2016) 'Reference-based phasing using the Haplotype Reference Consortium panel', *Nature Genetics*, 48(11), pp. 1443–1448. doi: 10.1038/NG.3679.

Loibl, S., Poortmans, P., Morrow, M., *et al.* (2021) 'Breast cancer', *The Lancet*, 397(10286), pp. 1750–1769. doi: 10.1016/S0140-6736(20)32381-3.

Long, J., Zhang, B., Signorello, L. B., *et al.* (2013) 'Evaluating genome-wide association study-identified breast cancer risk variants in African-American women', *PloS One*, 8(4). doi: 10.1371/JOURNAL.PONE.0058350.

Lonsdale, J., Thomas, J., Salvatore, M., *et al.* (2013) 'The Genotype-Tissue Expression (GTEx)

project', *Nature Genetics*, 45(6), pp. 580–585. doi: 10.1038/ng.2653.

Lukong, K. E., Ogunbolude, Y. and Kamdem, J. P. (2017) 'Breast cancer in Africa: prevalence, treatment options, herbal medicines, and socioeconomic determinants', *Breast Cancer Research and Treatment*, pp. 351–365. doi: 10.1007/s10549-017-4408-0.

Luo, J., Schumacher, M., Scherer, A., *et al.* (2010) 'A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data', *The Pharmacogenomics Journal*, 10(4), pp. 278–291. doi: 10.1038/tpj.2010.57.

Maher, B. (2008) 'Personal genomes: The case of the missing heritability', *Nature*, 456(7218), pp. 18–21. doi: 10.1038/456018A.

Martin, A. R., Gignoux, C. R., Walters, R. K., *et al.* (2017) 'Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations', *American Journal of Human Genetics*, 100(4), pp. 635–649.

Martin, A. R., Kanai, M., Kamatani, Y., *et al.* (2019) 'Clinical use of current polygenic risk scores may exacerbate health disparities', *Nature Genetics 2019 51:4*, 51(4), pp. 584–591. doi: 10.1038/s41588-019-0379-x.

Mavaddat, N., Michailidou, K., Dennis, J., *et al.* (2019) 'Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes', *American Journal of Human Genetics*, 104(1), pp. 21–34. doi: 10.1016/j.ajhg.2018.11.002.

Maxwell, K. N. and Nathanson, K. L. (2013) 'Common breast cancer risk variants in the post-COGS era: A comprehensive review', *Breast Cancer Research*. doi: 10.1186/bcr3591.

May, A., Hazelhurst, S., Li, Y., *et al.* (2013) 'Genetic diversity in black South Africans from Soweto', *BMC Genomics*, 14(644), pp. 1–12. doi: 10.1186/1471-2164-14-644.

McAuley, J. R., Bailey, K. M., Ekambaram, P., *et al.* (2019) 'MALT1 is a critical mediator of PAR1-driven NF-κB activation and metastasis in multiple tumor types', *Oncogene*, 38(49), pp. 7384–7398. doi: 10.1038/S41388-019-0958-4.

McCarthy, S., Das, S., Kretzschmar, W., *et al.* (2016) 'A reference panel of 64,976 haplotypes for genotype imputation', *Nature Genetics*, 48(10), pp. 1279–1283. doi: 10.1038/ng.3643.

McPherson, K., Steel, C. M. and Dixon, J. M. (2000) 'Breast cancer—epidemiology, risk factors, and genetics', *BMJ*, 321(7261), pp. 624–628. doi: 10.1136/BMJ.321.7261.624.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., *et al.* (2002) 'Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations', *Nature Genetics*, 31(1), pp. 55–59. doi: 10.1038/ng879.

Melchor, L. and Benítez, J. (2013) 'The complex genetic landscape of familial breast cancer', *Human Genetics*, 132(8), pp. 845–863. doi: 10.1007/s00439-013-1299-y.

Meng, X., Dong, Y., Yu, X., *et al.* (2017) 'MREG suppresses thyroid cancer cell invasion and proliferation by inhibiting Akt-mTOR signaling', *Biochemical and Biophysical Research Communications*, 491(1), pp. 72–78. doi: 10.1016/J.BBRC.2017.07.044.

van der Merwe, N. C. N., Hamel, N., Schneider, S.-R. R., *et al.* (2012) 'A founder BRCA2 mutation in non-Afrikaner breast cancer patients of the Western Cape of South Africa', *Clinical Genetics*, 81(2), pp. 179–184. doi: 10.1111/j.1399-0004.2010.01617.x.

Van der Merwe, N. C., Combrink, H. M. E., Ntaita, K. S., *et al.* (2022) 'Prevalence of Clinically Relevant Germline BRCA Variants in a Large Unselected South African Breast and Ovarian Cancer Cohort: A Public Sector Experience', *Frontiers in Genetics*, 13. doi: 10.3389/FGENE.2022.834265/PDF.

Michailidou, K., Hall, P., Gonzalez-neira, A., *et al.* (2013) 'Large-scale genotyping identifies 41 new loci associated with breast cancer risk', *Nature Genetics*, 45(4), pp. 353–361. doi: 10.1038/ng.2563.Large-scale.

Michailidou, K., Beesley, J., Lindstrom, S., *et al.* (2015) 'Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer', *Nature Genetics*, 47(4), pp. 373–380. doi: 10.1038/ng.3242.

Michailidou, K., Lindström, S., Dennis, J., *et al.* (2017) 'Association analysis identifies 65 new breast cancer risk loci', *Nature*, 551, pp. 92–94. doi: 10.1038/nature24284.

Michailidou, K. (2018) 'Meta-Analysis of Common and Rare Variants', *Methods in Molecular Biology*, 1793, pp. 73–88. doi: 10.1007/978-1-4939-7868-7_6.

Miki, Y., Swensen, J., Shattuck-Eidens, D., *et al.* (1994) 'A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.', *Science*, 266(5182), pp. 66–71.

Miller, S. A., Dykes, D. D. and Polesky, H. F. (1988) 'A simple salting out procedure for extracting DNA from human nucleated cells', *Nucleic Acids Research*, 16(3), p. 1215. doi: 10.1093/nar/16.3.1215.

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., *et al.* (2017) 'Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer', *Nature Genetics*, 49(12), pp. 1767–1778. doi: 10.1038/ng.3785.

Mocellin, S., Tropea, S., Benna, C., *et al.* (2018) 'Circadian pathway genetic variation and cancer risk: evidence from genome-wide association studies', *BMC Medicine*, 16(20), pp. 1–8. doi: 10.1186/S12916-018-1010-1.

Möller, S., Mucci, L. A., Harris, J. R., *et al.* (2016) 'The heritability of breast cancer among women in the nordic twin study of cancer', *Cancer Epidemiology Biomarkers and Prevention*, 25(1), pp. 145–150. doi: 10.1158/1055-9965.EPI-15-0913.

Momenimovahed, Z. and Salehiniya, H. (2019) 'Epidemiological characteristics of and risk factors for breast cancer in the world', *Breast Cancer: Targets and Therapy*, pp. 151–164. doi: 10.2147/BCTT.S176070.

Mulder, N., Abimiku, A. A., Adebamowo, S. N., *et al.* (2018) 'H3Africa: current perspectives.', *Pharmacogenomics and Personalized Medicine*, 11, pp. 59–66. doi: 10.2147/PGPM.S141546.

Munn, Z., Peters, M. D. J., Stern, C., *et al.* (2018) 'Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach', *BMC Medical Research Methodology*, 18(1), pp. 1–7. doi: 10.1186/s12874-018-0611-x.

Mvila, G. L., Postema, S., Marchal, G., *et al.* (2014) 'From the set-up of a screening program of breast cancer patients to the identification of the first BRCA mutation in the DR Congo', *BMC Public Health*, 14(1), pp. 1–10. doi: 10.1186/1471-2458-14-759.

Naj, A. C. (2019) 'Genotype Imputation in Genome-Wide Association Studies', *Current Protocols in Human Genetics*, 102(1). doi: 10.1002/CPHG.84.

Nakagawa, H. and Fujita, M. (2018) 'Whole genome sequencing analysis for cancer genomics and precision medicine', *Cancer Science*, 109(3), pp. 513–522. doi: 10.1111/CAS.13505.

Ndiaye, R., Diop, J. P. D., Bourdon-Huguenin, V., *et al.* (2020) 'Evidence for an ancient BRCA1 pathogenic variant in inherited breast cancer patients from Senegal', *npj Genomic Medicine*, 5(8), pp. 1–6. doi: 10.1038/s41525-020-0114-7.

Ngoma, T. and Ngoma, M. (2019) 'Cancer control in Africa: Is cancer research a luxury or necessity?', *eCancer Medical Science*, 13(947), pp. 1–5. doi: 10.3332/ecancer.2019.947.

Nguyen, J. V and Thomas, M. H. (2019) 'Beyond BRCA: Review of Hereditary Syndromes Predisposing to Breast Cancer', *Journal of Breast Imaging*, 1(2), pp. 84–91. doi: 10.1093/jbi/wbz014.

Nichols, H. B., Graff, M., Bensen, J. T., *et al.* (2021) 'Genetic variants in anti-Müllerian hormone-related genes and breast cancer risk: results from the AMBER consortium', *Breast Cancer Research and Treatment*, 185(2), pp. 469–478. doi: 10.1007/S10549-020-05944-4.

Oeffinger, K. C., Fontham, E. T. H., Etzioni, R., *et al.* (2015) 'Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society', *JAMA*, 314(15), pp. 1599–1614. doi: 10.1001/JAMA.2015.12783.

Okobia, M. N., Bunker, C. H., Garte, S. J., *et al.* (2008) 'Leptin receptor Gln223Arg polymorphism and breast cancer risk in Nigerian women: A case control study', *BMC Cancer*, 8, pp. 1–8. doi: 10.1186/1471-2407-8-338.

Okobia, M. N., Bunker, C. H., Garte, S. J., *et al.* (2009) 'Cytochrome P450 1B1 Val432Leu polymorphism and breast cancer risk in Nigerian women: A case control study', *Infectious Agents and Cancer*, 4(SUPPL. 1), pp. 1–8. doi: 10.1186/1750-9378-4-S1-S12.

Ouedraogo, S. Y., Zoure, A. A., Zeye, M. M. J., *et al.* (2022) 'BRCA1, BRCA2, TP53, PIK3CA, PTEN and AKT1 genes mutations in Burkina Faso breast cancer patients: prevalence, spectrum and novel variant', *Molecular Genetics and Genomics: MGG*. doi: 10.1007/S00438-022-01914-1.

Pacella-Norman, R., Urban, M. I., Sitas, F., *et al.* (2002) 'Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans.', *British Journal of Cancer*, 86(11), pp. 1751–1756.

doi: 10.1038/sj.bjc.6600338.

Palmer, J. R., Ambrosone, C. B. and Olshan, A. F. (2014) 'A collaborative study of the etiology of breast cancer subtypes in African American women: the AMBER consortium', *Cancer Causes & Control: CCC*, 25(3), pp. 309–319. doi: 10.1007/S10552-013-0332-8.

Palmer, J. R., Polley, E. C., Hu, C., *et al.* (2020) 'Contribution of Germline Predisposition Gene Mutations to Breast Cancer Risk in African American Women', *Journal of the National Cancer Institute*, 112(12), pp. 1213–1221. doi: 10.1093/jnci/djaa040.

Parys, J. B., Bultynck, G. and Vervliet, T. (2021) 'IP 3 Receptor Biology and Endoplasmic Reticulum Calcium Dynamics in Cancer', *Progress in Molecular and Subcellular Biology*, 59, pp. 215–237. doi: 10.1007/978-3-030-67696-4_11.

Perou, C. M., Sørlie, T., Eisen, M. B., *et al.* (2000) 'Molecular portraits of human breast tumours', *Nature*, 406(6797), pp. 747–752. doi: 10.1038/35021093.

Phipson, B., Smyth Walter, G. K. and Hall, E. (2010) 'Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn', *Statistical Applications in Genetics and Molecular Biology*, 9(1), pp. 1–12. doi: 10.2202/1544-6115.1585.

Picon-Ruiz, M., Morata-Tarifa, C., Valle-Goffin, J. J., *et al.* (2017) 'Obesity and adverse breast cancer risk and outcome: Mechanistic insights and strategies for intervention', *CA: A Cancer Journal for Clinicians*, 67(5), pp. 378–397. doi: 10.3322/CAAC.21405.

Pirie, G. H. and da Silva., M. (1986) 'Hostels for African Migrants in Greater Johannesburg', *GeoJournal*, 12(2), pp. 173–180.

Pritchard, J. K. and Przeworski, M. (2001) 'Linkage Disequilibrium in Humans: Models and Data', *The American Journal of Human Genetics*, 69(1), pp. 1–14. doi: 10.1086/321275.

Pruim, R. J., Welch, R. P., Sanna, S., *et al.* (2010) 'LocusZoom: regional visualization of genome-wide association scan results', *Bioinformatics*, 26(18), pp. 2336–2337. doi: 10.1093/bioinformatics/btq419.

Pumpalova, Y. S., Ayeni, O. A., Chen, W. C., *et al.* (2022) 'The Impact of Breast Cancer Treatment Delays on Survival Among South African Women', *The Oncologist*, 27(3), pp. e233–e243. doi:

10.1093/ONCOLO/OYAB054.

Rahman, N., Seal, S., Thompson, D., *et al.* (2007) 'PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene', *Nature Genetics*, 39(2), pp. 165–167. doi: 10.1038/ng1959.

Rather, G. M., Pramono, A. A., Szekely, Z., *et al.* (2021) 'In cancer, all roads lead to NADPH', *Pharmacology & Therapeutics*, 226. doi: 10.1016/J.PHARMTHERA.2021.107864.

Renwick, A., Thompson, D., Seal, S., *et al.* (2006) 'ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles', *Nature Genetics*, 38(8), pp. 873–875. doi: 10.1038/ng1837.

Rojas, K. and Stuckey, A. (2016) 'Breast Cancer Epidemiology and Risk Factors', *Clinical Obstetrics and Gynecology*, 59(4), pp. 651–672. doi: 10.1097/GRF.0000000000000239.

Rossing, M., Sørensen, C. S., Ejlertsen, B., *et al.* (2019) 'Whole genome sequencing of breast cancer', *APMIS*, 127, pp. 303–315. doi: 10.1111/APM.12920.

Roth, G. A., Abate, D., Abate, K. H., *et al.* (2018) 'Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017', *The Lancet*, 392(10159), pp. 1736–1788. doi: 10.1016/S0140-6736(18)32203-7.

Roy, R., Chun, J. and Powell, S. N. (2011) 'BRCA1 and BRCA2: different roles in a common pathway of genome protection', *Nature Reviews Cancer*, 12(1), pp. 68–78. doi: 10.1038/nrc3181.

Ruan, Y., Lin, Y. F., Feng, Y. C. A., *et al.* (2022) 'Improving polygenic prediction in ancestrally diverse populations', *Nature Genetics 2022 54:5*, 54(5), pp. 573–580. doi: 10.1038/s41588-022-01054-7.

Ruiz-Narváez, E. A., Sucheston-Campbell, L., Bensen, J. T., *et al.* (2016) 'Admixture mapping of African-American women in the AMBER Consortium identifies new loci for breast cancer and estrogen-receptor subtypes', *Frontiers in Genetics*, 7, pp. 1–10. doi: 10.3389/FGENE.2016.00170/PDF.

Sadeghi, F., Asgari, M., Matloubi, M., *et al.* (2020) 'Molecular contribution of BRCA1 and BRCA2 to genome instability in breast cancer patients: Review of radiosensitivity assays', *Biological Procedures Online*, 22(1), pp. 1–28. doi: 10.1186/S12575-020-00133-5/TABLES/6.

Sagna, T., Bonora, E., Ouedraogo, M. N. L., *et al.* (2019) 'Identification of brca1/2 p.ser1613gly, p.pro871leu, p.lys1183arg, p.glu1038gly, p.ser1140gly, p.ala2466val, p.his2440arg variants in women under 45 years old with breast nodules suspected of having breast cancer in burkina faso', *Biomolecular Concepts*, 10(1), pp. 120–127. doi: 10.1515/bmc-2019-0015.

Sampson, J. N., Wheeler, W. A., Yeager, M., *et al.* (2015) 'Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types', *Journal of the National Cancer Institute*, 107(12). doi: 10.1093/JNCI/DJV279.

Sardanelli, F., Podo, F., Santoro, F., *et al.* (2011) 'Multicenter surveillance of women at high genetic breast cancer risk using mammography, ultrasonography, and contrast-enhanced magnetic resonance imaging (the high breast cancer risk italian 1 study): Final results', *Investigative Radiology*, 46(2), pp. 94–105. doi: 10.1097/RLI.0B013E3181F3FCDF.

Savage, K. I. and Harkin, D. P. (2015) 'BRCA1, a "complex" protein involved in the maintenance of genomic stability', *The FEBS Journal*, 282(4), pp. 630–646. doi: 10.1111/FEBS.13150.

Sengupta, D., Choudhury, A., Fortes-Lima, C., *et al.* (2021) 'Genetic substructure and complex demographic history of South African Bantu speakers', *Nature Communications*, 12, pp. 1–13. doi: 10.1038/s41467-021-22207-y.

Sengupta, D., Botha, G., Meintjes, A., *et al.* (2023) 'Performance and accuracy evaluation of reference panels for genotype imputation in sub-Saharan African populations', *Cell Genomics*, 3(6), p. 100332. doi: 10.1016/J.XGEN.2023.100332.

Shah, C., Bauer-Nilsen, K., McNulty, R. H., *et al.* (2020) 'Novel radiation therapy approaches for breast cancer treatment', *Seminars in Oncology*, 47(4), pp. 209–216. doi: 10.1053/j.seminoncol.2020.05.003.

Shahid, T., Soroka, J., Kong, E. H., *et al.* (2014) 'Structure and mechanism of action of the BRCA2 breast cancer tumor suppressor', *Nature Structural and Molecular Biology*, 21(11), pp. 962–968.

doi: 10.1038/nsmb.2899.

Shield, K. D., Soerjomataram, I. and Rehm, J. (2016) 'Alcohol Use and Breast Cancer: A Critical Review', *Alcoholism, Clinical and Experimental Research*, 40(6), pp. 1166–1181. doi: 10.1111/ACER.13071.

Shirts, B. H., Casadei, S., Jacobson, A. L., *et al.* (2016) 'Improving performance of multigene panels for genomic analysis of cancer predisposition', *Genetics in Medicine*, 18(10), pp. 974–981. doi: 10.1038/gim.2015.212.

Shriner, D., Adeyemo, A., Gerry, N. P., *et al.* (2009) 'Transferability and Fine-Mapping of Genome-Wide Associated Loci for Adult Height across Human Populations', *PLoS ONE*, 4(12), pp. 1–9. doi: 10.1371/journal.pone.0008398.

Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) 'The Missing Diversity in Human Genetic Studies', *Cell*, 177(1), pp. 26–31. doi: 10.1016/J.CELL.2019.02.048.

Slamon, D., Eiermann, W., Robert, N., *et al.* (2011) 'Adjuvant trastuzumab in HER2-positive breast cancer', *The New England Journal of Medicine*, 365(14), pp. 1273–1283. doi: 10.1056/NEJMOA0910383.

Sliwa, K., Wilkinson, D., Hansen, C., *et al.* (2008) 'Spectrum of heart disease and risk factors in a black urban population in South Africa (the Heart of Soweto Study): a cohort study', *Lancet*, 371(9616), pp. 915–922. doi: 10.1016/S0140-6736(08)60417-1.

Smith, M. L., Baggerly, K. A., Bengtsson, H., *et al.* (2013) 'illuminaio: An open source IDAT parsing tool for Illumina microarrays', *F1000Research*, 2(264). doi: 10.12688/f1000research.2-264.v1.

Soghoian, D., Jayaraman, V., Silane, M., *et al.* (2005) 'Inositol 1,4,5-trisphosphate receptor phosphorylation in breast cancer', *Tumour Biology*, 26(4), pp. 207–212. doi: 10.1159/000086954.

Song, L., Liu, A. and Shi, J. (2019) 'SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics', *Bioinformatics*, 35(20), p. 4038. doi: 10.1093/BIOINFORMATICS/BTZ176.

Soo, C. C., Mukomana, F., Hazelhurst, S., *et al.* (2017) 'Establishing an academic biobank in a

resource-challenged environment', *South African Medical Journal*, 107(6), pp. 486–492. doi: 10.7196/SAMJ.2017.V107I6.12099.

Sørlie, T., Perou, C. M., Tibshirani, R., *et al.* (2001) 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), pp. 10869–10874. doi: 10.1073/pnas.191367098.

Speed, D., Cai, N., Johnson, M. R., *et al.* (2017) 'Reevaluation of SNP heritability in complex human traits', *Nature Genetics 2017 49:7*, 49(7), pp. 986–992. doi: 10.1038/ng.3865.

Speed, D., Holmes, J. and Balding, D. J. (2020) 'Evaluating and improving heritability models using summary statistics', *Nature Genetics 2020 52:4*, 52(4), pp. 458–462. doi: 10.1038/s41588-020-0600-y.

Stacey, S. N., Manolescu, A., Sulem, P., *et al.* (2007) 'Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer', *Nature Genetics*, 39(7), pp. 865–869. doi: 10.1038/ng2064.

Sung, H., Ferlay, J., Siegel, R. L., *et al.* (2021) 'Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries', *CA: A Cancer Journal for Clinicians*, 71(3), pp. 209–249. doi: 10.3322/CAAC.21660.

Swiel, Y., Brandenburg, J.-T., Hayat, M., *et al.* (2022) 'FPGA Acceleration of GWAS Permutation Testing', *bioRxiv*, pp. 1–10. doi: 10.1101/2022.03.11.483235.

Tarsounas, M. and Sung, P. (2020) 'The antitumorigenic roles of BRCA1–BARD1 in DNA repair and replication', *Nature Reviews Molecular Cell Biology*, 21, pp. 284–299. doi: 10.1038/s41580-020-0218-z.

Tavassoli, F. A. and Devilee, P. (2003) 'WHO Pathology and Genetics of Tumours of the Breast and Female Genital Organs', *IARC Press: Lyon 2003*, pp. 1–430.

Thacker, J. and Zdzienicka, M. Z. (2004) 'The XRCC genes: expanding roles in DNA double-strand break repair', *DNA Repair*, 3(8–9), pp. 1081–1090. doi: 10.1016/J.DNAREP.2004.04.012.

The 1000 Genomes Project Consortium (2015) 'A global reference for human genetic variation',

Nature, 526, pp. 68–74. doi: 10.1038/nature15393.

Tiofack, A. T. Z., Simo, G., Ofon, E. A., *et al.* (2020) 'The TP63 Gene Polymorphism rs17506395 is Associated with Early Breast Cancer in Cameroon', *Asian Pacific Journal of Cancer Prevention*, 21(8), pp. 2199–2208. doi: 10.31557/APJCP.2020.21.8.2199.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., *et al.* (2009) 'The genetic structure and history of Africans and African Americans.', *Science*, 324(5930), pp. 1035–1044. doi: 10.1126/science.1172257.

Toh, M. R., Low, C. E., Chong, S. T., *et al.* (2020) 'Missense PALB2 germline variant disrupts nuclear localization of PALB2 in a patient with breast cancer', *Familial Cancer*, 19(2), pp. 123–131. doi: 10.1007/S10689-020-00163-8/TABLES/2.

Torkamani, A., Wineinger, N. E. and Topol, E. J. (2018) 'The personal and clinical utility of polygenic risk scores', *Nature Reviews Genetics*, 19, pp. 581–590. doi: 10.1038/s41576-018-0018-x.

Torrorey-Sawe, R., van der Merwe, N., Mining, S. K., *et al.* (2020) 'Pioneering Informed Consent for Return of Research Results to Breast Cancer Patients Facing Barriers to Implementation of Genomic Medicine: The Kenyan BRCA1/2 Testing Experience Using Whole Exome Sequencing', *Frontiers in Genetics*, 11, pp. 1–12. doi: 10.3389/fgene.2020.00170.

Trayes, K. P. and Cokenakes, S. E. H. (2021) 'Breast Cancer Treatment', *American Family Physician*, 104(2), pp. 171–178.

Truong, T., Liquet, B., Menegaux, F., *et al.* (2014) 'Breast cancer risk, nightwork, and circadian clock gene polymorphisms', *Endocrine-Related Cancer*, 21(4), pp. 629–638. doi: 10.1530/ERC-14-0121.

Turnbull, C., Ahmed, S., Morrison, J., *et al.* (2010) 'Genome-wide association study identifies five new breast cancer susceptibility loci', *Nature Genetics*, 42(6), pp. 504–507. doi: 10.1038/ng.586.

Uno, Y., Uehara, S., Hosokawa, M., *et al.* (2014) 'Systematic identification and characterization of carboxylesterases in cynomolgus macaques', *Drug Metabolism and Disposition*, 42(12), pp. 2002–2006. doi: 10.1124/DMD.114.059972.

Urban, M., Banks, E., Egger, S., *et al.* (2012) 'Injectable and Oral Contraceptive Use and Cancers of the Breast, Cervix, Ovary, and Endometrium in Black South African Women: Case–Control Study', *PLoS Medicine*, 9(3), p. e1001182. doi: 10.1371/journal.pmed.1001182.

Uyisenga, J. P., Segers, K., Lumaka, A. Z., *et al.* (2020) 'Screening of germline mutations in young Rwandan patients with breast cancers', *Molecular Genetics & Genomic Medicine*, 8(11). doi: 10.1002/MGG3.1500.

Veronesi, U., Boyle, P., Goldhirsch, A., *et al.* (2005) 'Breast cancer', *The Lancet*, 365(9472), pp. 1727–1741. doi: 10.1016/S0140-6736(05)66546-4.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., *et al.* (2015) 'Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores', *American Journal of Human Genetics*, 97(4), pp. 576–592. doi: 10.1016/J.AJHG.2015.09.001.

Vinogradova, Y., Coupland, C. and Hippisley-Cox, J. (2020) 'Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases', *BMJ*, 371. doi: 10.1136/BMJ.M3873.

Visscher, P. M., Hill, W. G. and Wray, N. R. (2008) 'Heritability in the genomics era — concepts and misconceptions', *Nature Reviews Genetics 2008 9:4*, 9(4), pp. 255–266. doi: 10.1038/nrg2322.

Wang, H.-C., Lie, C.-S., Chiu, C.-F., *et al.* (2009) 'Significant association of DNA repair gene Ku80 genotypes with breast cancer susceptibility in Taiwan', *Anticancer Research*, 29(12), pp. 5251–5254.

Wang, Y., Guo, J., Ni, G., *et al.* (2020) 'Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations', *Nature Communications*, 11(1), pp. 1–9. doi: 10.1038/s41467-020-17719-y.

Watanabe, K., Taskesen, E., van Bochoven, A., *et al.* (2017) 'Functional mapping and annotation of genetic associations with FUMA', *Nature Communications*, 8. doi: 10.1038/s41467-017-01261-5.

Weeks, D. E. and Lathrop, G. M. (1995) 'Polygenic disease: methods for mapping complex

disease traits', *Trends in Genetics*, 11(12), pp. 513–519. doi: 10.1016/S0168-9525(00)89163-5.

Wells, A., Heckerman, D., Torkamani, A., *et al.* (2019) 'Ranking of non-coding pathogenic variants and putative essential regions of the human genome', *Nature Communications 2019 10:1*, 10(1), pp. 1–9. doi: 10.1038/s41467-019-13212-3.

Wen, Y. J., Zhang, H., Ni, Y. L., *et al.* (2018) 'Methodological implementation of mixed linear models in multi-locus genome-wide association studies', *Briefings in Bioinformatics*, 19(4), pp. 700–712. doi: 10.1093/BIB/BBW145.

Wild, C. P., Weiderpass, E. and Stewart, B. W. (2020) 'World cancer report: cancer research for cancer prevention', *Lyon, France: International Agency for Research on Cancer*, pp. 1–630. doi: Available from: http://publications.iarc.fr/586.

Willems, P., Claes, K., Baeyens, A., *et al.* (2008) 'Polymorphisms in nonhomologous end-joining genes associated with breast cancer risk and chromosomal radiosensitivity', *Genes, Chromosomes & Cancer*, 47(2), pp. 137–148. doi: 10.1002/GCC.20515.

Willer, C. J., Li, Y. and Abecasis, G. R. (2010) 'METAL: fast and efficient meta-analysis of genomewide association scans', *Bioinformatics*, 26(17), pp. 2190–2191. doi: 10.1093/bioinformatics/btq340.

Wooster, R., Bignell, G., Lancaster, J., *et al.* (1995) 'Identification of the breast cancer susceptibility gene BRCA2', *Nature*, 378(6559), pp. 789–792. doi: 10.1038/378789a0.

Xinhua, Z., Yiming, M., Weiwei, L., *et al.* (2021) 'Effect of RNF152 on NO induced apoptosis of colon cancer cells', *Chinese Journal of Oncology]*, 43(5), pp. 528–532. doi: 10.3760/CMA.J.CN112152-20201014-00898.

Yang, J., Benyamin, B., McEvoy, B. P., *et al.* (2010) 'Common SNPs explain a large proportion of the heritability for human height', *Nature Genetics*, 42(7), pp. 565–569.

Yang, J., Lee, S. H., Goddard, M. E., *et al.* (2011) 'GCTA: a tool for genome-wide complex trait analysis', *American Journal of Human Genetics*, 88, pp. 76–82. doi: 10.1016/J.ajhg.2010.11.011.

Yang, J., Zaitlen, N. A., Goddard, M. E., *et al.* (2014) 'Advantages and pitfalls in the application of mixed model association methods', *Nature Genetics*, 46(2), pp. 100–106. doi: 10.1038/ng.2876.

Yang, J., Zeng, J., Goddard, M. E., *et al.* (2017) 'Concepts, estimation and interpretation of SNP-based heritability', *Nature Genetics 2017 49:9*, 49(9), pp. 1304–1310. doi: 10.1038/ng.3941.

Yu, L. X., Liu, L. Y., Xiang, Y. J., *et al.* (2021) 'XRCC5/6 polymorphisms and their interactions with smoking, alcohol consumption, and sleep satisfaction in breast cancer risk: A Chinese multi-center study', *Cancer Medicine*, 10(8), pp. 2752–2762. doi: 10.1002/CAM4.3847.

Zhang, B., Fackenthal, J. D., Niu, Q., *et al.* (2009) 'Evidence for an ancient BRCA1 mutation in breast cancer patients of yoruban ancestry', *Familial Cancer*, 8(1), pp. 15–22. doi: 10.1007/s10689-008-9205-9.

Zhang, J., Fackenthal, J. D., Zheng, Y., *et al.* (2012) 'Recurrent BRCA1 and BRCA2 mutations in breast cancer patients of African ancestry', *Breast Cancer Research and Treatment*, 134(2), pp. 889–894. doi: 10.1007/s10549-012-2136-z.

Zhang, Y., Wei, H., Fan, L., *et al.* (2021) 'CLEC4s as Potential Therapeutic Targets in Hepatocellular Carcinoma Microenvironment', *Frontiers in Cell and Developmental Biology*, 9. doi: doi.org/10.3389/fcell.2021.681372.

Zheng, W., Long, J., Gao, Y.-T., *et al.* (2009) 'Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1', *Nature Genetics*, 41(3), pp. 324–328. doi: 10.1038/ng.318.

Zheng, Y., Walsh, T., Gulsuner, S., *et al.* (2018) 'Inherited breast cancer in Nigerian women', *Journal of Clinical Oncology*, 36(28), pp. 2820–2825. doi: 10.1200/JCO.2018.78.3977.

Zhou, L. P., Luan, H., Dong, X. H., *et al.* (2012) 'Association between XRCC5, 6 and 7 gene polymorphisms and the risk of breast cancer: a HuGE review and meta-analysis', *Asian Pacific Journal of Cancer Prevention*, 13(8), pp. 3637–3643. doi: 10.7314/APJCP.2012.13.8.3637.

Zhou, X. and Stephens, M. (2012) 'Genome-wide efficient mixed-model analysis for association studies', *Nature Genetics*, 44(7), pp. 821–824. doi: 10.1038/ng.2310.

Zhu, H. and Zhou, X. (2020) 'Statistical methods for SNP heritability estimation and partition: A review', *Computational and Structural Biotechnology Journal*, 18, pp. 1557–1568. doi: 10.1016/J.CSBJ.2020.06.011.

Zhu, X., Yan, S., Xiao, S., *et al.* (2020) 'Knockdown of ALPK2 inhibits the development and progression of Ovarian Cancer', *Cancer Cell International*, 20(1). doi: 10.1186/s12935-020-01347-z.

Zhu, Z., Teng, Z., van Duijnhoven, F. J. B., *et al.* (2017) 'Interactions between RASA2, CADM1, HIF1AN gene polymorphisms and body fatness with breast cancer: a population-based case-control study in China', *Oncotarget*, 8(58), pp. 98258–98269. doi: 10.18632/ONCOTARGET.21530.

Zoure, A. A., Slaoui, M., Bambara, H. A., *et al.* (2018) 'BRCA1 c.68_69delAG (exon2), c.181T>G (exon5), c.798_799delTT and 943ins10 (exon11) mutations in Burkina Faso', *Journal of Public Health in Africa*, 9(663), pp. 11–15. doi: 10.4081/jphia.2018.

Appendices

7. Appendix A: JCS questionnaire

MRC/NHLS/Wits CANCER EPIDEMIOLOGY RESEARCH GROUP (CERG)
JOHANNESBURG STUDY QUESTIONNAIRE FOR PATIENTS

Date of Birth Day Month Year 1 9 OR Age Years Study No: Sex (please tick) 1 M F 2

Diagnosis(Clinical) Path No ICD

Histology result, if available Hospital No. M

Initials of Interviewer ..... Date of Interview Day Month Year 2 0

BLOODS:
RED TOP yes no
PURPLE TOP yes no

Where were you born? (please tick)
1 Soweto 2 Jhb 3 Gauteng 4 N.Prov. 5 Mpum 6 NW 7 FS 8 ECape
9 W.Cape 10 N.Cape 11 Kwazulu Natal 12. Other ..... 1 Urban 2 Rural

Where do you normally live? (please tick)
1 Soweto 2 Jhb 3 Gauteng 4 N.Prov. 5 Mpum 6 NW 7 FS 8 ECape
9 W.Cape 10 N.Cape 11 Kwazulu Natal 12. Other ..... 1 Urban 2 Rural

How long have you lived there? Years

What is the highest standard you passed at school? (please tick) Gr10 Gr11 Gr12

1	2	3	4	5	6	7	8	9	10	11	12	13	14
None	Sub A Gr1	Sub B Gr2	Std 1 Gr3	Std 2 Gr4	Std 3 Gr5	Std 4 Gr6	Std 5 Gr7	Std 6 Gr8	Std 7 Gr9	Std 8 Form 3 NTC1	Std 9 Form 4 NTC2	Std 10 Form 5 NTC3	Univ tech

What are the walls of your home made of? (please tick)
1. brick/cement 2. Tin 3. Plastic 4. Wood 5. Other (specify) ..... 6. mud/clay

Where do you normally cook food in your house? 1 Inside Outside 2

What type of fuel do you MOSTLY use in your home for cooking? (please tick)
1. Wood 2. Charcoal 3. Coal 4. Anthracite 5. Paraffin 6. Gas 7. Electricity 8. Other ..... 9. dung

What type of fuel do you MOSTLY use in your home for keeping warm? (please tick)
1. Wood 2. Charcoal 3. Coal 4. Anthracite 5. Paraffin 6. Gas 7. Electricity 8. Other ..... 9. dung 10. no fuel

Where did you normally cook food in your house *20 years ago*? 1 Inside Outside 2

What type of fuel did you MOSTLY use in your home for cooking *20 years ago*?
1. Wood 2. Charcoal 3. Coal 4. Anthracite 5. Paraffin 6. Gas 7. Electricity 8. Other ..... 9. dung

What type of fuel did you MOSTLY use in your home for keeping warm *20 years ago*?
1. Wood 2. Charcoal 3. Coal 4. Anthracite 5. Paraffin 6. Gas 7. Electricity 8. Other ..... 9. dung 10.no fuel

In any of the houses you lived in, did the smoke ever make your eyes water?
1 Yes, more than 5 years No 2

Have you ever smoked cigarettes or a pipe regularly? (please tick) 1. Yes(now) 2. In the past 3. Never

If yes: In the past five to ten years, how many would you usually smoke in a day?

	1. Cigarettes	2. Hand rolled cigarettes	3. Pipes	4. Dagga
(number of)				

How old were you when you first started smoking regularly? Years old

If you have stopped smoking, how old were you when you stopped? Years old

Have you ever used snuff? (please tick) 1. Yes(now) 2. In the past 3. Never

In the past five to ten years, how often would you use snuff each day? times per day

Before you became ill, did you ever have your blood pressure measured? (Please tick)
1. Yes 2. No 3. Yes only during pregnancy

If yes, were you told it was high? (Please tick) 1. Yes 2. No 3. Yes only during pregnancy

Have you ever had diabetes? (Please tick) 1. Yes 2. No

22 January 2010

Before you became ill, about how much wine, beer or spirits did you drink on average each week?
(Please indicate glass/bottle size when you select block for each type of drink - eg. 3 large glasses of maize beer per week would be 3 e *if none then* 0

BEER			SPIRITS		WINE	CIDER; other	OTHER
sorghum (cartoon)	homemade (specify type)	commercial	Homemade	commercial		commercial alcoholic fruit drinks	SPECIFY

a: 2l (wine bottle)
b: 1l bottle (1 litre)
c: 1l carton
d: 750ml (wine bottle)
e: 500ml (beer glass)
f: 375ml (half jack/pint)
g: 340ml (beer can/dumpy)
h: 250ml glass
i: 200ml (wine) glass/nip
j : 100ml (small glass)
k 50ml (miniature airplane bottle)
l 25ml (tot)
mother (please specify)



Are you? *(please tick)*
1. ☐ Single (never been married) 2. ☐ Married(living as married) 3. ☐ Widowed 4. ☐ Separated

How many husbands/wives have you had?

FOR WOMEN ONLY: Have you ever been pregnant? *(please tick)* 1 Yes No 2

If yes, how many times have you been pregnant? ☐☐

If yes, how many times have you had a miscarriage? ☐☐

If yes, how many of your children were born alive? ☐☐

FOR MEN ONLY: How many children *(dead or alive)* do you have?

ALL: How many mothers / fathers do the children have?

ALL: How old is your oldest child now? *(include dead & alive)* ☐☐ Years old

ALL: How old is your youngest child now? *(include dead & alive)* ☐☐ Years old

ALL: How many boyfriends/girlfriends have you had?

FOR WOMEN ONLY
How old were you when your periods began? ☐☐ Years old

Age now (from previous page) _____

Age at first birth _____

Age at last birth _____

Have your periods ended? Yes ☐1 No ☐2 Not sure ☐3 If yes, how old were you when they ended? ☐☐ years old

If Yes, did your periods stop: ☑

1. naturally	2. due to surgery	3. due to medical treatment	4. due to IC use	5. other	6. unknown

	Oral	Injectable
Have you ever taken contraceptives?	1 Yes No 2	1 Yes No 2
If yes, how old were you when you started taking them?	☐☐	☐☐ Years old
How old were you when you stopped taking them?	☐☐	☐☐ Years old
If yes, for how long in total did you take them?	☐☐	☐☐ Years
Interruptions? ☑	1 Yes No 2	1 Yes No 2

Main reason for stopping, the last time you stopped using the method: ☐☐ ☐☐

1 Nausea 2 high BP 3 headaches 4 wt.Gain 5.changed menstrual pattern 6. menopause 7.sterilization 8.hysterectomy 9. no partner
10. to become pregnant 11 other 0.no specific reason

Have you ever had a 'pap' smear? ☑

0. Never	1. Yes	2. current illness only	9. Can't remember	If yes, how many?	Age at first pap(yrs)	Age at last pap (yrs)

FOR ALL: Describe your usual or past occupation (eg. Driver, machine operator).....

Describe what your usual workplace does or did (eg. A bank, a chemical factory, a gold mine.....

What language do you speak most at home? ☐

What language does/did your father speak? ☐

What language does/did your mother speak? ☐

1. Zulu 2. Xhosa 3. Sotho 4.Pedi 5. Tswana 6. Venda 7. Swazi 8. Shangaan 9. Ndebele
10 English 11. Afrikaans 12. Other

22 January 2010

Study number: _____

Cancer Epidemiology Research Group **Case-Control Study**

Antiretroviral Drug Usage

*Please ask **ALL PATIENTS** the following questions AFTER you have completed the standard questionnaire. Make sure that you emphasize that the information is voluntary and confidential.*

Have you ever been tested for HIV (before current test)? **Yes** $_1$ **No**$_0$ (circle one)

If YES, proceed: Were you tested: **before this illness**$_1$ **/ at time of current illness only**$_2$ (circle one)

Were you given the test results? **Yes**$_1$ **No**$_0$ (circle one)

If YES, proceed: Are you willing to disclose your status? **Yes**$_1$ **No**$_0$ (circle one)

If YES, proceed: Are you HIV: **positive**$_1$ **negative**$_0$ (circle one)

Date of last/current test (mm/yyyy): __________

If POSITIVE, proceed:
Have you ever taken anti-Retroviral medicines for your HIV infection? **Yes**$_1$ **No**$_0$ (circle one)

(Be careful to explain to the patient that we mean "Western" medicine/pills/capsules specifically for the HIV itself, NOT for opportunistic infections.)

If YES, Proceed:

From: _____ (start date) To: _____ (end date)

From: _____ (start date) To: _____ (end date)

From: _____ (start date) To: _____ (end date)

The ARV drugs were provided by (tick all that apply):

☐ Private practitioner$_1$ ☐ Public hospital/clinic$_2$

☐ NGO$_3$ ☐ Study/ clinical trial$_4$

Thank the patient for answering these extra questions even if they only answered the first one.

VERSION: February 2015 CERG Tonicah Maphanga 011489 9150

JOHANNESBURG HOSPITALS NHLS/ CERG CANCER EPIDEMIOLOGY STUDY

Interviewer check list

Note to the Interviewer: **Study no:**_____

This is a checklist for you and a Consent Form for the patient.
Please ask the questions in your own words.
Tick ☑ the boxes and fill in the blanks as you complete that task.

My name is _____*(name of interviewer)*

I work at the National Health Laboratory Service and would like to ask you some questions.
Please ask:

1. *Has the patient ever been diagnosed with a cancer prior to this cancer?* *yes* ☐ *no* ☐ *not known* ☐
   If yes, thank the patient for their time but DO NOT proceed with the interview.
2. *When was the patient diagnosed with the PRESENT CANCER?* ☐ *more than 6 months ago*
   ☐ *less than 6 months ago*
   ☐ *has not yet received diagnosis.*

If the patient was diagnosed more than 6 months ago, thank the patient but DO NOT INTERVIEW THE PATIENT.
If the patient was diagnosed less than 6 months ago, make sure that:

- *the patient has NOT had any chemotherapy or radiotherapy;*
- *the patient has NOT been interviewed for this study before*

.....*Tick the box once you have done this.* ☐

PATIENT CONSENT FORM

A: We would like you to take part in a study to find out if your exposure to substances in the environment, your circumstances and habits, your occupation, previous illnesses, and/or inherited factors play a role in the illness you have developed.

We have chosen a group of patients from this ward / clinic because of their current illness so that we can make comparisons between illnesses. We are asking you to spend 20 – 30 minutes answering questions about yourself, your family, and your work. *Do you have any questions so far?*

You may refuse to answer any questions if you choose. All answers will be treated confidentially. If you decline to participate or if you decline to answer some questions this will not affect your treatment in any way.

This is a consent form. It says in writing what I have just told you. Please sign and date it here to show that you have understood what I have told you and that you willingly agree to answer questions.

Signature_____ Date _____

B: We would also like to take 3 – 4 teaspoons of blood from you for future research into the causes of your illness. No names will be attached to the sample tubes.

The blood we take for this study will not be tested immediately; it will be frozen. There is no time limit to how long it will be stored. It may be tested for evidence of various infections including HIV; it may be tested for substances (cells, proteins, other body chemicals) which might make it easier for you to have become ill. It may also be tested for inherited factors which might make it more likely for you to have developed the illness you now have. These tests will help us to understand your and other patients' illness better and, perhaps, prevent others from getting it. We only need to take blood from you once. At the current time we cannot provide details of what will be looked at, as this is not yet known, but we give assurance that no research will be done without the approval of the Wits Human Research Ethics Committee as well as the applicable Research Ethics Committees at the analysis sites.

This is a consent form. It says in writing what I have just told you. Please sign and date it here to show that you have understood what I have told you and that you willingly agree to have your blood collected.

Signature_____ Date _____

Signature of witness _____Date _____
(in case verbal consent given)

For enquiries telephone: Tonicah Maphanga 011 489 9150
tonicah.maphanga@nhls.ac.za

February 2015

8. Appendix B: Ethics certificates

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Ms M Hayat, Prof M Ramsay & Dr J-T Brandenburg
School of Pathology
Department of Human Genetics
Sydney Brenner Institute for Medical Bioscience
University

E-mail: mahtaab.hayat@gmail.com

CC: Supervisor: Professor CG Mathew <Chris.G.Mathew@wits.ac.za>
and <HREC-Medical.ResearchOffice@wits.ac.za>

FROM: Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 23/10/2018

REF: *R14/49*

PROTOCOL NO: **M180997** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *The contribution of common genetic variants to breast cancer risk in South African Black populations*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to the Government funding of the University.

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Professor Michele Ramsay

CLEARANCE CERTIFICATE	**M121029**
PROJECT	WITS-INDEPTH Partnership: Genomic and Environment Risk Factors for Cardiometabolic Disease in Africans
INVESTIGATORS	Professor Michele Ramsay.
DEPARTMENT	School of Pathology/Division of Human Genetics
DATE CONSIDERED	26/10/2012
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 21/11/2012 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor :

DECLARATION OF INVESTIGATOR(S)
To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/~~We~~ fully understand the conditions under which I am/~~we~~ are authorized to carry out the abovementioned research and I/~~we~~ guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/~~we~~ undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

M. Ramsay 11 April 2013

UNIVERSITY OF THE
WITWATERSRAND
JOHANNESBURG

R14/49 Prof Michele Ramsay et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M180535

NAME: (Principal Investigator)	Prof Michele Ramsay et al
DEPARTMENT:	Sydney Brenner Institute for Molecular Bioscience MRC/Wits - Agincourt Unit, DPHRU, Dikgale HDSS, APHRC - Nairobi, CRUN - Nanoro, NHRC - Navrongo
PROJECT TITLE:	AWI-Gen Phase 2: Genomic and environmental risk factors for cardiometabolic disease in Africans
DATE CONSIDERED:	25/05/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	
APPROVED BY:	Professor CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	21/06/2018

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193. University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **May** and will therefore be due in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

M. Ramsay
Principal Investigator Signature

22 June 2018
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

R14/49 Prof Christopher Mathew et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M160807

NAME: (Principal Investigator)	Prof Christopher Mathew et al
DEPARTMENT:	Human Genetics Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Investigation of Genetic Risk Factors for African Breast, Cervical and Oesophageal Cancer
DATE CONSIDERED:	26/08/2016
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	21/10/2016

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/2nd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated,from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review in August and will therefore be due in the month August each year.

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

R14/49 Dr Elvira Singh et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1606103

NAME: (Principal Investigator)	Dr Elvira Singh et al
DEPARTMENT:	Medical Research Council National Health Laboratory Services University of the Witwatersrand
PROJECT TITLE:	Johannesburg Cancer Study
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Previously Dr Lara Stein, Cancer Case-Control Study (Previously M981119, M040445 and M090361)
SUPERVISOR:	
APPROVED BY:	Professor P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	09/12/2016

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in March and will therefore be due in the month of March each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES


TRREE
Zertifikat
Certificat
Certificado
Certificate
Promouvoir les plus hauts standards éthiques dans la protection des participants à la recherche biomédicale
Promoting the highest ethical standards in the protection of biomedical research participants
Certificat de formation - Training Certificate
Ce document atteste que - this document certifies that
Clinical Trials Centre
The University of Hong Kong
Mahtaab Hayat
a complété avec succès - has successfully completed
Introduction to Research Ethics
du programme de formation TRREE en évaluation éthique de la recherche
of the TRREE training programme in research ethics evaluation
Release Date: 2023/03/27
CID : BMk1138qxa
Professeur Dominique Sprumont
Coordinateur TRREE Coordinator
APPROVED BY
SIWF FMH
ISFM
Programmes de formation continue (2 crédits)
Continuing Education Programs (2 credits)
Foederatio Pharmaceutica Helvetiae
FPH
Programmes de formation continue
Continuing Education Programs
Programmes de formation postgraduée et continue
Ce programme est soutenu par - This program is supported by :
European and Developing Countries Clinical Trials Partnership (EDCTP) (www.edctp.org) - Swiss National Science Foundation (www.snf.ch) - Canadian Institutes of Health Research (http://www.cihr-irsc.gc.ca/e/2891.html) - Swiss Academy of Medical Science (SAMS/ASSM/SAMW) (www.samw.ch) - Commission for Research Partnerships with Developing Countries (www.kfpe.ch)
[REV : 20220217]

9. Appendix C: Plagiarism declaration

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Mahtaab Hayat (Student number: 603401) am a student registered for the degree of PhD in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: MHayat Date: 20 March 2023

10. Appendix D: Turn-It-In support letter and report

Research and Postgraduate Office
Faculty of Health Sciences
University of the Witwatersrand
Phillip V Tobias Health Sciences Building
29 Princess of Wales Terrace, Parktown
Johannesburg, South Africa, 2193

30 March 2023

RE: M HAYAT 603401: Turn-It-In report clarification for PhD Submission

Dear Colleagues,

This letter is in support of Ms M Hayat's PhD submission. The Turn-It-In Originality report for her thesis returned a similarity index of 30%. This is due to the inclusion of her original published manuscript as part of the thesis submission:

1. Hayat M, Chen WC, Brandenburg JT, Babb de Villiers C, Ramsay M and Mathew CG. (2021) Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations. JCO Glob Oncol. 7. https://doi.org/10.1200/GO.21.00089

Please refer to the originality report attached. The top two matches, accounting for 15% of the 30% of the similarity index, are all from her original work.

Yours sincerely,

Professor Christopher Mathew, FMedSci
Professor in Human Genetics
Sydney Brenner Institute for Molecular Bioscience
University of the Witwatersrand, South Africa
Email: christopher.mathew@kcl.ac.uk; chris.g.mathew@wits.ac.za

MHayat_phdthesis_28032023.pdf

ORIGINALITY REPORT

30%	24%	27%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	11%
2	Mahtaab Hayat, Wenlong Carl Chen, Jean-Tristan Brandenburg, Chantal Babb de Villiers, Michèle Ramsay, Christopher G. Mathew. "Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations", JCO Global Oncology, 2021 Publication	4%
3	core.ac.uk Internet Source	<1%
4	hdl.handle.net Internet Source	<1%
5	client.conf-manage.com Internet Source	<1%
6	open.uct.ac.za Internet Source	<1%
7	www.science.gov Internet Source	<1%

8	www.researchgate.net Internet Source	<1 %
9	aacrjournals.org Internet Source	<1 %
10	assets.researchsquare.com Internet Source	<1 %
11	profiles.stanford.edu Internet Source	<1 %
12	worldwidescience.org Internet Source	<1 %
13	kclpure.kcl.ac.uk Internet Source	<1 %
14	emea.support.illumina.com Internet Source	<1 %
15	scholar.sun.ac.za Internet Source	<1 %
16	Tara M. Friebel, Irene L. Andrulis, Judith Balmaña, Amie M. Blanco et al. " and pathogenic sequence variants in women of African origin or ancestry ", Human Mutation, 2019 Publication	<1 %
17	era.library.ualberta.ca Internet Source	<1 %

18	www.dtic.mil Internet Source	<1 %
19	Beatriz Guillen-Guio, Tamara Hernández-Beeftink, Itahisa Marcelino-Rodríguez, Héctor Rodríguez-Pérez et al. "Admixture mapping of asthma in southwestern Europeans with North African ancestry influences", American Journal of Physiology-Lung Cellular and Molecular Physiology, 2020 Publication	<1 %
20	www.nature.com Internet Source	<1 %
21	www.biorxiv.org Internet Source	<1 %
22	Submitted to King's College Student Paper	<1 %
23	"Handbook of Statistical Genomics", Wiley, 2019 Publication	<1 %
24	academic.oup.com Internet Source	<1 %
25	orca.cardiff.ac.uk Internet Source	<1 %
26	research-information.bris.ac.uk Internet Source	<1 %

27	discovery.ucl.ac.uk Internet Source	<1 %
28	Leila Dorling, Sara Carvalho, Jamie Allen, Anna González-Neira et al. "Breast Cancer Risk Genes — Association Analysis in More than 113,000 Women", New England Journal of Medicine, 2021 Publication	<1 %
29	archsix.com Internet Source	<1 %
30	"The Role of Genetics in Breast and Reproductive Cancers", Springer Science and Business Media LLC, 2010 Publication	<1 %
31	Jenna Lilyquist, Kathryn J. Ruddy, Celine M. Vachon, Fergus J. Couch. "Common Genetic Variation and Breast Cancer Risk—Past, Present, and Future", Cancer Epidemiology Biomarkers & Prevention, 2018 Publication	<1 %
32	ww5.komen.org Internet Source	<1 %
33	pdfs.semanticscholar.org Internet Source	<1 %
34	www.frontiersin.org Internet Source	<1 %

35	Cesar Fortes-Lima, Ezekia Mtetwa, Carina Schlebusch. "Africa, the Cradle of Human Diversity", Brill, 2022 Publication	<1 %
36	archive.org Internet Source	<1 %
37	Submitted to University of KwaZulu-Natal Student Paper	<1 %
38	repository.up.ac.za Internet Source	<1 %
39	uir.unisa.ac.za Internet Source	<1 %
40	www.kcl.ac.uk Internet Source	<1 %
41	library.unisel.edu.my Internet Source	<1 %
42	www.sabcs.org Internet Source	<1 %
43	Cooper Alastair Grace, João Luís Reis-Cunha, Daniel C. Jeffares. "Strong effects of parasite genotype on drug susceptibility in the Indian subcontinent", Cold Spring Harbor Laboratory, 2022 Publication	<1 %

44	Submitted to Karl Landsteiner Privatuniversität für Gesundheitswissenschaften GmbH Student Paper	<1 %
45	Mwiza Gideon Singini, Elvira Singh, Debbie Bradshaw, Wenlong Carl Chen et al. "HPV types 16/18 L1 E6 and E7 proteins seropositivity and cervical cancer risk in HIV-positive and HIV-negative black South African women", Infectious Agents and Cancer, 2022 Publication	<1 %
46	flex.flinders.edu.au Internet Source	<1 %
47	Submitted to University of Hong Kong Student Paper	<1 %
48	www.research.manchester.ac.uk Internet Source	<1 %
49	Guimin Gao, Fangyuan Zhao, Thomas U. Ahearn, Kathryn L. Lunetta et al. "Polygenic Risk Scores for Prediction of Breast Cancer Risk in Women of African Ancestry: a Cross-Ancestry Approach", Cold Spring Harbor Laboratory, 2021 Publication	<1 %
50	Shaun Aron, Paballo Abel Chauke, Verena Ras, Sumir Panji, Katherine Johnston, Nicola Mulder. "The Development of a Sustainable	<1 %

	Bioinformatics Training Environment Within the H3Africa Bioinformatics Network (H3ABioNet)", Frontiers in Education, 2021 Publication	
51	Submitted to Laureate Higher Education Group Student Paper	<1 %
52	Submitted to University of Hertfordshire Student Paper	<1 %
53	Wenlong Carl Chen, Robyn Kerr, Andrew May, Babongile Ndlovu et al. "The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term Storage at −30°C", Biopreservation and Biobanking, 2018 Publication	<1 %
54	www.hirsla.lsh.is Internet Source	<1 %
55	Morris, John Allan. "The Genetic and Epigenetic Determinants of Osteoporosis", McGill University (Canada), 2021 Publication	<1 %
56	Julie R. Palmer, Christine B. Ambrosone, Andrew F. Olshan. "A collaborative study of the etiology of breast cancer subtypes in African American women: the AMBER consortium", Cancer Causes & Control, 2013 Publication	<1 %

57	Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush et al. "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes", The American Journal of Human Genetics, 2018 Publication	<1 %
58	Palwende Romuald Boua, Jean-Tristan Brandenburg, Ananyo Choudhury, Scott Hazelhurst et al. "Novel and Known Gene-Smoking Interactions With cIMT Identified as Potential Drivers for Atherosclerosis Risk in West-African Populations of the AWI-Gen Study", Frontiers in Genetics, 2020 Publication	<1 %
59	Brehima Diakite, Yaya Kassogue, Guimogo Dolo, Jun Wang et al. "p.Arg72Pro polymorphism of P53 and Breast Cancer Risk: A meta-analysis of case-control studies", Research Square Platform LLC, 2019 Publication	<1 %
60	Newcombe, Paul J., David V. Conti, and Sylvia Richardson. "JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects", Genetic Epidemiology, 2016. Publication	<1 %
61	Sung, Yi-Chen Vanessa. "Mechanisms of MET-Dependent Tumour Initiation and Progression", McGill University (Canada), 2021	<1 %

	Publication	
62	estia.hua.gr Internet Source	<1%
63	repositories.lib.utexas.edu Internet Source	<1%
64	scholar.ufs.ac.za Internet Source	<1%
65	Aurina Arnatkeviciute, Ben D. Fulcher, Mark A. Bellgrove, Alex Fornito. "Where the genome meets the connectome: Understanding how genes shape human brain connectivity", NeuroImage, 2021 Publication	<1%
66	Godfred Agongo, Lucas Amenga-Etego, Engelbert A. Nonterah, Cornelius Debpuur et al. "Candidate Gene Analysis Reveals Strong Association of CETP Variants With High Density Lipoprotein Cholesterol and PCSK9 Variants With Low Density Lipoprotein Cholesterol in Ghanaian Adults: An AWI-Gen Sub-Study", Frontiers in Genetics, 2020 Publication	<1%
67	Jeanne P. Uyisenga, Karin Segers, Aimé Z. Lumaka, Pacifique Mugenzi et al. "Screening of germline mutations in young Rwandan patients with breast cancers", Molecular Genetics & Genomic Medicine, 2020	<1%

	Publication	
68	Tatiane Yanes, Mary-Anne Young, Bettina Meiser, Paul A. James. "Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field", Breast Cancer Research, 2020 Publication	<1 %
69	jultika.oulu.fi Internet Source	<1 %
70	s3-eu-west-1.amazonaws.com Internet Source	<1 %
71	Jian Yang, Jian Zeng, Michael E Goddard, Naomi R Wray, Peter M Visscher. "Concepts, estimation and interpretation of SNP-based heritability", Nature Genetics, 2017 Publication	<1 %
72	Michael N Okobia. "Cytochrome P450 1B1 Val432Leu polymorphism and breast cancer risk in Nigerian women: a case control study", Infectious Agents and Cancer, 2009 Publication	<1 %
73	Submitted to University of Dundee Student Paper	<1 %
74	helda.helsinki.fi Internet Source	<1 %
75	karuspace.karu.ac.ke Internet Source	

		<1 %
76	Belbin, Gillian Morven. "Leveraging Population Genomics for Biomedical Discovery in a Healthcare System.", Icahn School of Medicine at Mount Sinai, 2017 Publication	<1 %
77	Carrie Printz. "Study to examine genetic factors behind increased risk of breast cancer in African Americans", Cancer, 2017 Publication	<1 %
78	Cerani, Agustin. "Combining Mendelian Randomization and Metabolomics-Based Case-Control Studies for the Identification of Metabolic Biomarkers of Disease", McGill University (Canada), 2021 Publication	<1 %
79	Francisco C. Ceballos, Scott Hazelhurst, David W. Clark, Godfred Agongo et al. "Autozygosity influences cardiometabolic disease-associated traits in the AWI-Gen sub-Saharan African study", Nature Communications, 2020 Publication	<1 %
80	Ray-Jones, Helen. "Defining the key biological and genetic mechanisms involved in psoriasis", The University of Manchester (United Kingdom), 2020 Publication	<1 %

81	Spear, Melissa. "The Effect of Genetic Ancestry on the Genetic Architecture of Complex Traits in Admixed Populations", University of California, San Francisco, 2020 Publication	<1 %
82	d-nb.info Internet Source	<1 %
83	purehost.bath.ac.uk Internet Source	<1 %
84	repub.eur.nl Internet Source	<1 %
85	www.nitrc.org Internet Source	<1 %
86	"Genetic Epidemiology", Springer Science and Business Media LLC, 2018 Publication	<1 %
87	Coignet, Marie V.. "Analysis of Reproductive Aging in African American Women A Replication Study & Exome Array Analysis in the Amber Consortium.", State University of New York at Buffalo, 2018 Publication	<1 %
88	Flint, Jonathan; Greenspan, Ralph J; Kendler, Kenneth S.. "How Genes Influence Behavior", OUP Oxford, 2020 Publication	<1 %

89	Sergey A. Kornilov, Elena L. Grigorenko. "Molecular Genetics Methods for Developmental Scientists", Wiley, 2016 Publication	<1 %
90	Submitted to University College London Student Paper	<1 %
91	Submitted to University of Mauritius Student Paper	<1 %
92	Yoanna S Pumpalova, Oluwatosin A Ayeni, Wenlong Carl Chen, Ines Buccimazza et al. "The Impact of Breast Cancer Treatment Delays on Survival Among South African Women", The Oncologist, 2022 Publication	<1 %
93	biblio.ugent.be Internet Source	<1 %
94	dokumen.pub Internet Source	<1 %
95	eprints.maynoothuniversity.ie Internet Source	<1 %
96	ir.canterbury.ac.nz Internet Source	<1 %
97	Abram B. Kamiza, Sounkou M. Toure, Marijana Vujkovic, Tafadzwa Machipisa et al. "Transferability of genetic risk scores in African populations", Nature Medicine, 2022	<1 %

	Publication	
98	Feng, Ye. "Post-Gwas Methods in Large Scale Studies of Breast Cancer in African Americans.", University of Southern California, 2019 Publication	<1 %
99	Isabelle Romero Novelli. "Nutritional factors and metabolic biomarkers: impact in risk and survival of women with breast cancer", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2022 Publication	<1 %
100	Kara N Maxwell, Katherine L Nathanson. "Common breast cancer risk variants in the post-COGS era: a comprehensive review", Breast Cancer Research, 2013 Publication	<1 %
101	Li, Jingmei. "Genetic Determinants of Breast Cancer Risk.", Karolinska Institutet (Sweden), 2021 Publication	<1 %
102	Panarella, Michela . "Extending Rare Variant Association Tests in Affected Siblings to Account for Background Genetic Risk.", University of Toronto (Canada), 2019 Publication	<1 %
	Submitted to University of Cambridge	

103	Student Paper	<1%
104	"Public Health Assistance to Sub-Saharan Africa: An Overview.", Points of View, 2016 Publication	<1%
105	Chang, Chi-Dan Vicky. "Investigating the Association Between Iron and Breast Cancer Risk and Factors Associated with Iron Status Among Canadian Women", University of Toronto (Canada), 2020 Publication	<1%
106	Clare Turnbull, Nazneen Rahman. "Genetic Predisposition to Breast Cancer: Past, Present, and Future", Annual Review of Genomics and Human Genetics, 2008 Publication	<1%
107	Guimin Gao, Fangyuan Zhao, Thomas U Ahearn, Kathryn L Lunetta et al. "Polygenic Risk Scores for Prediction of Breast Cancer Risk in Women of African Ancestry: a Cross-Ancestry Approach", Human Molecular Genetics, 2022 Publication	<1%
108	Jean-Tristan Brandenburg, Lindsay Clark, Gerrit Botha, Sumir Panji, Shakuntala Baichoo, Christopher Fields, Scott Hazelhurst. "H3AGWAS: a portable workflow for genome	<1%

	wide association studies", BMC Bioinformatics, 2022 Publication	
109	Jiaping Chen, Yue Jiang, Xiaoan Liu, Zhenzhen Qin et al. "Genetic variants at chromosome 9p21, 10p15 and 10q22 and breast cancer susceptibility in a Chinese population", Breast Cancer Research and Treatment, 2011 Publication	<1%
110	Submitted to Oxford Brookes University Student Paper	<1%
111	aran.library.nuigalway.ie Internet Source	<1%
112	Submitted to CSU, Fullerton Student Paper	<1%
113	Dewald Eygelaar, Elizabeth J. van Rensburg, Fourie Joubert. "Germline sequence variants contributing to cancer susceptibility in South African breast cancer patients of African ancestry", Scientific Reports, 2022 Publication	<1%
114	JULIA J. RUCKLIDGE. "Psychiatric, Psychosocial, and Cognitive Functioning of Female Adolescents With ADHD", Journal of the American Academy of Child & Adolescent Psychiatry, 05/2001 Publication	<1%

115	Mefford, Joel. "Applications of Cross-Validated Genetic Predictions from Linear Mixed Models and Other Results in Statistical Genetics.", University of California, San Francisco Publication	<1 %
116	Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, Mark J. Daly. "Clinical use of current polygenic risk scores may exacerbate health disparities", Nature Genetics, 2019 Publication	<1 %
117	Babongile C. Ndlovu, Mazvita Sengayi - Muchengeti, Caradee Y. Wright, Wenlong C. Chen, Lazarus Kuonza, Elvira Singh. "Skin cancer risk factors among Black South Africans—The Johannesburg Cancer Study, 1995–2016", Immunity, Inflammation and Disease, 2022 Publication	<1 %
118	Emmanuel R. Ezeome, King-David T. Yawe, Omobolaji Ayandipo, Olawale Badejo et al. "The African Female Breast Cancer Epidemiology Study Protocol", Frontiers in Oncology, 2022 Publication	<1 %
119	Ly, Reynold. "Metformin Pharmacogenomics: Functional Genomics of CDC25b and TPD54 Regulation of Metformin Anticancer Response	<1 %

	in Breast Cancer.", College of Medicine - Mayo Clinic, 2018 Publication	
120	Paulo C.M. Lyra-Junior, Nayara G. Tessarollo, Isabella S. Guimaraes, Taciane B. Henriques et al. "Chapter 5 GWAS in Breast Cancer", IntechOpen, 2017 Publication	<1 %
121	Submitted to Pima Community College Student Paper	<1 %
122	Submitted to University of Keele Student Paper	<1 %
123	Submitted to University of Stellenbosch, South Africa Student Paper	<1 %
124	Submitted to University of Sydney Student Paper	<1 %
125	Yu Peng, Fubin Liu, Yating Qiao, Peng Wang, Han Du, Changyu Si, Xixuan Wang, Kexin Chen, Fangfang Song. "Genetically Modified Circulating Levels of Advanced Glycation End-Products and Their Soluble Receptor (AGEs-RAGE Axis) with Risk and Mortality of Breast Cancer", Cancers, 2022 Publication	<1 %
126	Yumna Khan, Najeeb Ullah Khan, Ijaz Ali, Samiullah Khan, Aakif Ullah Khan, Aqib Iqbal,	<1 %

Brian D Adams. "Significant association ofBRCA1(rs1799950),BRCA2(rs144848) andTP53(rs1042522) polymorphism with breast cancer risk in Pashtun population of Khyber Pakhtunkhwa, Pakistan", Research Square Platform LLC, 2023
Publication

127	pure.rug.nl Internet Source	<1 %
128	www.tandfonline.com Internet Source	<1 %
129	Anderson, Kevin Michael. "Integrative Molecular Bases of Brain Function and Organization", Yale University, 2021 Publication	<1 %
130	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. "Evaluation of 19 susceptibility loci of breast cancer in women of African ancestry", Carcinogenesis, 2012. Publication	<1 %
131	Joshy George, Kylie L. Gorringe, Gordon K. Smyth, David D. L. Bowtell. "Chapter 2 Identifying Associations Between Genomic Alterations in Tumors", Springer Science and	<1 %

Business Media LLC, 2013
Publication

132	Lawal AbdulRazzaq Oluwagbemiga, Atoyebi Oluwole, Adesunkanmi AbdulRasheed Kayode. "Seventeen years after BRCA1: what is the BRCA mutation status of the breast cancer patients in Africa? – a systematic review", SpringerPlus, 2012 Publication	<1 %
133	Monica McGrath. "Novel breast cancer risk alleles and endometrial cancer risk", International Journal of Cancer, 12/15/2008 Publication	<1 %
134	Pawan Kumar Raghav, Aditya Raghav, Anjali Lathwal, Archit Saxena, Zoya Mann, Manisha Sengar, Raja Rajalingam. "Experimental and clinical data analysis for identification of COVID-19 resistant ACE2 mutations", Scientific Reports, 2023 Publication	<1 %
135	Pere Gelabert, Susanna Sawyer, Anders Bergström, Ashot Margaryan et al. "Genome-scale sequencing and analysis of human, wolf, and bison DNA from 25,000-year-old sediment", Current Biology, 2021 Publication	<1 %
136	Tani Sagna, Elena Bonora, Marie Nabonswindé Lamoussa Ouedraogo, Daniela	<1 %

	Fusco et al. "Identification of BRCA1/2 p.Ser1613Gly, p.Pro871Leu, p.Lys1183Arg, p.Glu1038Gly, p.Ser1140Gly, p.Ala2466Val, p.His2440Arg variants in women under 45 years old with breast nodules suspected of having breast cancer in Burkina Faso", Biomolecular Concepts, 2019 Publication	
137	www.af.ac.me Internet Source	<1%
138	www.liteforexsa.com Internet Source	<1%
139	"Breast Imaging", Springer Science and Business Media LLC, 2022 Publication	<1%
140	Ani Manichaikul, Lauren C. Peres, Xin - Qun Wang, Mollie E. Barnard et al. "Identification of novel epithelial ovarian cancer loci in women of African ancestry", International Journal of Cancer, 2019 Publication	<1%
141	Ava Willoughby, Paul Andreassen, Amanda Toland. "Genetic Testing to Guide Risk-Stratified Screens for Breast Cancer", Journal of Personalized Medicine, 2019 Publication	<1%

142	David L Jaye. "Expression of the plasmacytoid dendritic cell marker BDCA-2 supports a spectrum of maturation among CD4+ CD56+ hematodermic neoplasms", Modern Pathology, 12/2006 Publication	<1 %
143	Gileta, Alexander Frank. "Genetic Analyses of Pavlovian Conditioning in Outbred Models.", The University of Chicago Publication	<1 %
144	Hall, Elizabeth Scarlett. "Applying Polygenic Models to Disentangle Genotype-Phenotype Associations across Common Human Diseases", University of Toronto (Canada), 2021 Publication	<1 %
145	Lachlan T. Strike, Baptiste Couvy-Duchesne, Narelle K. Hansell, Gabriel Cuellar-Partida, Sarah E. Medland, Margaret J. Wright. "Genetics and Brain Morphology", Neuropsychology Review, 2015 Publication	<1 %
146	Melissa Henry, Raphaële Harvey, Lawrence M. Chen, Michael Meaney et al. "Genetic predisposition to depression and inflammation impacts symptom burden and survival in patients with head and neck	<1 %

cancer: A longitudinal study", Journal of Affective Disorders, 2023
Publication

147	Saeideh Torabi Dalivandan, Jasmine Plummer, Simon A. Gayther. "Risks and Function of Breast Cancer Susceptibility Alleles", Cancers, 2021 Publication	<1 %
148	Turkowski, Kari Leah. "Genetic Architecture, Polygenic Risk Scores, and Artificial Intelligence Guided Risk Stratification in Congenital Long Qt Syndrome", College of Medicine - Mayo Clinic, 2021 Publication	<1 %
149	Yosuke Kawai, Yusuke Watanabe, Yosuke Omae, Reiko Miyahara et al. "Exploring the genetic diversity of the Japanese Population: Insights from a Large-Scale Whole Genome Sequencing Analysis", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
150	Zhang, Jian. "Unraveling the Genetics of Cancer Using Whole-Exome Sequencing.", McGill University (Canada), 2021 Publication	<1 %
151	citeseerx.ist.psu.edu Internet Source	<1 %

152	diposit.ub.edu Internet Source	<1 %
153	eprints.whiterose.ac.uk Internet Source	<1 %
154	open.library.ubc.ca Internet Source	<1 %
155	patents.justia.com Internet Source	<1 %
156	www.karger.com Internet Source	<1 %
157	(2-10-14) http://152.98.160.29/contents/p/staff/Berndt_etal_NatG Internet Source	<1 %
158	A Cecile J W Janssens. "Validity of polygenic risk scores: are we measuring what we think we are?", Human Molecular Genetics, 2019 Publication	<1 %
159	Dron, Jacqueline S.. "The Genetic Determinants of Complex Lipid and Lipoprotein Phenotypes", The University of Western Ontario (Canada), 2022 Publication	<1 %
160	Frank Qian, Ye Feng, Yonglan Zheng, Temidayo O. Ogundiran et al. "Genetic variants in microRNA and microRNA biogenesis pathway genes and breast cancer	<1 %

risk among women of African ancestry", Human Genetics, 2016
Publication

161	Gleason, Kevin J.. "Methods for Integrative Multi-Omics Association Analysis Using Summary Statistics.", The University of Chicago, 2020 Publication	<1 %
162	Submitted to Intercollege Student Paper	<1 %
163	Issofou Njifen. "The impact of health capital investments on mortality from cardiovascular disease, cancer and diabetes in Africa: a pooling data analysis", Research Square Platform LLC, 2022 Publication	<1 %
164	Joshua Hoffman, Laura Fejerman, Donglei Hu, Scott Huntsman et al. "Identification of novel common breast cancer risk variants at the 6q25 locusamong Latinas", Breast Cancer Research, 2019 Publication	<1 %
165	Kathleen Boesze-Battaglia, Hongman Song, Maxim Sokolov, Concepcion Lillo et al. " The Tetraspanin Protein Peripherin-2 Forms a Complex with Melanoregulin, a Putative Membrane Fusion Regulator ", Biochemistry, 2007	<1 %

Publication

166	Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley et al. "Association analysis identifies 65 new breast cancer risk loci", Nature, 2017 Publication	<1 %
167	M. Ramsay, N. Crowther, E. Tambo, G. Agongo et al. "H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries", Global Health, Epidemiology and Genomics, 2016 Publication	<1 %
168	Nicole van der Merwe, Armand Peeters, Fredrieka Pienaar, Juanita Bezuidenhout, Susan van Rensburg, Maritha Kotze. "Exome Sequencing in a Family with Luminal-Type Breast Cancer Underpinned by Variation in the Methylation Pathway", International Journal of Molecular Sciences, 2017 Publication	<1 %
169	Paquet, eric Raphael. "Towards Individualized Assessment of Patient Benefit from Treatment and Prognosis in Breast Cancer.", McGill University (Canada), 2021 Publication	<1 %

170	Patel, Sejal. "Integrating Neuroimaging and Genetics Data in a Bioinformatics-Based Approach for the Study of Alzheimer's Disease.", University of Toronto (Canada), 2019 Publication	<1 %
171	Pereira Goncalves, Ana Cristina. "Oxidative Stress versus Epeginetics: Role in Susceptibility, Development, and Progression of Myeloid Neoplasms.", Universidade de Coimbra (Portugal), 2020 Publication	<1 %
172	Rubel, Meagan A.. "Effects of Diet and Parasites on The Gut Microbiota of Diverse Sub-Saharan Africans.", University of Pennsylvania, 2020 Publication	<1 %
173	Sarah Nadeau, Christian W. Thorball, Roger Kouyos, Huldrych F. Günthard et al. "A phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits", Cold Spring Harbor Laboratory, 2021 Publication	<1 %
174	Sawyer, Elinor, Rebecca Roylance, Christos Petridis, Mark N. Brook, Salpie Nowinski, Efterpi Papouli, Olivia Fletcher, Sarah Pinder, Andrew Hanby, Kelly Kohut, Patricia Gorman,	<1 %

Michele Caneppele, Julian Peto, Isabel dos Santos Silva, Nichola Johnson, Ruth Swann, Miriam Dwek, Katherine-Anne Perkins, Cheryl Gillett, Richard Houlston, Gillian Ross, Paolo De Ieso, Melissa C. Southey, John L. Hopper, Elena Provenzano, Carmel Apicella, Jelle Wesseling, Sten Cornelissen, Renske Keeman, Peter A. Fasching, Sebastian M. Jud, Arif B. Ekici, Matthias W. Beckmann, Michael J. Kerin, Federick Marme, Andreas Schneeweiss, Christof Sohn, Barbara Burwinkel, Pascal Guénel, Therese Truong, Pierre Laurent-Puig, Pierre Kerbrat, Stig E. Bojesen, Børge G. Nordestgaard, Sune F. Nielsen, Henrik Flyger, Roger L. Milne, Jose Ignacio Arias Perez, Primitiva Menéndez, Javier Benitez, Hermann Brenner, Aida Karina Dieffenbach, Volker Arndt, Christa Stegmaier, Alfons Meindl, Peter Lichtner, Rita K. Schmutzler, Magdalena Lochmann, Hiltrud Brauc. "Genetic Predisposition to In Situ and Invasive Lobular Carcinoma of the Breast", PLoS Genetics, 2014.
Publication

---

175 Sylvanus Toikumo, Rachel Vickers-Smith, Zeal Jinwala, Heng Xu et al. "The genetic architecture of pain intensity in a sample of 598,339 U.S. veterans", Cold Spring Harbor Laboratory, 2023 <1 %

	Publication	
176	Wenlong C. Chen, Jean-Tristan Brandenburg, Ananyo Choudhury, Mahtaab Hayat et al. "Abstract 31: Genome-wide association study of African esophageal squamous cell carcinoma", Cancer Research, 2021 Publication	<1 %
177	Yazhen Chen, Fangmeng Fu, Yuxiang Lin, Lin Qiu, Minjun Lu, Jiantang Zhang, Wei Qiu, Peidong Yang, Na Wu, Meng Huang, Chuan Wang. "The precision relationships between eight GWAS-identified genetic variants and breast cancer in a Chinese population", Oncotarget, 2016 Publication	<1 %
178	arrow.tudublin.ie Internet Source	<1 %
179	asgct.org Internet Source	<1 %
180	bura.brunel.ac.uk Internet Source	<1 %
181	cancerbrc.org Internet Source	<1 %
182	docplayer.com.br Internet Source	<1 %
183	etheses.bham.ac.uk Internet Source	

		<1 %
184	link.springer.com Internet Source	<1 %
185	lucris.lub.lu.se Internet Source	<1 %
186	oncology.medicinematters.com Internet Source	<1 %
187	opinvisindi.is Internet Source	<1 %
188	orbi.uliege.be Internet Source	<1 %
189	repositorio.ufmg.br Internet Source	<1 %
190	repository.ju.edu.et Internet Source	<1 %
191	s4be.cochrane.org Internet Source	<1 %
192	serval.unil.ch Internet Source	<1 %
193	www.wjgnet.com Internet Source	<1 %
194	"Cancer Epidemiology", Springer Science and Business Media LLC, 2009	<1 %

Publication

195	"Cellular Biology of the Endoplasmic Reticulum", Springer Science and Business Media LLC, 2021 Publication	<1 %
196	"The 2018 Annual Meeting of the International Genetic Epidemiology Society", Genetic Epidemiology, 2018 Publication	<1 %
197	Bruno César Feltes, Joice de Faria Poloni, Itamar José Guimarães Nunes, Sara Socorro Faria, Marcio Dorn. "Multi-Approach Bioinformatics Analysis of Curated Omics Data Provides a Gene Expression Panorama for Multiple Cancer Types", Frontiers in Genetics, 2020 Publication	<1 %
198	Chan, Ada Jyn-Sa. "Investigation of Genetic and Clinical Heterogeneity in Subtypes of Autism Spectrum Disorder", University of Toronto (Canada), 2020 Publication	<1 %
199	Charles, Bashira A., Daniel Shriner, and Charles N. Rotimi. "Accounting for Linkage Disequilibrium in Association Analysis of Diverse Populations : Leveraging Linkage Disequilibrium", Genetic Epidemiology, 2014. Publication	<1 %

200	Choi, Jihoon. "Bioinformatics Approaches for Investigating Missing Heritability in Complex Traits", Queen's University (Canada), 2021 Publication	<1 %
201	Clare Turnbull, Sarah Hines, Anthony Renwick, Deborah Hughes et al. "Mutation and association analysis of GEN1 in breast cancer susceptibility", Breast Cancer Research and Treatment, 2010 Publication	<1 %
202	Daniel W. Nebert, Ge Zhang, Elliot S. Vesell. "From Human Genetics and Genomics to Pharmacogenetics and Pharmacogenomics: Past Lessons, Future Directions", Drug Metabolism Reviews, 2008 Publication	<1 %
203	Dhriti Sengupta, Ananyo Choudhury, Cesar Fortes-Lima, Shaun Aron et al. "Genetic-substructure and complex demographic history of South African Bantu speakers", Cold Spring Harbor Laboratory, 2020 Publication	<1 %
204	Diego Barba, Ariana León-Sosa, Paulina Lugo, Daniela Suquillo-Yépez et al. "Breast Cancer, Screening and Diagnostic Tools: All You Need to Know", Critical Reviews in Oncology/Hematology, 2020 Publication	<1 %

205	Evans M Mathebula, Dhriti Sengupta, Nimmisha Govind, Vincent A Laufer et al. "A genome-wide association study for rheumatoid arthritis replicates previous HLA and non-HLA associations in a cohort from South Africa", Human Molecular Genetics, 2022 Publication	<1 %
206	Hang-Hang Luan, Li-Sha Luo, Zhi-Yan Lu. "Historical Trends in Incidence of Breast Cancer in Shanghai, Hong Kong and Los Angeles, 1973–2012: A Joinpoint and Age-Period-Cohort Analysis", International Journal of Public Health, 2021 Publication	<1 %
207	Jimmy Mullaert, Matthieu Bouaziz, Yoann Seeleuthner, Benedetta Bigio et al. "Taking population stratification into account by local permutations in rare - variant association studies on small samples", Genetic Epidemiology, 2021 Publication	<1 %
208	Kristy E. Driver, Honglin Song, Fabienne Lesueur, Shahana Ahmed et al. "Association of single-nucleotide polymorphisms in the cell cycle genes with breast cancer in the British population", Carcinogenesis, 2008 Publication	<1 %

209	Lihua Li. "TNRC9/LOC643714 polymorphisms are not associated with breast cancer risk in Chinese women", European Journal of Cancer Prevention, 08/2009 Publication	<1 %
210	Lona Durazo, Frida. "The Genetic Architecture of Pigmentation Traits in Modern Human Populations", University of Toronto (Canada), 2021 Publication	<1 %
211	Long, Jirong, Ben Zhang, Lisa B. Signorello, Qiuyin Cai, Sandra Deming-Halverson, Martha J. Shrubsole, Maureen Sanderson, Joe Dennis, Kyriaki Michailiou, Douglas F. Easton, Xiao-Ou Shu, William J. Blot, and Wei Zheng. "Evaluating Genome-Wide Association Study-Identified Breast Cancer Risk Variants in African-American Women", PLoS ONE, 2013. Publication	<1 %
212	Marc Donald Wilfried Adico, Abdou Azaque Zouré, Herman Karim Sombié, Touwendpoulimdé Isabelle Kiendrebeogo et al. " Involvement of ( ) and (,) polymorphisms of DNA repair genes in breast cancer occurrence in Burkina Faso ", Molecular Genetics & Genomic Medicine, 2023 Publication	<1 %

213	Mayosi, B.M.. "The burden of non-communicable diseases in South Africa", The Lancet, 20090912/18 Publication	<1 %
214	Michèle Ramsay, Nigel J. Crowther, Godfred Agongo, Stuart A. Ali et al. "Regional and sex-specific variation in BMI distribution in four sub-Saharan African countries: The H3Africa AWI-Gen study", Global Health Action, 2019 Publication	<1 %
215	Oldenburg, R.A.. "Genetic susceptibility for breast cancer: How many more genes to be found?", Critical Reviews in Oncology and Hematology, 200708 Publication	<1 %
216	Ostrom, Quinn T.. "Leveraging Demographic Differences in Incidence for Discovery and Validation of Risk Variants in Glioma.", Case Western Reserve University, 2020 Publication	<1 %
217	Qianqian Zhu, Lori Shepherd, Kathryn L. Lunetta, Song Yao et al. "Trans-ethnic follow-up of breast cancer GWAS hits using the preferential linkage disequilibrium approach", Oncotarget, 2016 Publication	<1 %

218	Raymond, Biaty. "Use of Whole-Genome Sequence Data for Genomic Prediction Across Populations and Species.", Wageningen University and Research, 2021 Publication	<1 %
219	Shetty, Priya Bhatia. "Genetic Epidemiology of Hypertension in Populations: Applications of Modified Methods.", Case Western Reserve University, 2020 Publication	<1 %
220	Sinziana Ionescu, Alin Codrut Nicolescu, Marian Marincas, Octavia-Luciana Madge, Laurentiu Simion. "An Update on the General Features of Breast Cancer in Male Patients—A Literature Review", Diagnostics, 2022 Publication	<1 %
221	Tapia-Ortiz, Edgar. "Reduction of Obesity Associated Breast Cancer Risk in a Phase II Clinical Trial of Metformin.", The University of Arizona, 2021 Publication	<1 %
222	Taraszka, Kodi Nicole. "Understanding the Genetic Architecture of Complex Traits through Meta-Analysis", University of California, Los Angeles, 2022 Publication	<1 %

223	Wei Zheng, Jirong Long, Yu-Tang Gao, Chun Li et al. "Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1", Nature Genetics, 2009 Publication	<1 %
224	Yangjing Zhao, Hui Wang, Jingdong Zhou, Qixiang Shao. "Glutathione Peroxidase GPX1 and Its Dichotomous Roles in Cancer", Cancers, 2022 Publication	<1 %
225	Yingying Yuan, Yachun Wang, Xiaodong Niu, Yungang Han, Wenbo Li, Meijin Cheng, Zheng Li, Jiao Tan, Yue Zhao, Wei Wang. "Association of lncRNA H19 polymorphisms with cancer susceptibility: An updated meta-analysis based on 53 studies", Frontiers in Genetics, 2022 Publication	<1 %
226	doczz.net Internet Source	<1 %
227	eprints.iums.ac.ir Internet Source	<1 %
228	eprints.qut.edu.au Internet Source	<1 %
229	escholarship.org Internet Source	<1 %

230	herkules.oulu.fi Internet Source	<1 %
231	kuscholarworks.ku.edu Internet Source	<1 %
232	opus.bibliothek.uni-wuerzburg.de Internet Source	<1 %
233	ourarchive.otago.ac.nz Internet Source	<1 %
234	pericles.pericles-prod.literatumonline.com Internet Source	<1 %
235	scholarship.claremont.edu Internet Source	<1 %
236	tel.archives-ouvertes.fr Internet Source	<1 %
237	theses.hal.science Internet Source	<1 %
238	theses.ncl.ac.uk Internet Source	<1 %
239	ukm.aju.airlinemeals.net Internet Source	<1 %
240	www.imperial.ac.uk Internet Source	<1 %
241	www.journals.ac.za Internet Source	<1 %

242	www.macs.hw.ac.uk Internet Source	<1 %
243	www.mdpi.com Internet Source	<1 %
244	www.medlib.am Internet Source	<1 %
245	www.medrxiv.org Internet Source	<1 %
246	www.nice.org.uk Internet Source	<1 %
247	www.repository.cam.ac.uk Internet Source	<1 %
248	www.science.org Internet Source	<1 %
249	www.thessgac.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

11. Appendix E: Declaration: Student's contribution to article and agreement of co-authors

I, **Mahtaab Hayat**, student number **603401**, declare that this Thesis/Dissertation/Research Report is my own work and that I contributed adequately towards research findings published in the article(s) stated below which are included in my Thesis/Dissertation/Research Report.

Signature of Student MHayat **Date**…03 March 2023

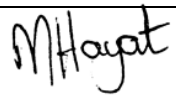
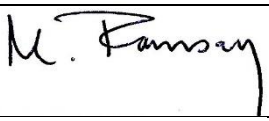
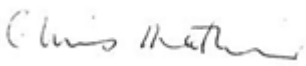
Name of Primary Supervisor Professor Christopher G Mathew

Signature of Primary Supervisor …… **Date**…03 March 2023……

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of his/**her Thesis**/Dissertation/Research Report. In cases where the student is not the 1st author of a published article, the primary supervisor must explain (under comments) why the student is entitled to use the paper for his/**her** degree purposes.

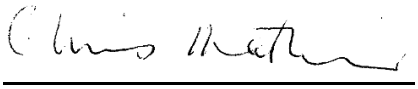
Article 1: Title: Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations

Journal name, year, volume and page numbers: JCO Global Oncology, 2021, volume 7, pg 1462-1471

Authors	Name	Signature	Date
1st author	Mahtaab Hayat		18/07/2022
2nd author	Wenlong Carl Chen		07/03/2023
3rd author	Jean-Tristan Brandenburg		06/03/2023
4th author	Chantal Babb de Villiers		19/07/2022
5th author	Michèle Ramsay		06/03/2023
6th author	Christopher G Mathew		06/03/2023

Comments by primary supervisor:

I confirm that the above article was conceived and written by Mahtaab Hayat and that she is the first author of the article. She is therefore fully entitled to include this in her PhD thesis.



Professor Christopher Mathew

12. Appendix F: Published article (included in chapter 2)

Check for updates

BREAST CANCER

review articles

Genetic Susceptibility to Breast Cancer in Sub-Saharan African Populations

Mahtaab Hayat, MSc[1,2]; Wenlong Carl Chen, MSc[1,3,4]; Jean-Tristan Brandenburg, PhD[1]; Chantal Babb de Villiers, PhD[2]; Michèle Ramsay, PhD[1]; and Christopher G. Mathew, PhD[1,2,5]

INTRODUCTION

Breast cancer (BC) is the most common cancer among African women and is the second most common cause of cancer deaths in Africa as estimated by GLOBOCAN 2018.[1] Cancer incidence is increasing in Africa[1] and is associated largely with lifestyle changes, resulting from increased urbanization.[2] Ferlay et al[3] predicted a 49% increase in BC incidence across Africa between 2018 and 2040, an increase faster than any other region in the world. The International Association of Cancer Registries cancer incidence report includes data for only 20 of 46 sub-Saharan African (SSA) countries,[4] and it is likely that the incidence of BC is under-reported owing to lack of infrastructure for cancer diagnosis in resource-constrained countries in SSA.[5] Also, owing to the large burden of communicable and other non-communicable diseases in SSA, most resources are directed at controlling these diseases.[6,7] Cancer research, especially molecular research, in SSA is often deprioritized against other pressing public health concerns.[8]

The epidemiology of BC is complex and the causes are multifactorial, with genetic risk factors and the environment playing important roles in cancer development.[9] Genetic risk factors contributing to the development of BC include rare pathogenic variants[10] or many common variants that, in combination, increase the risk of developing BC.[11] Familial or hereditary BC is associated with rare variants, which significantly increase the risk of developing BC and occur in genes such as *BRCA1*, *BRCA2*, *PTEN*, *FGFR2*, *CHEK2*, *TP53*, *PALB2*, *ATM*, and *XRCC2*. Hereditary BC accounts for only 5%-7% of all cases of BC.[2] The contribution of rare variants with moderate to high penetrance for BC risk has been reviewed recently by Nguyen and Thomas.[12] The other category is BC that occurs in the general population without evidence of a significant family history and results from a combination of genetic and environmental risk factors. The genetic risk is conferred by the presence of multiple genetic variants such as single-nucleotide polymorphisms (SNPs), which are relatively common in the population being studied, with allele frequencies of 1% or more.

Each variant individually confers only a small increase in risk of disease, but the presence of many such variants together significantly increases the risk of developing BC.[11,13] The contribution of common variants to BC risk has been investigated through the use of genome-wide association studies (GWASs), with at least 172 known genetic loci associated at genome-wide significance with the risk of BC.[14]

A majority of studies into the molecular etiology of BC have been performed in non-African populations, although there is substantial emerging literature on the genetics of BC in African-American populations.[15,16] However, published data on resident Black African populations from continental Africa are much more limited. This scoping review aims to summarize genetic studies on BC risk that have been conducted in resident Black SSA populations and to identify potential gaps in knowledge and research opportunities.

METHODS

The research question for this review was, *what knowledge exists on the association of germline genetic variants with familial or sporadic breast cancer in Black populations resident in sub-Saharan Africa?* This review follows the framework for scoping studies according to Arksey and O'Malley.[17] We chose to carry out a scoping review rather than a systematic review on the basis of the criteria described in the study by Munn et al.[18] A scoping review is more appropriate since there is a limited body of evidence in SSA, and our purpose was to scope the available evidence on BC genetics in SSA and to identify gaps in knowledge and limitations in methodologic approaches.

PubMed and SCOPUS were used to search the literature, and only original research articles written in English were included. All studies published on or before April 1, 2020, were included. Review articles were excluded. Only studies including data from SSA populations were included. A comprehensive combination of keywords was used to identify genetic studies on participants resident in SSA countries (Appendix Table A1).

The selection of appropriate studies was carried out according to the following criteria: original research articles that studied inherited variants associated with

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on September 1, 2021 and published at ascopubs.org/journal/go on October 8, 2021: DOI https://doi.org/10.1200/GO.21.00089

Creative Commons Attribution Non-Commercial No Derivatives 4.0 License

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

CONTEXT

Key Objective

Both rare and common genetic variants contribute to the risk of breast cancer (BC). Much is known about the genetics of BC in populations of European origin. Our objective was to review studies of genetic susceptibility to BC in the populations of sub-Saharan Africa to identify knowledge gaps and propose study designs to address these.

Knowledge Generated

A total of 22 studies were identified in 10 of 46 countries in sub-Saharan Africa. Rare variants predicted to be pathogenic were detected most commonly in the *BRCA1*, *BRCA2*, and *PALB2* genes. There were few studies of common genetic variants, none of which provided strong evidence for association with BC risk.

Relevance

Knowledge of genetic factors contributing to BC risk has important implications for patients and their families both for genetic counseling and for potential access to new therapies, which improve survival and reduce side effects.

BC in resident SSA populations. Screening of titles and abstracts for all papers was done independently by the lead authors (M.H. and W.C.C.), and papers were included on the basis of consensus agreements between them. If there was uncertainty on deciding whether a study was relevant, the full article was read to make a final decision. Studies of South African women of European, Indian, or mixed ancestry were excluded. If the abstract or full text to a study could not be accessed, it was excluded from this study. The following information was extracted: PMID, year of study, population, sample size, key findings (variants and genes), the predicted pathogenicity of variants, methodology, and a brief summary of the study. Rare variants were considered of interest if predicted to be deleterious by variant effect predictor tools and/or significantly enriched in BC cases compared with controls.

RESULTS

Using the comprehensive list of keywords to search, PubMed identified a total of 1,029 articles (Fig 1). Scanning the titles and the abstracts led to exclusion of 1,002 studies as they were not BC genetic studies, did not include samples from SSA, or were reviews, leaving 27 that met the broad inclusion criteria. A further four studies were excluded after full-text evaluation as they were conducted in African-American populations, and our aim was specifically to review studies of BC genetics in SSA populations. One study was removed as it only investigated tumor DNA. The search in SCOPUS did not yield any additional eligible articles.

The remaining 22 papers were included and assessed in this review.[19-40] Of these, 16 were rare variant studies,[19-34] and the remaining six investigated a few common variants in small sample sizes.[35-40] The studies used protein truncation tests, variant-specific polymerase chain reaction assays, sequencing of genes, exons, or coding regions, or gene panel sequencing to identify variants. There were no GWASs among the 22 papers. The studies in this review included populations from 10 of the 46 SSA countries:

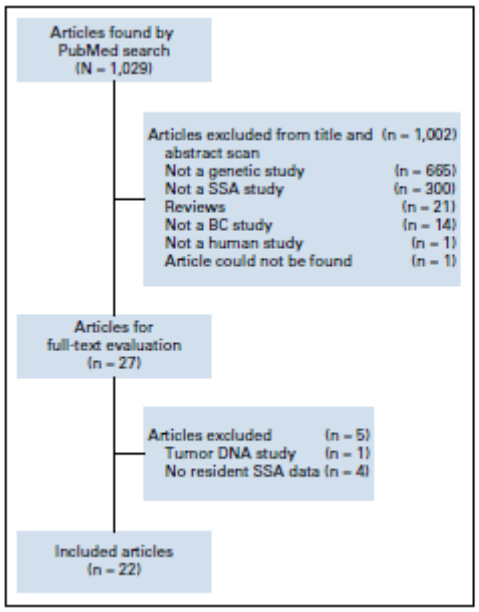


FIG 1. Process of selection of papers. The selection criteria were germline genetic studies focusing on BC and carried out in an SSA population(s). BC, breast cancer; SSA, sub-Saharan African.

Burkina Faso,[22,27,37] Cameroon,[21] Democratic Republic of Congo,[25] Ethiopia,[38] Kenya,[19] Nigeria,[23,28,30,32,33,35,36] Rwanda,[29,39,40] Senegal,[20,26,34] South Africa,[24,31] and Uganda.[21]

The Contribution of Rare Variants to BC Risk in SSA

Among the 22 studies identified, 16 were rare variant studies, which involved the screening of exons of genes of interest for variants with moderate to high penetrance and low population frequencies. A summary of the rare variant studies including countries of origin of participants and

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

sample sizes is shown in Table 1. Probable pathogenic variants, which have been detected in two or more patients, are listed in Table 2. Some of the rare variant studies (6 of 16) genotyped only a few specific variants, which had been identified previously in other populations. In 14 of the 16 studies, variants were identified in one or more known BC susceptibility genes, with sample sizes ranging from 1 to 1,136 patients with BC. The largest and most comprehensive of these, which built on previous studies in the Nigerian population, used the BROCA gene panel[41] to sequence the exons of 25 genes for variants in 1,136 Nigerian women with BC and in 997 Nigerian female controls.[23] Variants predicted to be deleterious were detected in 16 genes, including *BRCA1*, *BRCA2*, *PALB2*, and *TP53*. Deleterious variants were detected in 14.7% of patients with BC and in 1.8% of controls. Carriers of mutations in *BRCA1* and *TP53* had a significantly younger age at diagnosis than noncarriers, and *BRCA1* mutation carriers were more likely to have TNBC. The majority of patients for whom tumor stage was available were stage III (41.8%) or stage IV (42.6%), but no association of tumor stage with mutation status was reported. Four other previous studies of *BRCA1* and *BRCA2* in Nigerian patients with BC are summarized in Table 1, and mutations detected are listed in Table 2.

A South African study sequenced *BRCA1*, *BRCA2*, and *PALB2* from several ancestral groups, which included 85 South African Black patients with BC,[24] and found pathogenic variants in *BRCA1*, *BRCA2*, and *PALB2* in this ethnic subset. The group of patients with triple-negative breast cancer had a higher mutation frequency than the group of patients diagnosed before age 50 years. In another South African study, which included 16 women of Xhosa ancestry with BC, the same deleterious variant in *BRCA2* (p.Ile1924ArgTer) was detected in four of 16 patients and is suggested to be a founder variant in both Black and mixed ancestry populations in the Western Cape Province.[31] A study of patients with BC from Cameroon and Uganda sequenced 12 known or candidate BC genes in 196 cases and 185 controls unselected for age at diagnosis or family history[21] and identified pathogenic or likely pathogenic variants in 15.8% of BC cases and 1.6% of controls. Their findings included two patients with a mutation in *ATM* and single patient with mutations in *BARD1*, *CDH1*, and *TP53*. One of three studies in Senegal identified a duplication of 10 nucleotides in the *BRCA1* gene, c.815_824dup10, in 15 index cases from families with hereditary BC.[20] This duplication was located on the same microsatellite haplotype in all cases, had previously been reported in BC cases from France, Spain, and the United States, particularly in people of African or Hispanic descent, and is thus a founder mutation likely to be of West African origin. Whole-exome sequencing of 13 patients with BC in Kenya[19] detected five variants of uncertain clinical significance in *BRCA1* and *BRCA2* and a pathogenic variant in *BRCA2*. A study in the Democratic Republic of Congo detected a pathogenic *BRCA1* variant in all three affected members of a BC family.[25] However, sequencing of *BRCA1* and *BRCA2*

TABLE 1. Studies of Rare Variants in BC Genes in SSA

Methodology	Country	BC Cases	Controls	Variants Detected[a]	Reference
Whole-exome sequencing	Kenya	13	0	1	19
BRCA1/*BRCA2*, Sanger sequencing	Senegal	27	90	1	20
CHEK2 sequencing	Rwanda	41	42	0	29
BRCA1/*BRCA2* sequencing	Nigeria	434	0	29	30
Gene panel	South Africa (Xhosa)	16	0	1	31
Allele-specific PCR	Nigeria	365	117	1	32
SSCP and protein truncation test	Nigeria	70	0	3	33
BRCA2 exon sequencing	Senegal	33	96	1	34
Gene panel sequencing	Cameroon and Uganda	196	185	34	21
Sequencing of coding regions of genes	Burkina Faso	9	0	0	22
BROCA gene panel sequencing	Nigeria	1,136	611	106	23
Coding gene sequencing and MLPA	South Africa (Black)	85	0	6	24
BRCA1/*BRCA2* sequencing	DRC	3	0	1	25
BRCA1 sequencing	Senegal	1	0	1	26
BRCA1 sequencing (exons 2, 5, and 11)	Burkina Faso	15	0	0	27
SNAPSHOT and sequencing	Nigeria	356	0	12	28

Abbreviations: BC, breast cancer; MLPA, multiplex ligation–dependent probe amplification; PCR, polymerase chain reaction; SSA, sub-Saharan African; SSCP, single strand conformation polymorphism.

[a]No. of pathogenic variants detected.

 © 2021 by American Society of Clinical Oncology

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

TABLE 2. Predicted Pathogenic Variants Detected in Two or More African Patients

Gene	Variant	Variant Effect	No. of Cases	Country	Reference
BRCA1	c.80+1G>A	Splice site	2	Nigeria	23
BRCA1	c.133_134del	p.Lys45IleTer	5	Nigeria	23
BRCA1	c.191G>A	p.Cys64Tyr	3	Nigeria	30
BRCA1			4	Nigeria	23
BRCA1	c.303T>G	p.Tyr101Ter	4	Nigeria	28
BRCA1			10	Nigeria	23
BRCA1	c.815_824dup	p.Thr276Ter	15	Senegal	20
BRCA1	c.838dup	p.Ala280GlyTer	3	Nigeria	23
BRCA1	c.1504_1508del	p.Leu502AlaTer	2	Nigeria	30
BRCA1			3	Nigeria	23
BRCA1	c.1623dup	p.Asn542GluTer	3	Nigeria	23
BRCA1			4	Nigeria	30
BRCA1	c.2192_2196del	p.Lys731ArgTer	2	Nigeria	23
BRCA1	c.2329del	p.Tyr777MetTer	4	Nigeria	23
BRCA1	c.2864C>A	p.Ser955Ter	3	Nigeria	23
BRCA1	c.3268C>T	p.Gln1090Ter	2	Nigeria	30
BRCA1			4	Nigeria	23
BRCA1			2	Nigeria	33
BRCA1	c.3637_3638del	p.Glu1213ArgTer	2	Nigeria	23
BRCA1	c.4122_4123del	p.Ser1374ArgTer	3	Nigeria	30
BRCA1			3	Nigeria	23
BRCA1	c.4240dup	p.Leu1414ProTer	2	Nigeria	30
BRCA1	c.4484G>T	p.Arg1495Met	3	Cameroon	21
BRCA1	c.4986+6T>C	Splice site	2	Cameroon	21
BRCA1	c.4992C>T	Splice site	2	Nigeria	23
BRCA1	c.5095C>T	p.Arg1699Trp	2	Nigeria	23
BRCA1	c.5177_5180del	p.Arg1726LysTer	2	Nigeria	23
BRCA1	c.5324T>G	p.Met1775Arg	3	Nigeria	30
BRCA1			7	Nigeria	23
BRCA2	c.93G>A	p.Trp31Ter	2	Nigeria	23
BRCA2	c.1310_1313del	p.Lys437IleTer	3	Nigeria	30
BRCA2			4	Nigeria	23
BRCA2	c.1796_1800del	p.Ser599Ter	3	Nigeria	23
BRCA2	c.2402_2412del	p.Asn801IleTer	2	Nigeria	30
BRCA2			3	Nigeria	23
BRCA2	c.2808_2811del	p.Ala938ProTer	2	Nigeria	23
BRCA2	c.5241_5242insTA	p.Ser1748Ter	2	Nigeria	23
BRCA2	c.5286T>G	p.Tyr1762Ter	2	Nigeria	23
BRCA2	c.5771_5774del	p.Ile1924ArgTer	4	South Africa (Black)	31
BRCA2	c.8351+2T>C	Splice site	3	Nigeria	23
BRCA2	c.8817_8820del	p.Lys2939AsnTer	2	Nigeria	30
BRCA2	c.9875C>T	p.Pro3292Leu	2	South Africa (Black)	24
BARD1	c.1396-1G>A	p.Tyr101Ter	2	Nigeria	23
ATM	c.7913G>A	p.Trp2638Ter	2	Cameroon	21
PALB2	c.2485C>T	p.Gln829Ter	2	Nigeria	23
RAD51C	c.186_187del	p.Gln62HisTer	2	Nigeria	23

in patients with BC in Burkina Faso identified only benign variants in these genes[22,27] and a screen for germline variants in *CHEK2* in Rwandan BC cases and controls was negative.[29]

In summary, a total of 36 predicted pathogenic variants were detected in two or more patients with BC. These variants, together with their country of origin, are listed in Table 2. Some *BRCA1* variants were detected in up to 15 patients, and, as noted above, there are suggestive founder pathogenic variants in the Senegalese and South African populations. Overall, pathogenic variants were detected in at least two patients in the *ATM*, *BARD1*, *BRCA1*, *BRCA2*, *PALB2*, and *RAD51C* genes. The study by Zheng et al[23] in Nigeria suggests that variants in several other genes may also contribute to BC risk, but the mutations were too rare to draw firm conclusions on causality.

The number of predicted pathogenic variants found in each gene and the SSA country where the studies were conducted are summarized in Figure 2. The preponderance of variants of Nigerian origin reflects the much larger and more comprehensive variant studies that have been conducted in this population.[23]

The Contribution of Common Variants to BC Risk in SSA

Our search did not find any GWAS of BC performed exclusively in resident SSA populations. The genetic association studies carried out in SSA populations were candidate gene studies, which investigated a limited number of common variants previously associated with BC in other populations. These were genotyped in relatively small numbers of BC cases and controls and tested for association with BC and are summarized in Table 3. Okobia et al[36] genotyped a *CYP1B1* polymorphism Val432Leu (c.1294C>G) in BC cases and controls from Southern Nigeria and found association of the heterozygous Val/Leu genotype with BC (odds ratio [OR] = 1.59; 95% CI, 1.01 to 2.52). The same group genotyped the c.668A>G single-nucleotide polymorphism (SNP; Gln223Arg) in the leptin receptor gene (*LEPR*) and reported a suggestive association of the 223Arg allele with disease risk in premenopausal women (OR = 1.8; 95% CI, 1.0 to 3.2; $P = .07$).[35] A study of the glutathione S-transferase null variants in the *GSTM1* and *GSTT1* genes in cases and controls from Burkina Faso found no association of the *GSTM1 null* variant with BC, whereas the *GSTT1 null* allele was associated with an increase in BC risk (OR = 2.42; 95% CI, 1.17 to 5.02; $P = .01$).[37] Investigation of an initiation codon variant (Met1Arg, rs2228570) in the vitamin D receptor (*VDR*) gene in 392 Ethiopian BC cases and controls reported an association of the Arg/Arg (G/G) genotype with an increased risk of BC (OR = 1.44; 95% CI, 1.01 to 2.06; $P = .04$).[38] A lack of association of a functional SNP in the *GPX1* gene (p.Pro198Leu) and the *TP53* functional SNP (p.Pro72Arg) with BC was reported in the Rwandese population,[39,40] but the sample sizes were very small.

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

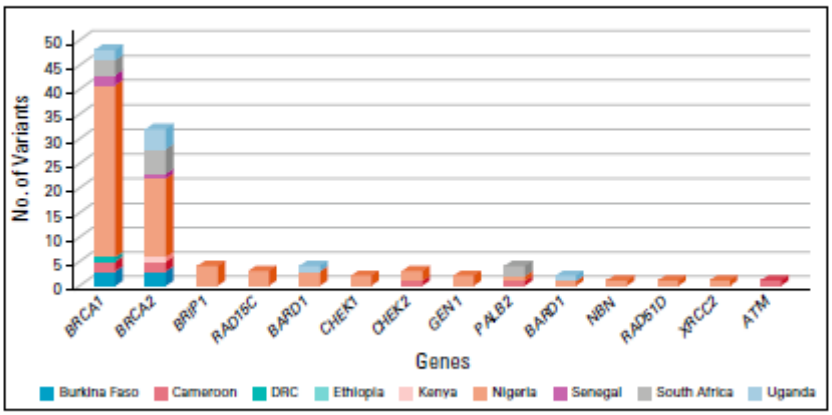


FIG 2. Variant distribution in genes and countries. This shows how many predicted pathogenic variants were identified in breast cancer susceptibility genes in each reported sub-Saharan African country. DRC, Democratic Republic of Congo.

DISCUSSION

Our initial PubMed and SCOPUS search yielded 1,029 articles from which 22 were selected for this review. These studies investigated the association of germline variants with BC in resident SSA populations. A large majority of studies focused on the detection and analysis of rare variants in known BC susceptibility genes in cases and, in some studies, also in controls. Our search did not find any GWAS for BC in resident African populations, but we did find a small number of association studies of candidate genes and common variants in cases and controls. Although the focus of our review was on sub-Saharan Africa, we discuss the findings below in the context of the extensive literature on BC genetics in African-American populations, which provide important insights relevant to continental Africa.

The studies included in this review either tested for specific variants known to be associated with BC risk in non-African populations or used DNA sequencing of the coding regions of known BC genes to explore the contributions of rare germline variants to the etiology of BC in their local populations. These germline variants are rare in the general population but have moderate to high penetrance, and thus, individuals with these variants are at substantially higher risk of developing BC compared with the general population.[42] Most of the studies focused on the key BC predisposition genes, *BRCA1* and *BRCA2*, which play important roles in DNA repair pathways.[43,44]

The variants of particular interest listed in Table 2 indicate that at least six genes, *BRCA1*, *BRCA2*, *PALB2*, *ATM*, *BARD1*, and *RAD51C*, collectively make a significant contribution to BC risk in the African populations that were studied. The variants disrupt normal protein function either by early termination of mRNA translation by the introduction of stop codons or by missense variants, which may affect key functional elements of the protein.[45] However, we note that no functional assays were performed in any of the studies reviewed here and all potential functional effects were predicted in silico. There is also the problem of variants of unknown significance involving missense variants

TABLE 3. Association Studies of Common Variants in SSA BC

Gene	Methodology	Variant	Country	BC Cases	Controls	ORs (95% CI)[a]	Reference
LEPR	PCR-RFLP	p.Gln223Arg	Nigeria	209	209	1.8 (1.0 to 3.2)	35
CYP1B1	PCR-RFLP	p.Leu432Val	Nigeria	250	250	1.59 (1.0 to 2.52)	36
GSTT1	PCR	Deletion (null)	Burkina Faso	80	100	2.42 (1.17 to 5.02)	37
VDR	PCR-RFLP	p.Met51Thr	Ethiopia	392	193	1.44 (1.01 to 2.06)	38
GPX1	Exon sequencing	p.Pro198Leu	Rwanda	41	42	Not reported[b]	39
TP53	Allele-specific PCR	p.Pro72Arg	Rwanda	40	39	Not reported[b]	40

Abbreviations: BC, breast cancer; OR, odds ratio; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; SSA, sub-Saharan African.

[a]ORs and 95% CIs for strongest results.

[b]No association with BC overall. No ORs reported.

 © 2021 by American Society of Clinical Oncology

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

in known BC susceptibility genes that fail current predictive algorithms for pathogenicity, which will be amplified by inclusion of diverse and understudied African genomes.[46] The strategy of, for example, Zheng et al,[23] in which genes were sequenced in large numbers of cases and controls, is helpful in this regard, as it demonstrated a significant excess of many variants in cases vs controls, but the cost of sequencing large numbers of samples may be unaffordable for many African researchers. The contribution of other genes in which potentially pathogenic variants were found is not yet clear as the number of cases with variants in each of these genes was very small. However, for some of these genes, such as *ATM*, there is strong epidemiologic and molecular evidence of causality in other populations.[47] Another question of interest is whether specific pathogenic variants are found in more than one African country or ethnic group, such as is the case for European populations. There is currently insufficient evidence to resolve this question in African populations, but it could be addressed by larger-scale studies, such as the Nigerian study of Zheng et al,[23] being carried out in other African populations.

All the genetic association studies for common variants which we identified involved genotyping a small number of SNPs in one or two candidate genes and in small sample sizes. Some marginally significant associations were reported, but such studies would have very low power to detect the typically small effect sizes for individual variants in genetic susceptibility to disease in a complex trait. A further difficulty with candidate gene studies is that they do not provide the means for correction of population structure in the samples, which is an important requirement for a robust genetic association study.

Although the abovementioned studies are limited in scope and of low statistical power, recently, there have been several well-powered genetic studies in BC cases and controls of African ancestry, not resident in Africa. For example, a meta-analysis of two GWASs and a replication study in women of African ancestry identified 3q26.21 as a novel susceptibility locus for estrogen receptor (ER)–negative BC.[36] This study involved genotyping of more than 6,657 BC cases and 7,713 controls on a genome-wide SNP array. BC risk variants identified in other populations were tested for association in another large study of women of African ancestry[48] and at least 12 regions were identified, which showed evidence of association in this population. These findings suggest that there is at least partial transferability of BC risk loci between different populations and ethnicities. This phenomenon was also observed for other cancers such as prostate and esophageal cancers where some but not all risk loci were transferable across African, European, and Asian populations.[49-51] These studies suggest that the GWAS approach may be fruitful for studying the genetics of BC in resident African populations.

Whole-genome sequencing of many ethnolinguistic groups in Africa has confirmed the extraordinary degree of genomic diversity across the African continent.[52] A comprehensive analysis of the contribution of germline variants to BC risk in continental Africa will therefore require studies in a wide range of populations and regions. Another important consideration is the study design, which will depend on whether the aim is to identify rare variants of moderate to high penetrance in genes such as *BRCA1* and *BRCA2*, likely associated with familial cancers, or to identify variants with low penetrance, which collectively make a large contribution to BC risk in the general population.

To identify rare and novel germline variants of intermediate to high penetrance, DNA sequencing, rather than genotyping of known risk variants across the regions and populations of Africa, will be required since the limited data available suggest extensive diversity of BC risk variants across the African continent. The study design should also include an equivalent set of well-matched controls, which, together with algorithms for predicting functional effects in silico, will help to identify variants that are likely to be pathogenic. The sequencing strategy will depend on available financial and bioinformatic resources. Whole-exome sequencing would allow interrogation of the entire coding genome but is still relatively costly and requires bioinformatic skills to call, annotate, and interpret variants. Sequencing a panel of genes known to be involved in BC susceptibility, such as the BROCA panel,[41] would reduce costs and simplify the bioinformatic analysis but only investigate genes already implicated in other populations. The probability of detecting pathogenic variants in such genes would be increased by enriching the BC samples for younger patients and those with a family history of BC, as these are groups in which such variants are most likely to be identified. Finally, a substantial sample size will be needed to obtain a robust assessment of the contribution of these rare variants to BC risk in local populations. An exemplar that includes these desirable features is the study by Zheng et al[23] in the Nigerian population. Larger sample sizes would also allow tests of whether germline mutations were associated with clinical features such as hormone receptor status, tumor stage, and survival. A large study of germline mutations in African-American women, for example, showed a twofold higher incidence of pathogenic mutations in ER-negative BC compared with ER-positive BC.[15] Detection of germline mutations in genes such as *BRCA1* and *BRCA2* has important implications for clinical care such as the risk of further cancers in the patient, genetic counseling to other family members, and potential access to new classes of drugs such as poly (ADP-ribose) polymerase inhibitors, which can improve survival and reduce side effects in mutation carriers.[53]

Regarding the detection of common risk variants for BC, small sample sizes such as those reported to date in SSA populations do not have the power to detect associations of

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

common variants with small effects. A shortcoming of most candidate gene studies is that a limited number of variants are genotyped, which do not capture the full genetic diversity of the gene and genome.[54] Also, genotyping a small number of variants does not allow correction for the population structure, which is particularly important in populations with a high degree of genetic diversity and substructure. This requires either a genome-wide study or at least a genotyping array containing thousands of common variants to detect and control for population structure. The route to significant discoveries of common BC risk variants in African populations is to carry out GWAS using SNP arrays, which are designed to capture a substantial component of African genetic diversity, and by genotyping thousands of cases and controls to achieve the statistical power required to detect variants with modest effect sizes. We are currently carrying out a GWAS using the H3Africa array on more than 2,700 BC cases and 1,100 controls from the Johannesburg Cancer Study[55] in the Black South African population as part of the ERICA-SA study,[56] which aims to identify evolving risk factors for cancer in African populations.

A major difficulty regarding the desirable study designs outlined above is the resource-intensive nature of molecular cancer research such as next-generation sequencing and genome-wide genotyping in terms of funding and appropriately skilled personnel, both of which are relatively scarce in Africa. Significant progress in improving our understanding of the genetics of BC in Africa will require a substantial investment from governments or other sources to generate the grant income and training programs needed to achieve this goal.

In conclusion, although some studies on the genetics of BC have been carried out in populations of African origin, few comprehensive large-scale studies have included populations resident in SSA. This scoping review has identified that more extensive studies of the contribution of both rare and common variants to BC risk are needed in a broader range of African countries. The knowledge gap creates an opportunity for African researchers to gain new insights into the molecular epidemiology of BC on the African continent.

Both rare and common genetic variants contribute to the risk of BC. A wealth of information exists on genes and variants associated with BC risk in populations of European origin, but knowledge of the genetics of BC susceptibility in SSA is limited. Our scoping review found relevant published data from only 10 of 46 SSA countries, and most studies focused on detection of rare variants in known BC susceptibility genes such as *BRCA1* and *BRCA2*. Few studies looked for association of common variants with BC in the general population, and these had neither the statistical power nor the scope to identify common risk variants for BC or to correct for population structure differences. No GWASs for BC have been performed in resident SSA populations. Thus, there is limited information on the contribution of variants to BC in most SSA countries. More extensive screening of BC gene panels for high-penetrance variants is needed in SSA, and GWASs of large case-control samples are required to identify common variants contributing to BC risk in SSA populations. This knowledge gap represents important and exciting opportunities for African cancer research.

AFFILIATIONS

[1]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa
[4]Non-communicable Diseases Research Division, Wits Health Consortium (PTY) Ltd, Johannesburg, South Africa
[5]Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom

CORRESPONDING AUTHOR

Christopher G. Mathew, PhD, Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, 9 Jubilee Rd, Parktown, Johannesburg 2193, South Africa; e-mail: christopher.mathew@kcl.ac.uk.

DISCLAIMER

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the South African Medical Research Council or the South African National Department of Health or the UK Medical Research Council from the UK Government's Newton Fund.

SUPPORT

Supported by the South African Medical Research Council (with funds received from the South African National Department of Health) and the UK Medical Research Council (with funds from the UK Government's Newton Fund; MRC-RFA-SHIP 01-2015).

AUTHOR CONTRIBUTIONS

Conception and design: Mahtasb Hayat, Wenlong Carl Chen, Michèle Ramsay, Christopher G. Mathew
Administrative support: Chantal Babb de Villiers
Collection and assembly of data: Mahtasb Hayat, Wenlong Carl Chen, Christopher G. Mathew
Data analysis and interpretation: Mahtasb Hayat, Wenlong Carl Chen, Jean-Tristan Brandenburg, Chantal Babb de Villiers, Christopher G. Mathew
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/go/authors/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Michèle Ramsay

Consulting or Advisory Role: Genentech/Roche

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

This project forms part of an international research program aimed at identifying evolving risk factors for cancer in African populations (ERICA-SA; http://www.mrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa).

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, et al: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394-424, 2018
2. Lukong KE, Ogunbolude Y, Kamdem JP: Breast cancer in Africa: Prevalence, treatment options, herbal medicines, and socioeconomic determinants. Breast Cancer Res Treat 166:351-365, 2017
3. Ferlay J, Colombet M, Soerjomataram I, et al: Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 144:1941-1953, 2019
4. Christopher PW, Weiderpass E, Bernard WS: World Cancer Report: Cancer Research for Cancer Prevention. Lyon, France, International Agency for Research on Cancer, 2020, pp 1-630
5. Cumber SN, Nchanji KN, Tsoka-Gwegweni JM: Breast cancer among women in sub-Saharan Africa: Prevalence and a situational analysis. South Afr J Gynaecol Oncol 9:35-37, 2017
6. Roth GA, Abate D, Abate KH, et al: Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet 392:1736-1788, 2018
7. Gouda HN, Charlson F, Sorsdahl K, et al: Burden of non-communicable diseases in sub-Saharan Africa, 1990-2017: Results from the Global Burden of Disease Study 2017. Lancet Glob Heal 7:e1375-e1387, 2019
8. Ngoma T, Ngoma M: Cancer control in Africa: Is cancer research a luxury or necessity? Ecancermedicalscience 13:947, 2019
9. Momenimovahed Z, Salehiniya H: Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer Targets Ther 11:151-164, 2019
10. Dorling L, Carvalho S, Allen J, et al: Breast cancer risk genes—Association analysis in more than 113,000 women. N Engl J Med 384:428-439, 2021
11. Maxwell KN, Nathanson KL: Common breast cancer risk variants in the post-COGS era: A comprehensive review. Breast Cancer Res 15:212, 2013
12. Nguyen JV, Thomas MH: Beyond BRCA: Review of hereditary syndromes predisposing to breast cancer. J Breast Imaging 1:84-91, 2019
13. Weeks DE, Lathrop GM: Polygenic disease: Methods for mapping complex disease traits. Trends Genet 11:513-519, 1995
14. Lilyquist J, Ruddy KJ, Vachon CM, et al: Common genetic variation and breast cancer risk—Past, present, and future. Cancer Epidemiol Biomarkers Prev 27:380-394, 2018
15. Palmer JR, Polley EC, Hu C, et al: Contribution of germline predisposition gene mutations to breast cancer risk in African American women. J Natl Cancer Inst 112:1213-1221, 2020
16. Huo D, Feng Y, Haddad S, et al: Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Hum Mol Genet 25:ddw305, 2016
17. Arksey H, O'Malley L: Scoping studies: Towards a methodological framework. Int J Soc Res Methodol Theor Pract 8:19-32, 2005
18. Munn Z, Peters MDJ, Stern C, et al: Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 18:1-7, 2018
19. Torrorey-Sawe R, van der Merwe N, Mining SK, et al: Pioneering informed consent for return of research results to breast cancer patients facing barriers to implementation of genomic medicine: The Kenyan BRCA1/2 testing experience using whole exome sequencing. Front Genet 11:1-12, 2020
20. Ndiaye R, Diop JPD, Bourdon-Huguenin V, et al: Evidence for an ancient *BRCA1* pathogenic variant in inherited breast cancer patients from Senegal. NPJ Genomic Med 5:1-6, 2020
21. Adedokun B, Zheng Y, Ndom P, et al: Prevalence of inherited mutations in breast cancer predisposition genes among women in Uganda and Cameroon. Cancer Epidemiol Biomarkers Prev 29:359-367, 2020
22. Sagna T, Bonora E, Ouedraogo MNL, et al: Identification of BRCA1/2 p.Ser1613gly, p.Pro871leu, p.Lys1183arg, p.Glu1038gly, p.Ser1140gly, p.Ala2466val, p.His2440arg variants in women under 45 years old with breast nodules suspected of having breast cancer in Burkina Faso. Biomol Concepts 10:120-127, 2019
23. Zheng Y, Walsh T, Gulsuner S, et al: Inherited breast cancer in Nigerian women. J Clin Oncol 36:2820-2825, 2018
24. Francies FZ, Wainstein T, De Leeneer K, et al: BRCA1, BRCA2 and PALB2 mutations and CHEK2 c.1100delC in different South African ethnic groups diagnosed with premenopausal and/or triple negative breast cancer. BMC Cancer 15:912, 2015
25. Mvila GL, Postema S, Marchal G, et al: From the set-up of a screening program of breast cancer patients to the identification of the first BRCA mutation in the DR Congo. BMC Public Health 14:1-10, 2014
26. Diez O, Pelegrí A, Gadea N, et al: Novel BRCA1 deleterious mutation (c.1949_1950delTA) in a woman of Senegalese descent with triple-negative early-onset breast cancer. Oncol Lett 2:1287-1289, 2011
27. Zoure AA, Siaoui M, Bambara HA, et al: BRCA1 c.68_69delAG (exon2), c.181T>G (exon5), c.798_799delTT and 943ins10 (exon11) mutations in Burkina Faso. J Public Health Africa 9:11-15, 2018
28. Zhang J, Fackenthal JD, Zheng Y, et al: Recurrent BRCA1 and BRCA2 mutations in breast cancer patients of African ancestry. Breast Cancer Res Treat 134:889-894, 2012
29. Habyarimana T, Attaleb M, Mugenzi P, et al: CHEK2 germ line mutations are lacking among familial and sporadic breast cancer patients in Rwanda. Asian Pac J Cancer Prev 19:375-379, 2018
30. Fackenthal JD, Zhang J, Zhang B, et al: High prevalence of BRCA1 and BRCA2 mutations in unselected Nigerian breast cancer patients. Int J Cancer 131:1114-1123, 2012

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

31. van der Merwe NCN, Hamel N, Schneider S-RR, et al: A founder BRCA2 mutation in non-*Afrikaner* breast cancer patients of the Western Cape of South *Africa*. Clin Genet 81:179-184, 2012
32. Zhang B, Fackenthal JD, Niu Q, et al: Evidence for an ancient BRCA1 mutation in breast cancer patients of Yoruban ancestry. Fam Cancer 8:15-22, 2009
33. Gao Q, Adebamowo CA, Fackenthal J, et al: Protein truncating BRCA1 and BRCA2 mutations in *African* women with pre-menopausal breast cancer. Hum Genet 107:192-194, 2000
34. Diop JPD, Diallo RN, Bourdon-Huguenin V, et al: Novel BRCA2 pathogenic variant c.5219 T>G ; p.(Leu1740Ter) in a consanguineous Senegalese family with hereditary breast cancer. BMC Med Genet 20:1-7, 2019
35. Okobia MN, Bunker CH, Garte SJ, et al: Leptin receptor Gln223Arg polymorphism and breast cancer risk in Nigerian women: A case control study. BMC Cancer 8:1-8, 2008
36. Okobia MN, Bunker CH, Garte SJ, et al: Cytochrome *P450 1B1* Val432 Leu polymorphism and breast cancer risk in Nigerian women: A case control study. Infect Agent Cancer 4:1-8, 2009
37. Kiendrebeogo IT, Zoure AA, Sorgho PA, et al: Glutathione S-transferase M1 (GSTM1) and T1 (GSTT1) variants and breast cancer risk in Burkina Faso. Biomol Concepts 10:175-183, 2019
38. Ahmed JH, Makonnen E, Fotoohi A, et al: Vitamin D status and association of VDR genetic polymorphism to risk of breast cancer in Ethiopia. Nutrients 11:1-14, 2019
39. Habyarimana T, Bakri Y, Mugenzi P, et al: Association between glutathione peroxidase 1 codon 198 variant and the occurrence of breast cancer in Rwanda. Mol Genet Genomic Med 6:268-275, 2018
40. Habyarimana T, Attaleb M, Mugenzi P, et al: Association of p53 codon 72 polymorphism with breast cancer in a Rwandese population. Pathobiology 85:186-191, 2018
41. Shirts BH, Casadei S, Jacobson AL, et al: Improving performance of multigene panels for genomic analysis of cancer predisposition. Genet Med 18:974-981, 2016
42. Kuchenbaecker KB, Hopper JL, Barnes DR, et al: Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA 317:2402-2416, 2017
43. Shahid T, Soroka J, Kong EH, et al: Structure and mechanism of action of the BRCA2 breast cancer tumor suppressor. Nat Struct Mol Biol 21:962-968, 2014
44. Deng C, Brodie S: Roles of BRCA1 and its interacting proteins. Bioessays 22:728-737, 2000
45. Tarsounas M, Sung P: The antitumorigenic roles of BRCA1–BARD1 in DNA repair and replication. Nat Rev Mol Cell Biol 21:284-299, 2020
46. Sirugo G, Williams SM, Tishkoff SA: The missing diversity in human genetic studies. Cell 177:26-31, 2019
47. Ahmed M, Rahman N: ATM and breast cancer susceptibility. Oncogene 25:5906-5911, 2006
48. Feng Y, Rhie SK, Huo D, et al: Characterizing genetic susceptibility to breast cancer in women of *African* ancestry. Cancer Epidemiol Biomarkers Prev 26:1016-1026, 2017
49. Chang BL, Spangler E, Gallagher S, et al: Validation of genome-wide prostate cancer associations in men of *African* descent. Cancer Epidemiol Biomarkers Prev 20:23-32, 2011
50. Shriner D, Adeyemo A, Gerry NP, et al: Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. PLoS One 4:e8398, 2009
51. Chen WC, Bye H, Matejcic M, et al: Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South *African* Black population. Carcinogenesis 40:513-520, 2019
52. Choudhury A, Aron S, Botigué LR, et al: High-depth African genomes inform human migration and health. Nature 586:741-748, 2020
53. Yadav S, Couch FJ: Germline genetic testing for breast cancer risk: The past, present, and future. Am Soc Clin Oncol Ed Book 39:61-74, 2019
54. Duncan LE, Ostacher M, Ballon J: How Genome-Wide Association Studies (GWAS) made traditional candidate gene studies obsolete. Neuropsychopharmacology 44:1518-1523, 2019
55. Pacella-Norman R, Urban MI, Sitas F, et al: Risk factors for oesophageal, lung, oral and laryngeal cancers in black South *Africans*. Br J Cancer 86:1751-1756, 2002
56. ERICA-SA. http://www.mrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa

 © 2021 by American Society of Clinical Oncology

Downloaded from ascopubs.org by 197.90.153.125 on March 27, 2023 from 197.090.153.125
Copyright © 2023 American Society of Clinical Oncology. See https://ascopubs.org/go/authors/open-access for reuse terms.

APPENDIX

TABLE A1. Keywords and Combinations of Keywords Used to Search the Literature Using PubMed

Search No.	Keywords	Fields Searched
#1	breast cancer and variant	Field: Title/Abstract
#2	breast cancer and variants	Field: Title/Abstract
#3	breast cancer and mutation	Field: Title/Abstract
#4	breast cancer and mutations	Field: Title/Abstract
#5	breast cancer and germline	Field: Title/Abstract
#6	breast cancer and hereditary	Field: Title/Abstract
#7	breast cancer and inherited	Field: Title/Abstract
#8	breast cancer and genetic	Field: Title/Abstract
#9	breast cancer and genetics	Field: Title/Abstract
#10	breast cancer and genomic	Field: Title/Abstract
#11	breast cancer and genomics	Field: Title/Abstract
#12	breast cancer and genome-wide	Field: Title/Abstract
#13	breast cancer and GWAS	Field: Title/Abstract
#14	breast cancer and associated	Field: Title/Abstract
#15	#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14	
#16	(africa) OR (african) OR (angola) OR (benin) OR (botswana) OR (burkina faso) OR (burundi) OR (cameroon) OR (central african republic) OR (chad) OR (comoros) OR (congo) OR (democratic republic of congo) OR (DRC) OR (djibouti) OR (equatorial guinea) OR (eritrea) OR (ethiopia) OR (gabon) OR (gambia) OR (ghana) OR (guinea) OR (bissau) OR (ivory coast) OR (kenya) OR (lesotho) OR (liberia) OR (malawi) OR (mali) OR (mauritania) OR (mozambique) OR (namibia) OR (niger) OR (nigeria) OR (rwanda) OR (senegal) OR (sierra leone) OR (somalia) OR (south africa) OR (sudan) OR (swaziland) OR (tanzania) OR (togo) OR (uganda) OR (zaire) OR (zambia) OR (zimbabwe) OR (central africa) OR (west africa) OR (east africa) OR (southern africa) OR (sub saharan africa), Field: Title/Abstract	
#17	#15 AND #16	

The entire search looked like this: ((((((

13. Appendix G: Letter of permission to include adaptation of GWAS workflow

9 Jubilee Road
Parktown
Johannesburg
South Africa
2193

30/07/2023

Dear Dr JT Brandenburg

RE: Request to reproduce material in a thesis

I am a PhD student at the University of the Witwatersrand and I am correcting my thesis. I would like your permission to include in my thesis an adaptation of the diagram of the workflow you have created on your GitHub page for the H3Africa GWAS pipeline.

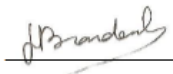
I will include this letter in the appendices of my thesis.

As my co-supervisor, you have given me verbal permission. However, for the documentation of this permission, please sign below.

Kind regards,

Mahtaab Hayat

I, Dr JT Brandenburg, grant Mahtaab Hayat permission to include an adapted diagram of the H3A GWAS pipeline workflow in her thesis.



Dr JT Brandenburg

31/07/2023

Date