

https://dspace.org/

The genetic aetiology of oesophageal cancer in the South African black population

Chen, Wenlong Carl

2022
https://hdl.handle.net/10539/35655

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

THE GENETIC AETIOLOGY OF OESOPHAGEAL CANCER IN THE SOUTH AFRICAN BLACK POPULATION

Wenlong Carl Chen



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

Johannesburg, 2022

DECLARATION

I, Wenlong Carl Chen, declare that this thesis is of my own, unaided work. Where others have contributed and assisted, they have been duly acknowledged. 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 any other University.



Signed on the 17th day of August 2022

DEDICATION

I dedicate this work to my family, Yan, Amy and Michelle, who have supported me throughout this journey. Thank you.

"The meaning of life is to give life meaning."
— Viktor E. Frankl

This work is also dedicated to Dr Elvira Singh, a pioneer, a leader, and a true cancer advocate. Gone too soon.

Hamba Kahle, Elvira. (08 July 1976 – 26 February 2022)

PRESENTATIONS ARISING FROM THIS RESEARCH PROJECT

Conference Proceedings:

Chen WC, Bye H, Matejcic M, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mathew CG. The Genetic Aetiology of Oesophageal Cancer in South African Black Populations. **Poster presentation** at the AACR International Conference: New Frontiers in Cancer Research, 18-22 January **2017**, Cape Town, South Africa.

Chen WC, Bye H, Matejcic M, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mathew CG. The Genetic Aetiology of Oesophageal Cancer in South African Black Populations. **Poster presentation** at the Cancer Research Symposium, University of the Witwatersrand, 9-12 February **2017**, Johannesburg, South Africa.

Chen WC, Bye H, Matejcic M, Amar A, Khew YW, Beynon V, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mather CG. Identifying genetic markers of susceptibility to African Oesophageal Cancer. **Oral presentation** at the CANSA Research in Action Conference, 3-5 July **2018**, Pretoria, South Africa.

Chen WC, Bye H, Matejcic M, Amar A, Khew YW, Beynon V, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mather CG. The genetic aetiology of oesophageal cancer in South African Black populations. **Poster presentation** at the Molecular Biosciences Research Thrust Postgraduate Research Day, University of the Witwatersrand, 29 November **2018**, Johannesburg, South Africa.

Chen WC, Bye H, Matejcic M, Amar A, Khew YW, Beynon V, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mather CG. The genetic aetiology of oesophageal cancer in South African Black populations. **Poster presentation** at the Faculty of Health Sciences biennial Research Day, University of the Witwatersrand, 6 September **2018**, Johannesburg, South Africa.

Chen WC, Bye H, Matejcic M, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI, Mathew CG. The Genetic Aetiology of Oesophageal Cancer in South African

Black Populations. **Poster presentation** at the American Association for Cancer Research, Annual Meeting 2018, 14-18 April **2018**, Chicago, Illinois.

Chen WC, Babb de Villiers C, Kerr R, Mathew CG. The genetic aetiology of Oesophageal squamous cell carcinoma in the South African black population. **Oral presentation** at the Faculty of Health Sciences Biennial Research Day **2020**, University of Witwatersrand, 15th October 2020, Johannesburg, South Africa.

Chen WC, Brandenburg JT, Ramsay M, Parker MI, Lewis CM, Mathew CG. Genome-wide Association Study of African Esophageal Squamous Cell Carcinoma. **Oral presentation** at the AORTIC Cancer Genomic Conference, 18 & 19th February **2021**, Virtual.

Chen WC, Mathew CG for ERICA-SA. Genome-wide association study of African oesophageal squamous cell carcinoma. **Oral presentation** at the general session and invited speaker presentation at the Oesophageal cancer session, AORTIC bi-annual meeting, 5-10th November **2021**, Virtual.

Chen WC, Mathew CG. Hunting causal genes in common African cancers. **Oral presentation** at the Sydney Brenner Institute for Molecular Biosciences Research Symposium, University of the Witwatersrand, 11th October **2021**, Johannesburg, South Africa.

PEER-REVIEWED PUBLICATIONS ARISING FROM THIS RESEARCH PROJECT

The research findings arising from this study were drafted and presented in three separate manuscripts. Two of these have been published in international peer-reviewed scientific journals. The third manuscript has been submitted for publication. The signed co-author agreement/supervisor declaration for the use of each manuscript can be found in Appendix A.

Chen WC, Kerr R, May A, Ndlovu B, Sobalisa A, Duze ST, Joseph L, Mathew CG & Babb de Villiers (**2018**) The integrity and yield of genomic DNA isolated from whole blood following long-term storage at -30°C. *Biopreservation and Biobanking*, Apr 2018, 106-113. https://doi.org/10.1089/bio.2017.0050

Chen WC, Bye H, Matejcic M, Amar A, Govender D, Khew YW, Beynon V, Kerr R, Singh E, Prescott NJ, Lewis CM, Babb de Villiers C, Parker MI & Mathew CG (**2019**) Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. *Carcinogenesis*, 40(4) Apr 2019, 513-520. https://doi.org/10.1093/carcin/bgz026

Chen WC, Brandenburg JT, Choudhury A, Hayat M, Sengupta D, Swiel Y, Babb de Villiers C, Ferndale F, Aldous C, Soo C, Lee S, Curtis C, Newton R, Waterboer T, Sitas F, Bradshaw D, Abnet CC, Ramsay M, Parker MI, Singh E, Lewis CM & Mathew CG (**2022**) Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations. Submitted to *Nature Communications*.

RELATED PEER-REVIEWED PUBLICATIONS

Chen WC, Singh E, Muchengeti M, Bradshaw D, Mathew CG, Babb de Villiers C, Lewis CM, Waterboer T, Newton R & Sitas F (**2020**) Johannesburg Cancer Study (JCS): contribution to knowledge and opportunities arising from 20 years of data collection in an African setting. *Cancer Epidemiology*, 65(101701) Apr 2020. https://doi.org/10.1016/j.canep.2020.101701

Ferndale F, Moodley M, **Chen WC**, Wadee R, Wright CA, Parker MI, Willem P & Mathew CG (**2021**) Processing and analysis of tissue samples from esophageal cancer patients in an African setting. *Biopreservation and Biobanking*, epub ahead of print, Aug 2021. https://doi.org/10.1089/bio.2021.0030

Ferndale L, Ayeni OA, **Chen WC** & Aldous C (**2022**) Factors Influencing survival time in patients treated for oesophageal cancer with palliative intent in South Africa. *World Journal of Surgical Oncology*, under review & preprint. https://doi.org/10.21203/rs.3.rs-1074694/v1

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

Simba H, Tromp G, Sewram V, Mathew CG, **Chen WC** & Kuivaniemi H (**2022**) Esophageal Cancer Genomics in Africa: Recommendations for Future Research. *Frontiers in Genetics, section Human and Medical Genomics*, published online: 25 March 2022. https://doi.org/10.3389/fgene.2022.864575

GRANTS AND AWARDS

2016: Recipient of a National Research Foundation (NRF) DAAD-NRF In-Country Scholarship (R100 000).

2017: Recipient of the Cancer Research Symposium, University of the Witwatersrand, best poster award (R5 000).

2017: Recipient of a Boehringer Ingelheim Fonds travel grant (€2 350).

2018: Recipient of an NRF Thuthuka PhD track research grant for three years (R250 000 per annum).

2019: Recipient of a University of the Witwatersrand Staff Bursary for PhD fees.

2021: Recipient of a National Institute of Health (NIH), United States, R01 administrative supplement grant 3R01CA238433-03S1 ($50 000).

ABSTRACT

Oesophageal squamous cell carcinoma (OSCC) is a common disease in eastern and southern Africa with a complex aetiology. Little is known about the contribution of genetic factors to OSCC risk in African populations. Recent genome-wide association studies (GWAS) in Asian and European populations identified multiple genetic risk loci for OSCC, but few findings from other populations have been replicated in African OSCC. This thesis aims to identify genetic risk factors for OSCC in the black South African population.

High molecular weight DNA of good quality and quantity is needed to conduct large-scale genetic studies. We had access to a large resource of frozen blood samples from OSCC cases and controls collected by the Johannesburg Cancer Study (JCS). The integrity and yield of genomic DNA isolated from 2,758 whole blood following long-term storage at -30°C were assessed using spectrophotometry, fluorometric assay for double-stranded DNA and agarose gel electrophoresis. The mean DNA yield per sample was 114 µg from whole blood volumes of 0.5 ml to 4 ml. The mean A260/280 ratio and median A260/280 ratios were both 1.8. This analysis showed that high-quality DNA can be extracted from whole blood samples that are stored at -30°C for up to 19 years.

The contributions of genetic factors to the risk of African OSCC were investigated for 7 risk loci previously identified in non-African populations. Association studies were performed in a total of 1,471 OSCC cases and 1,791 controls from Black South Africans in two study sample groups from the Western Cape and Gauteng province. Thereafter a meta-analysis was done for 11 variants that had been genotyped in both studies. A single nucleotide polymorphism (SNP) in the *CHEK2* gene, rs1033667, was significantly associated with OSCC (P = 0.002; OR = 1.176; 95% CI: 1.06 - 1.30). However, SNPs in the *CASP8/ALS2CR12*, *TMEM173*, *PLCE1*, *ALDH2*, *ATP1B2/TP53* and *RUNX1* loci were not associated with the disease. The lack of association of these six loci with OSCC in South African populations may reflect different genetic risk factors in non-African and African populations or differences in the genetic architecture of African genomes. The association of the SNP in the *CHEK2* gene provided support for the contribution of a common genetic variant at this locus to the risk of OSCC.

A genome-wide association study (GWAS) was then conducted in 1,686 African OSCC patients and 3,217 controls to broaden the search for genetic risk factors. Samples were genotyped using the H3A custom microarray containing ~2.3 million genetic markers enriched for African content. One novel, genome-wide significant risk locus was identified on chromosome 9, upstream of the *FAM120A* gene (rs12379660, $P = 4.58x10^{-8}$, odds ratio = 1.28, 95% confidence interval = 1.22-1.34) and one African-specific risk locus was identified on chromosome 2 (rs142741123, $P = 5.49x10^{-8}$) within the *MYO1B* gene. A meta-analysis was performed for the African OSCC GWAS with a Chinese OSCC GWAS (2,013 OSCC patients, 2,701 controls) and identified three genome-wide significant loci on chromosome 9 at *FAM120A* (rs12379660, $P_{meta} = 9.36x10^{-10}$), chromosome 10 at *PLCE1* (rs7099485, $P_{meta} = 1.48x10^{-8}$) and chromosome 22 at *CHEK2* (rs1033667, $P_{meta} = 1.47x10^{-9}$).

This study, which is the first GWAS done in an African population, provided evidence of a substantial genetic contribution to OSCC risk in Africa and identified both shared and distinct risk loci for OSCC in populations of diverse ancestry. This work is the first indication of the molecular basis of OSCC in Africa and shows that larger studies in African populations are justified to identify more of the genetic contribution to the risk of this disease.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and heartfelt appreciation for the many people who have travelled with me on this amazing journey:

To my supervisor, Professor Christopher Mathew, thank you for your generous guidance, support and mentorship. Thank you for guiding me through my journey and development into a cancer researcher. This journey would not have been possible without your support.

To my co-supervisor, Dr Chantal Babb de Villiers, thank you for this opportunity for me to embark on this amazing PhD journey. Thank you for believing in me and thank you for bringing me into the NCR family.

To my co-supervisor, Dr Robyn Kerr, thank you for being by my side and supporting me through my post-graduate journey, from MSc to PhD.

To Dr Elvira Singh and Dr Maureen Joffe, thank you for your mentorship and for moulding me into a better person and a better scientist. Thank you for your generosity that enabled me to pursue this PhD.

To Dr Jean-Tristan Brandenburg, thank you for all your help on the technical aspects of this project.

To my colleagues at the SBIMB and the NCR, thank you for welcoming me into the family and walking this journey with me.

To the study participants, thank you for your generosity. Without your participation, this study would not have been possible.

Lastly, to the funders who have made this study possible, thank you.

TABLE OF CONTENTS

DECLARATION ii
DEDICATION iii
PRESENTATIONS ARISING FROM THIS RESEARCH PROJECT iv
PEER-REVIEWED PUBLICATIONS ARISING FROM THIS RESEARCH PROJECT vi
RELATED PEER-REVIEWED PUBLICATIONS vii
GRANTS AND AWARDS viii
ABSTRACT ix
ACKNOWLEDGEMENTS xi
TABLE OF CONTENTS xii
LIST OF FIGURES xv
PREFACE xxi
1. CHAPTER 1: Introduction 1
1.1. Oesophageal Cancer in Africa 1
1.2. The genetic contribution to African oesophageal squamous cell carcinoma: Candidate gene studies 4
1.3. Genome-wide association studies in human disease 7
1.3.1. The Human Hereditary and Health In Africa Consortium 8
1.4. Genome-wide association studies in oesophageal squamous cell carcinoma 10
1.5. Study Aim and Objectives 11
1.6. References: 12
2. CHAPTER 2: Summary of Methods 19
2.1. Overview of study design 19
2.2. Objective 1: Assessing the quality of DNA samples 19
2.3. Objective 2: Replication of known OSCC genetic risk factors 20
2.4. Objectives 3 and 4: Novel OSCC genetic risk factor discovery and transethnic meta-analysis 20
2.5. Ethical clearance and certification: 22
2.6. References: 23
3. CHAPTER 3: The Integrity and Yield of Genomic DNA Isolated from Whole Blood following long-term storage at -30°C. 25
3.1. Abstract 25

3.2. Introduction 26
3.3. Materials and Methods 27
3.4. Results 30
3.5. Discussion 36
3.6. Acknowledgements 39
3.7. Author contributions 40
3.8. Author Disclosure Statement 40
3.9. References 40
4. CHAPTER 4: Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. 43
4.1. Abstract 43
4.2. Introduction 44
4.3. Materials and Methods 45
4.4. Results 49
4.5. Discussion 54
4.6. Funding 57
4.7. Acknowledgements 58
4.8. Supplementary 59
4.9. References 61
5. CHAPTER 5: Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations 65
5.1. Abstract 65
5.2. Introduction 66
5.3. Results 68
5.4. Discussion 77
5.5. Methods 80
5.6. Data Availability 84
5.7. Acknowledgements 84
5.8. Author contributions 85
5.8. Competing interests 86
5.9. Supplementary 86
5.10. References: 96
6. CHAPTER 6: Discussion and Conclusions 103

6.1. Discussion 103
6.2. Conclusion 111
CONSOLIDATED REFERENCES 112
APPENDIX 125

LIST OF FIGURES

Figure 1.1. Oesophageal cancer. Oesophageal cancer commonly occurs in the cells lining the oesophagus. (www.mayoclinic.org)..... 2
Figure 1.2. Oesophageal gastroscopy image of advanced dysphagia. 2
Figure 1.3. Estimated age-standardized incidence rates (World) in 2020, for oesophageal cancer, both sexes, all ages. GLOBOCAN 2021. Sung et al., 2021..... 3
Figure 1.4. NHGRI-EBI GWAS Catalog published Genome-Wide Associations as of July 2019. $P \leq 5x10^{-8}$ for 17 trait categories. 8
Figure 1.5. The AWI-Gen study's participating Research Centres and enrolment sites..... 9
Figure 3.1. Gel image showing the integrity of genomic DNA from 4 years to 19 years of blood sample storage..... 32
Figure 3.2. %dsDNA vs storage duration. 33
Figure 3.3. Range and distribution of DNA yields per 1ml of blood, extracted from frozen whole blood stored over 4 to 19 years. 35
Figure 3.4. A260/280 ratios distribution of all samples across the duration of blood sample storage. 35
Figure 3.5. A260/230 ratios distribution of all samples across the duration of blood sample storage. 36
Figure 4.1. Schematic of the overall study design. 47
Figure 4.2. Comparison of the allele frequencies of the 19 SNPs of interest by population. 53
Figure 5.1. Admixture and Principal component plots of OSCC cases and controls..... 69
Figure 5.2. Quantile-Quantile and Manhattan plot of the African OSCC GWAS..... 70
Figure 5.3. Locuszoom plots of the African OSCC GWAS hits..... 71
Figure 5.4. Meta-analysis Manhattan plot for African OSCC and Chinese OSCC. 73
Figure 5.5. Meta-analysis Locus zoom plots. 75

LIST OF TABLES

Table 1.1. OSCC Genetic susceptibility studies conducted in Africa. (adapted from Simba et al., 2019) 6
Table 1.2. Genetic variants showing significant summary OR for OSCC meta-analysis. (adapted from Li et al., 2020) 10
Table 3.1. Descriptive statistics for the samples used for genomic DNA isolation per year of storage. 31
Table 4.1. Characteristics of OSCC cases and controls in the SAB populations. 50
Table 4.2. Case-control association study analysis for SNPs genotyped in 7 known risk loci for OSCC in the SAB populations. 51
Table 4.3. Meta-analysis of 11 variants genotyped in the SAB populations. 52
Table 5.1. Top SNPs associated with African oesophageal squamous cell carcinoma from Chromosome 9 and Chromosome 2. 74
Table 5.2. Top signals from African-Chinese meta-analysis for oesophageal squamous cell carcinoma. 76

LIST OF SUPPLEMENTARY FIGURES

Supplementary figure 5.1. The eigenvalues for Principal component analysis of OSCC GWAS..... 86
Supplementary figure 5.2. Cluster plot for rs12379660.....87
Supplementary figure 5.3. Study schematic and workflow outline of the African OSCC GWAS study. 88
Supplementary figure 5.4. Cross-validation errors for Admixture analysis, for K=1 to K=9. Optimal K=6, with the lowest cross-validation error of 0.51286. 89

LIST OF SUPPLEMENTARY TABLES

Supplementary Table 4.1. Case-control association study analysis for SNPs genotyped in 7 known risk loci for OSCC in the SAB populations for dominant inheritance model and recessive inheritance model. 59
Supplementary Table 4.2. Meta-analysis of 11 variants genotyped in the SAB populations under the dominant inheritance model. 60
Supplementary Table 4.3. Meta-analysis of 11 variants genotyped in the SAB populations under recessive inheritance model. 60
Supplementary Table 4.4. Comparison of minor allele frequencies of the 19 SNPs across the UCT, JCS study cohort and the YRI, CHB and CEU populations from the 1000 Genome Project Phase 3 data. 61
Supplementary Table 5.1. OSCC GWAS study participants by case-control status. 90
Supplementary Table 5.2. Imputation score (Impute2 score) of all imputed SNPs in the African OSCC GWAS dataset. 90
Supplementary Table 5.3. Allele frequencies of the top SNPs in Chromosome 2 from the OSCC GWAS and the 1KG. 91
Supplementary Table 5.4. Linkage disequilibrium independent SNPs from 31 OSCC associated loci with P values < $5x10^{-6}$. 92
Supplementary Table 5.5. Replication of known OSCC risk SNPs and loci 93
Supplementary Table 5.6. eQTL mapping for the top GWAS hit on Chromosome 9. 95
Supplementary Table 5.7. eQTL mapping for the Meta-analysis top hits on Chromosomes 10 and 22. 95
Supplementary Table 5.8. Data quality control parameters. 96

ABBREVIATIONS

1KG	– 1000 Genomes Consortium
ADC	– Adenocarcinoma
AFR	– African
AfrECC	– African Esophageal Cancer Consortium
AIMS	– Ancestry informative markers
ASR	– Age-standardized incidence rate
AWI-Gen	– Africa Wits-INDEPTH partnership for Genomics studies
AWS	– Amazon web service
CANSA	– Cancer Association of South Africa
CERG	– Cancer epidemiology research group
CEU	– Northern European (population)
CHB	– Han Chinese (population)
CHN	– Chinese
CI	– Confidence interval
CV	– Cross-validation
DAAD	– German Academic Exchange Service
dH_2o	– Distilled water
DNA	– Deoxyribonucleic acid
dsDNA	– Double-stranded DNA
DST	– Department of Science and Technology
EAF	– Effect-allele frequency
EDTA	– Ethylenediaminetetraacetic acid
EGA	– European Genome-Phenome archive
eQTL	– Expression quantitative trait locus
ERICA-SA	– Evolving Risk Factors for Cancers in African Populations
FFPE	– Formalin-Fixed Paraffin-Embedded
FPGA	– Field-Programmable Gate Arrays
FUMA	– Functional mapping and annotation
gDNA	– Genomic DNA
GH	– Greys Hospital
GIPD	– Grants, Innovation and Product Development (unit)
GLM	– Generalized linear modelling
GRM	– Genetic relationship/kinship matrix
GTEx	– Genotype-Tissue Expression (project)
GWAS	– Genome-wide association study/studies
GxE	– Gene x environmental
h2g	– Genetic heritability estimate
H3A	– Human Heredity & Health in Africa (consortium)
HIV	– Human immunodeficiency virus
HPV	– Human papilloma virus

HREC – Human Research Ethics Committee
HWE – Hardy–Weinberg equilibrium
INDEPTH – International Network for the Demographic Evaluation of Populations and their Health in low and middle-income countries
JCS – Johannesburg Cancer Study
LD – Linkage disequilibrium
LMM – Linear-mixed model
MAF – Minor allele frequency
mRNA – Messenger RNA
NCR – National cancer registry of South Africa
NHLS – National Health Laboratory Service
NIH – National Institute of Health
NRF – National Research Foundation
OADC – Oesophageal adenocarcinoma
OC – Oesophageal Cancer
OR – Odds ratio
OSCC – Oesophageal squamous cell carcinoma
PC – Principal component
PCR – Polymerase Chain Reaction
PhD – Doctor of Philosophy
QC – Quality control
QPCR – Quantitative Polymerase Chain Reaction
RFLP – Restriction fragment length polymorphism
RNA – Ribonucleic acid
SAB – South African Black (population)
SAMRC – South African Medical Research Council
SAP – Shrimp alkaline phosphatase
SBIMB – Sydney Brenner Institute for Molecular Biosciences Biobank
SCC – Squamous cell carcinoma
SNP – Single nucleotide polymorphisms
ssDNA – Single-stranded DNA
TE – Tris EDTA
UCT – University of Cape Town
UKZN – University of the KwaZulu-Natal
WES – whole-exome sequencing
Wits – University of the Witswatersrand
WTCCC – Wellcome Trust Case Control Consortium
YRI – Yoruban (population)

PREFACE

This body of work is the culmination of the significant contributions made by academic and intellectual greats that came before me. Through the decades of collective efforts, the vast wealth of knowledge and resources collected through the Johannesburg Cancer Study provided a robust foundation for us to further the scientific research and better understand the devastating disease of oesophageal squamous cell cancer. I was fortunate to have this opportunity to pursue a PhD degree and add to the knowledge pool to enable future work.

This PhD thesis is presented in the "divided-block" format which includes three manuscripts. Two manuscripts have been published in respective internationally peer-reviewed scientific journals. The third manuscript has been submitted for publication.

This thesis is divided into six chapters:

Chapter 1: Introduction

This chapter describes the literature on the genetics of oesophageal squamous cell cancer and the study's aims and objectives.

Chapter 2: Summary of Methods

This chapter summarises the methodological approaches used in each manuscript. As the methods are described in total in each manuscript, this chapter serves as an overview.

Chapter 3: The integrity and yield of genomic DNA isolated from whole blood following long-term storage at -30°C

This is the first published manuscript and the first results chapter. In this chapter, the quality and quantity of genomic DNA isolated from archival blood samples were assessed. Large-scale genetic and genomic analysis requires high-quality genomic DNA, thus having good quality DNA is essential. We demonstrated the high quality of the DNA and verified that it would be suitable for the planned genomic analysis, particularly for GWAS.

Chapter 4: Association of genetic variants in *CHEK2* with oesophageal squamous cell carcinoma in the South African Black population

This is the second published manuscript. This chapter describes our work testing previous known genetic risk loci for oesophageal squamous cell cancer identified in the Chinese population for replication and association with OSCC risk in the South African Black population.

Chapter 5: Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations

This is the third manuscript, which has just been submitted for publication. This chapter describes our work identifying African-specific genetic risk loci and trans-ethnic genetic risk loci for oesophageal squamous cell cancer through a GWAS.

Chapter 6: Discussion and Conclusions

This chapter contains the integrated discussion and consolidated findings for the three manuscripts, as well as the limitations and future work. This chapter ends with an overall conclusion of what has been achieved.

1. CHAPTER 1: Introduction

1.1. Oesophageal Cancer in Africa

Oesophageal Cancer (OC) is the eighth most common cancer worldwide with over 600,000 new cases diagnosed each year and is responsible for more than 540,000 deaths annually (Sung et al., 2021). OC develops from the inner lining of the oesophagus (Figure 1.1) and cancer can develop anywhere within the oesophagus. OC can be classified as either adenocarcinoma (ADC) or squamous cell carcinoma (SCC). Historically, oesophageal ADC (OADC) and oesophageal SCC (OSCC) were considered the same disease, yet they are distinctly different. ADC is an uncontrolled abnormal proliferation of glandular tissue, whereas SCC arises from the rapid proliferation of squamous epithelial cells. Recently, the authors of The Cancer Genome Atlas (TCGA) provided molecular evidence of the intrinsic differences between OADC and OSCC thus confirming that OADC and OSCC are two distinct tumour types with different pathogenesis, epidemiology, tumour biology and prognosis requiring different therapeutic strategies (Siewert & Ott, 2007; Lagisetty, Beer & Chang, 2017; TCGA, 2017). OADC is primarily associated with gastroesophageal reflux disease and Barret's oesophagus and occurs mainly in industrialized countries (Enzinger & Mayer, 2003).

The aetiology of OSCC is complex and less clear, with the developing world accounting for 88% of its global incidence (Arnold et al., 2015). The development of OSCC is largely asymptomatic and the clinical presentation is often late. Patients often present with dysphagia (difficulty in swallowing) (Figure 1.2), unintended and unexplained weight, chest pain, heartburn and/or coughing or hoarseness (Reichenbach et al., 2019). Diagnosis is made by a combination of imaging, gastroscopy and biopsy with histopathology. High-risk areas for OSSC include Linxian in China, northern Iran and sub-Saharan Africa (Figure 1.3). The age-standardized incidence rate (ASR) in Southern and Eastern Africa is 13 cases per 100,000 (Sung et al., 2021), whereas in rural populations of the Eastern Cape the ASR is 23.2 per 100,000 in males and 14.5 per 100,000 in females (Forman et al., 2013). OC is the most common cancer amongst males and the second most common cancer amongst females in the Eastern Cape (Forman et al., 2013). Other regions in South Africa with a high prevalence of OC include Gauteng and Kwazulu-Natal (NCR, 2021).

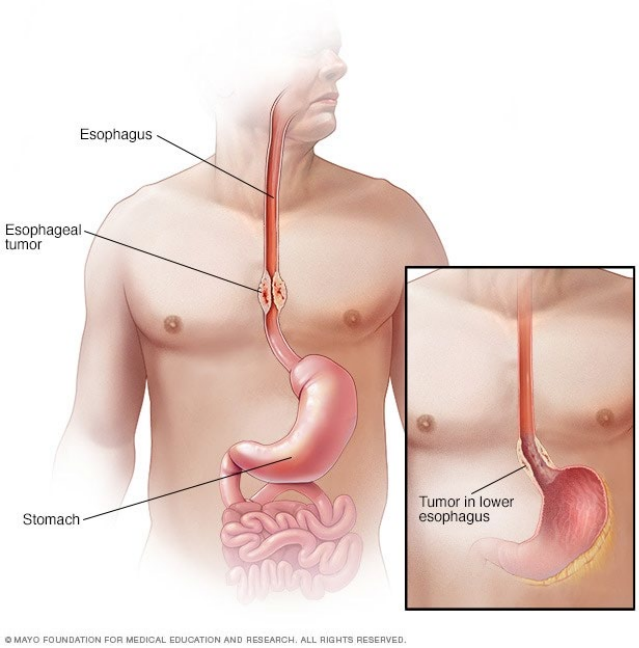


Figure 1.1. Oesophageal cancer. Oesophageal cancer commonly occurs in the cells lining the oesophagus. (www.mayoclinic.org)

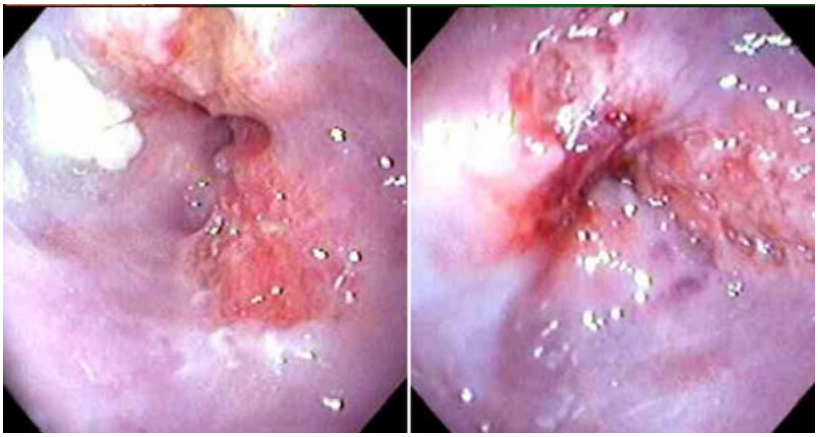


Figure 1.2. Oesophageal gastroscopy image of advanced dysphagia.

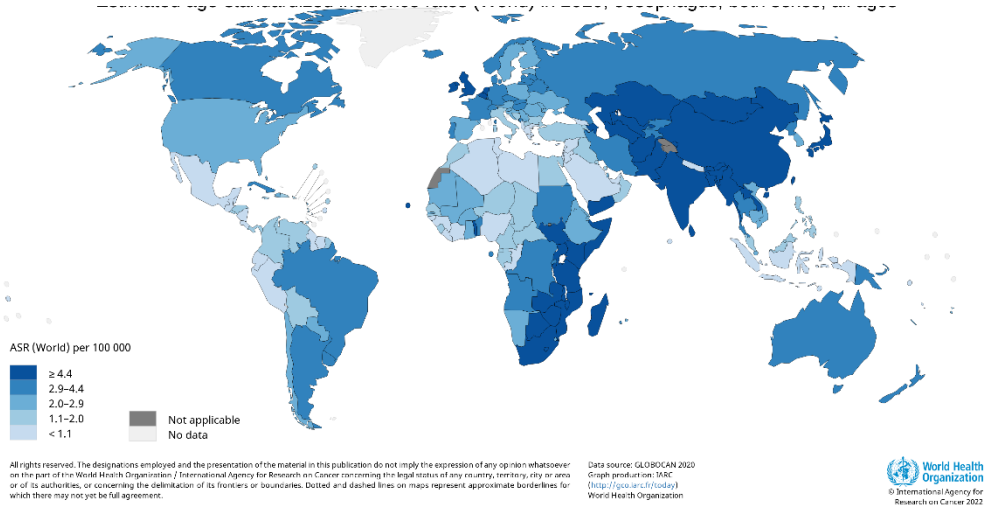


Figure 1.3. Estimated age-standardized incidence rates (World) in 2020, for oesophageal cancer, both sexes, all ages. GLOBOCAN 2021. Sung et al., 2021

The prognosis of OSCC in Africa is very poor. Early-stage OC can be treated with curative intent. This can be achieved by surgery, or neo-adjuvant chemo/radiation therapy followed by surgery. Late-stage treatments are generally palliative including the use of adjuvant chemo/radiation therapies. A common treatment option for OC in LMICs is the use of a stent to open the oesophageal blockage to allow the patient to swallow (Ferndale et al., 2019; Loots, Ferndale & Clarke, 2019). Patients treated with stents in Kenya and Malawi, for example, had a median survival of only 250 (just over 8 months) and 210 days (just under 7 months) respectively after treatment (White et al., 2009; Thumbs et al., 2012), while South African patients have a median survival of 112 days (3.7 months) (Ferndale et al., 2021). The success of treatment is dependent on numerous factors such as stage of diagnosis, location of the tumour, the patient's physical state and response to treatments. In HICs, targeted immunotherapies for OC are available (Sun et al, 2021). The high incidence and poor prognosis for OC constitute a major ongoing public health problem in South and East Africa.

The aetiology of OSCC is complex; implicated risk factors in Africa include low socioeconomic status, poor nutrition, tobacco smoking, alcohol use, smoke inhalation from cooking fires, consumption of fumonisin-contaminated maize, dental fluorosis and the presence of human papillomavirus (HPV) and human immunodeficiency virus (HIV) (Murphy et al., 2017; Menya, Kigen, et al., 2019; Menya, Maina, et al., 2019), but no single major cause has been identified. In South Africa, case-control studies in urban Soweto confirmed the association with smoking and alcohol, particularly smoking pipe tobacco

and consuming home-brewed beer (Segal, Reinach & de Beer, 1988; Isaacson et al., 2015). The Johannesburg Cancer Study (JCS) reported that the combination of smoking and regular alcohol consumption increased the risk further, with an odds ratio (OR) of 4.4 (95% confidence interval (CI): 3.2-6.1) (Pacella-Norman et al., 2002; Chen et al., 2020). Infection with HPV has been implicated in OSCC risk (Schäfer et al., 2013), but a rigorous analysis of tumour tissues from OSCC patients in China, South Africa and Iran for the presence of 51 human alpha-papillomaviruses did not support an aetiological role for these viruses in OSCC (Halec et al, 2016). HIV infection was reported to confer an increase in risk for OSCC in Zambia, with an OR of 2.3 (95% CI: 1.0-5.1), and no risk associated with HPV infection (Kayamba et al., 2015).

1.2. The genetic contribution to African oesophageal squamous cell carcinoma: Candidate gene studies

An important but unresolved question in the aetiology of OSCC is whether genetic factors contribute significantly to disease risk. If so, the nature of the genes and genetic variants involved could offer important clues to the underlying molecular pathways to pathogenesis, as has been demonstrated for other common, complex disorders such as Crohn's disease (reviewed in Mathew, 2008). The identification of genetic risk variants for OSCC might also have potential in the stratification of disease risk in high-risk populations where risk genotypes can be combined with environmental risk factors such as smoking and alcohol to produce composite risk prediction scores.

Simba et al. (2019) reviewed the current state of genetic research in OSCC in Africa, which is summarised in table 1. To date, 23 genetic studies on OSCC have been conducted in resident African populations, with the majority of the studies done in South Africa. Of the 23 studies, 16 investigated germline genetic susceptibility to OSCC. These 16 studies utilized genotyping techniques such as TaqMan assays, RFLP (restriction fragment length polymorphism), Sanger sequencing or iPLEX, to replicate known OSCC genetic risk factors identified in non-African populations (Zaahl et al., 2005; Dandara, Ballo & Parker, 2005; Li, Dandara & Parker, 2005, 2010; Chelule et al., 2006; Dandara et al., 2006; Li et al., 2008; Bye et al., 2011, 2012; Matejcic et al., 2011, 2015; Strickland et al., 2012; Vogelsang et al., 2012; Eltahir et al., 2012). Six studies investigate the role of somatic mutations in OSCC (Victor et al., 1990; Gamieldien et al., 1998; Dietzsch & Parker, 2002; Dietzsch, Laubscher & Parker,

2003; Vos et al., 2003; Naidoo et al., 2005; Patel et al., 2011; Liu et al., 2017), and two studies investigated both germline and somatic variants (Dietzsch, Laubscher & Parker, 2003; Vos et al., 2003).

The genetic susceptibility studies tested single nucleotide polymorphisms (SNPs) in 37 genes for association with African OSCC. Of the genes tested, 25 SNPs in 20 genes showed some evidence of association with OSCC, namely *ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, GSTT2B, MGMT, MLH3, MSH3, NAT2, PTGS2, PLCE1, PMS1, RUNX1, SLC11A1, and TP53* (Simba et al., 2019). It is important to note that all 25 SNP associations were detected in South African populations. This included earlier work by the Mathew research group at King's College London which tested SNPs that had been associated with OSCC in genome-wide association studies (GWAS) from other populations for association with OSCC in the South African Black (SAB) population in cases and controls from Xhosa-speakers in the Cape Province (Bye et al., 2011, 2012; Matejcic et al., 2011). Most SNPs showed no evidence of association with OSCC, although preliminary evidence was found for an association of a missense variant Arg548Leu in the *PLCE1* gene, with OR = 0.74 (95% CI: 0.60-0.93), P = 0.008 (Bye et al., 2012).

These studies had several limitations. They were limited in scope in that they focused on the analysis of a limited number of genetic variants from a small number of candidate genes. The sample sizes were relatively small and thus underpowered to detect the modest increases in risk generally conferred by genetic variants (Zhang et al, 2020). Also, studies that genotyped only a small number of genetic variants do not have the means to adjust for population structure in cases and controls, which is particularly important in the context of very diverse African genomes. Thus, the results of such studies could generate false-positive signals driven by population substructure (Sengupta et al, 2021). In the context of the previous studies performed, an OSCC GWAS can overcome the limitations of targeted candidate gene studies in African OSCC by adopting a genome-wide approach.

Table 1.1. OSCC Genetic susceptibility studies conducted in Africa. (adapted from Simba et al., 2019)

Study	Location	Year	Population	Sample size		Analysis method	PMID
				Cases	Ctrl		
Bye et al., 2011	South Africa	2011	Black	358	477	TaqMan Assay	21926110
			Mixed ancestry	201	427	TaqMan Assay	
Bye et al., 2012	South Africa	2012	Black	407	849	TaqMan Assay and KASP	22865593
			Mixed ancestry	257	860	TaqMan Assay and KASP	
Chelule et al., 2006	South Africa	2006	Black	70	261	PCR-RFLP	17264406
Dandara et al., 2005	South Africa	2005	Black	142	178	PCR-RFLP	15978331
			Mixed ancestry	99	94	PCR-RFLP	
Dandara et al., 2006	South Africa	2006	Black	145	194	PCR-RFLP	16272171
			Mixed ancestry	100	94	PCR-RFLP	
Dietzsch et al., 2003	South Africa	2003	Black and mixed ancestry	582	226	PCR and PAGE	12925954
Eltahir et al., 2012	Sudan	2012	?	18	235	PCR-RFLP	23053979
Li et al., 2005	South Africa	2005	Black and mixed ancestry	189	198	PCR-SSCP and DNA sequencing	15899651
Li et al., 2008	South Africa	2008	Black	142	178	PCR- RLFP	18254707
			Mixed ancestry	101	100	PCR-RFLP	
Li et al., 2010	South Africa	2010	Black	145	194	PCR-RFLP	20540773
			Mixed ancestry	100	94	PCR- RFLP	
Matejcic et al., 2011	South Africa	2011	Black	330	479	TaqMan assay and gel electrophoresis	22216261
			Mixed ancestry	232	428	TaqMan assay and gel electrophoresis	
Matejcic et al., 2015	South Africa	2015	Black	463	480	TaqMan assay	26447020
			Mixed ancestry	269	288	TaqMan Assay	
Strickland et al., 2012	South Africa	2012	Black	96	88	HEX SSCP and DNA sequencing	21901748
Vogelsang et al., 2012	South Africa	2012	Black	345	344	Allele-specific quantitative PCR	22623965
			Mixed ancestry	205	266	Allele-specific quantitative PCR	
Vos et al., 2003	South Africa	2003	Black	74	118	SSCP and DNA sequencing	12550754
Zaahl et al., 2005	South Africa	2005	Mixed ancestry	105	110	SSCP and DNA sequencing	15860357

1.3. Genome-wide association studies in human disease

GWAS is based on the common variant-common disease hypothesis which predicts that common risk variants exist in all human populations and influence the risk of a particular disease. Each variant will have a small effect on the risk of disease, but the cumulative effect of the inheritance of many risk alleles for a disease will make a substantial difference in disease risk. The GWAS itself is a hypothesis-free association study that makes no assumptions as to which gene or variants may contribute to disease risk. Instead, millions of SNPs from across the entire genome are genotyped in large numbers of cases and controls and each SNP is then tested for association with the phenotype. GWAS leverages the phenomenon of linkage disequilibrium (LD), whereby segments of DNA are inherited non-independently. Thus, several variants/SNPs are passed on together as a single block, and genotyping one of the variants, known as the "tag" SNP, will inform the genotypes of the remaining variants within that block without having to genotype every variant. This allows a GWAS microarray containing several hundred thousand to several million SNPs to effectively provide coverage of the entire genome (Sud, Kinnersley & Houlston, 2017). While the first GWAS was published in 2005 (Klein et al., 2005), it was the Wellcome Trust Case Control Consortium (WTCCC) that demonstrated the robust application of GWAS for large-scale case-control studies in 2007 (Burton et al., 2007). By combining large study cohorts, as well as the utilization of shared population controls, the WTCCC identified novel genetic risk loci for seven common diseases. The GWAS Catalog (https://www.ebi.ac.uk/gwas/home), a central repository for the storage and curation of GWAS summary statistics, summarized by its 10th anniversary in 2018, has curated data from more than 5687 studies, for more than 71,673 variants – trait associations from 3,567 publications (Buniello et al., 2019) (Figure 1.4).

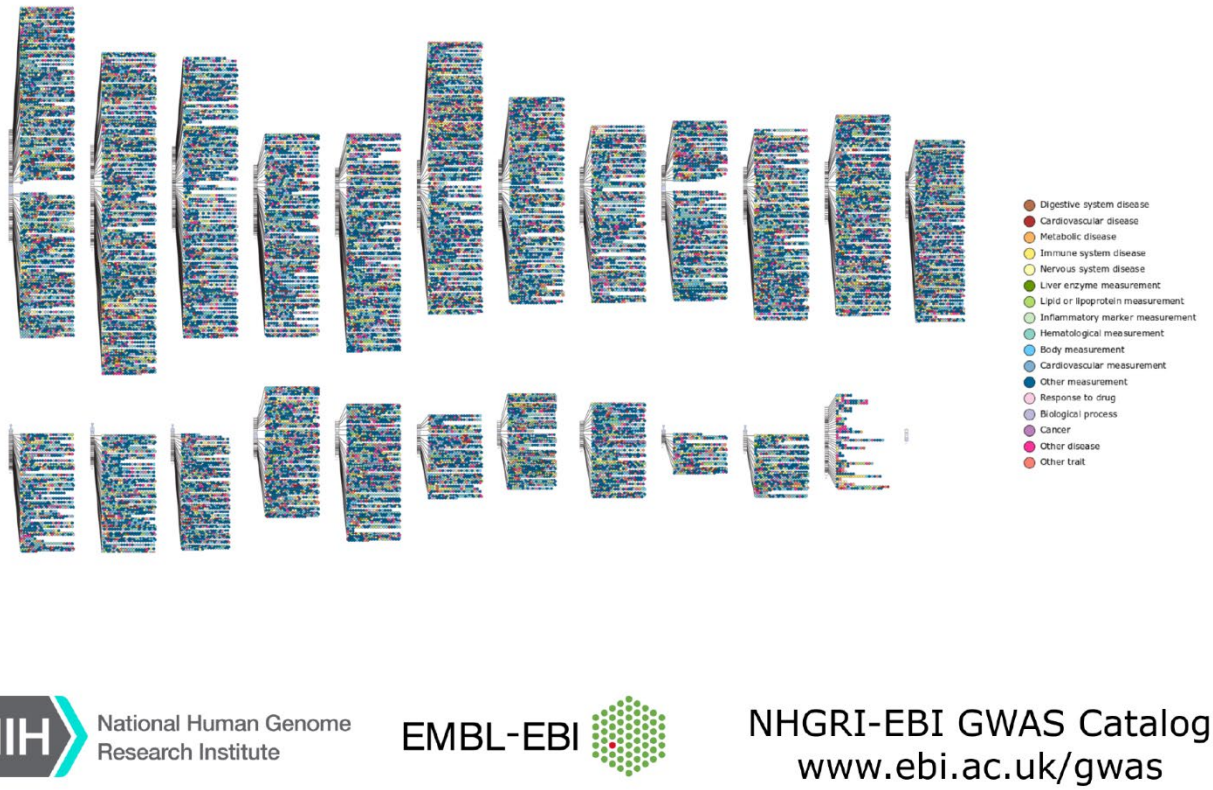


Figure 1.4. NHGRI-EBI GWAS Catalog published Genome-Wide Associations as of July 2019. P ≤ $5x10^{-8}$ for 17 trait categories.

In the African context, GWAS in African populations are limited but are growing in number. African populations have lower LD across the genome compared to European populations. This has the disadvantage that a genotyping of a larger number of variants may be required to detect associations but has the advantage that it enables fine-mapping of disease risk loci that are less tractable in non-African populations (Rotimi et al., 2017).

1.3.1. The Human Hereditary and Health In Africa Consortium

The Human Heredity & Health in Africa (H3Africa) consortium is an NIH and Wellcome Trust-funded initiative that "facilitates fundamental research into diseases on the African continent while also developing infrastructure, resources, training, and ethical guidelines to support a sustainable African research enterprise – led by African scientists, for the African people." (Rotimi et al, 2014). Through the H3Africa consortium (Mulder et al., 2018), a custom ~2.3 million SNP microarray has been developed. This microarray was developed to capture more of the vast genetic diversity seen in African populations and

increase tagging of African variations and capture common variants seen in multiple African populations. This array also increased coverage of known disease-associated genomic regions and captures known SNPs of interest for specific H3Africa projects. It is, therefore, more efficient in capturing African genetic diversity and better suited for GWAS applications in African populations.

One of the many studies within the H3Africa consortium is the Africa Wits-INDEPTH partnership for Genomic studies or AWI-Gen. AWI-Gen is a partnership between the University of the Witswatersrand (Wits) and the International Network for the Demographic Evaluation of Populations and their Health in low and middle-income countries (INDEPTH). The AWI-Gen study aims to examine genomic and environmental factors that influence body composition, body fat distribution and cardiometabolic disease risk in middle-aged African populations.

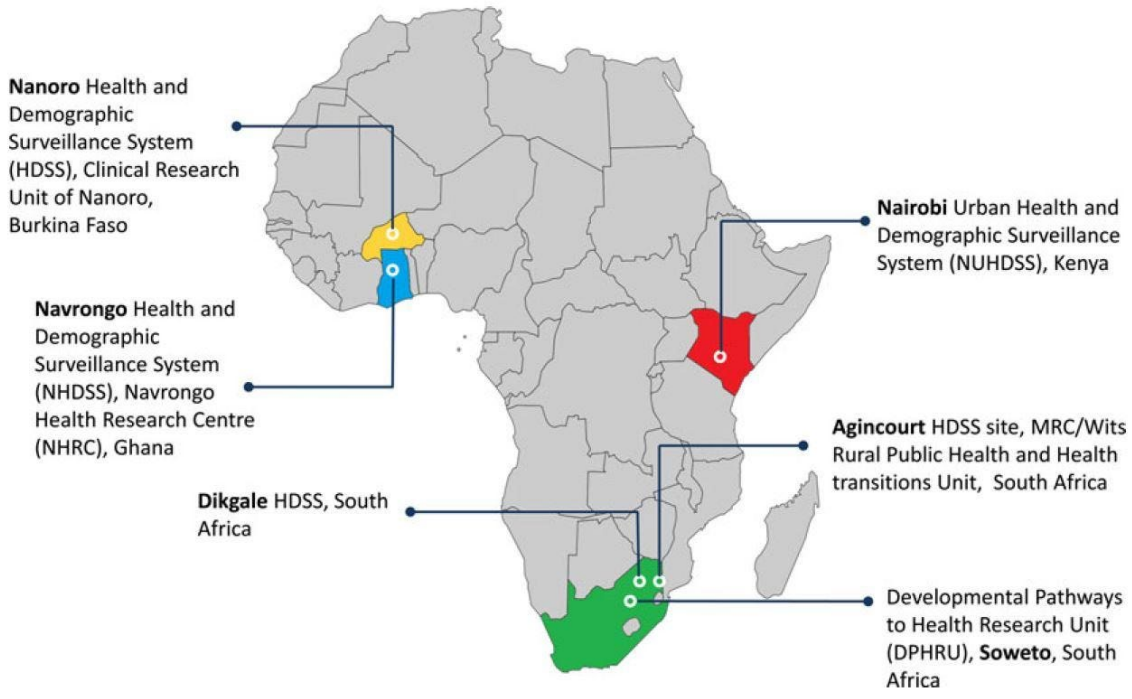


Figure 1.5. The AWI-Gen study's participating Research Centres and enrolment sites.

The AWI-Gen project enrolled and genotyped ~12,000 African participants from six study centres across Africa on the H3Africa Custom microarray (Boua et al., 2020, 2022) (Figure 1.5). Utilizing this custom array, the AWI-Gen project (https://www.wits.ac.za/research/sbimb/research/awi-gen/) has published two GWAS in African populations (Boua et al., 2022, Choudhury et al., 2022). Following the model set

out by the WTCCC, the AWI-Gen project provides access to genome-wide level array data from their study samples. This will allow other African studies to leverage valuable shared population control data to effectively scale African GWAS.

1.4. Genome-wide association studies in oesophageal squamous cell carcinoma

In the context of OSCC, six GWAS and one whole-exome sequencing (WES) study have been performed in Japanese, Chinese and European populations (Cui et al., 2009; Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014), but none in African populations. These studies have reported robust associations at multiple loci, including *ALDH2*, *ADH1B*, *PLCE1*, *PDE4D*, *TMEM173*, *HEATR3*, *TP53*, *MTMR3*, *CASP8*, *RUNX1*, *XPB1/CHEK2*, *CCHCR1*, *TCN2*, *TNXB*, *LTA*, *CYP26B1* and *FASN* (Cui et al., 2009; Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014; Chang et al., 2018). Of these loci, the *PLCE1* locus is the most significant association and has been replicated in several OSCC GWAS in the Chinese population (Abnet et al., 2010, Wu et al., 2014). In a recent comprehensive systematic meta-analysis of the oesophageal cancer candidate gene studies and GWAS, Li et al (2020) summarized the current state of knowledge of known genetic risk factors for both OSCC and OADC. Eleven variants showed a significant association with OSCC with strong or moderate risk in the Asian and European populations (Table 1.2) (Li et al., 2020).

Table 1.2. Genetic variants showing significant summary OR for OSCC meta-analysis. (adapted from Li et al., 2020)

			Number evaluated		Oesophageal cancer risk		*Heterogeneity	
Gene	**Variant**	**Alleles tested**	**Cases**	**Controls**	**OR (95% CI)**	**P-value**	**P-value**	**Evidence of association**
ADH1B	rs1229984	G vs. A	9,203	17,589	1.69 (1.52–1.89)	1.15×10^{-20}	$<1 \times 10^{-3}$	Strong
CHEK2	rs738722	T vs. C	3,153	4,298	1.27 (1.17–1.37)	1.33×10^{-09}	0.467	Strong
CYP1A1	rs1048943	G vs. A	1,947	2,765	1.36 (1.20–1.54)	1.10×10^{-6}	0.161	Strong
CYP2E1	rs3813867	C vs. G	708	712	0.69 (0.58–0.83)	5.59×10^{-5}	0.409	Moderate
PLCE1	rs2274223	G vs. A	24,779	32,211	1.33 (1.27–1.40)	4.68×10^{-30}	0.001	Strong
PTEN	rs701848	C vs. T	955	1,085	1.37 (1.21–1.57)	1.96×10^{-6}	0.778	Strong
LTA	rs909253	G vs. A	4,029	5,133	0.83 (0.78–0.88)	4.33×10^{-9}	0.519	Strong
MTHFR	rs1801133	T vs. C	3,309	4,028	1.30 (1.12–1.51)	4.37×10^{-4}	$<1 \times 10^{-3}$	Moderate
CDKN1A	rs2395655	A vs. G	1,444	1,610	0.75 (0.61–0.92)	0.006	0.273	Strong
ECRG1	rs12422149	A vs. G	1,473	1,762	1.60 (1.13–2.26)	0.008	0.294	Moderate
MIR34BC	rs4938723	C vs. T	2,121	2,408	0.72 (0.59–0.87)	7.55×10^{-4}	0.741	Strong

*Heterogeneity is the measure of the variability of outcomes in the meta-analysis

The initial conclusion from these studies is that there are substantial genetic contributions to OSCC risk in some populations. The identification of such risk factors may provide an understanding of the genetic aetiology of the disease. However, most studies are focused on Asian and European populations. GWAS for OSCC in other populations are needed to fully understand the genetic risk effects for OSCC.

This leads to the need for a substantially larger case-control study that includes cases and controls from major African Black population groups to confirm these interesting preliminary findings both from African replication studies and non-African GWAS. The knowledge generated will form the basis of a more extensive investigation of genetic susceptibility to OSCC in the black African populations, especially from South Africa. The motivation for this study is the limited knowledge of the extent of the genetic contribution to OSCC in the SAB population in comparison to non-African populations, and the high level of disease burden for this cancer in the Black populations of South Africa (Di Pardo et al., 2016; Loots, Ferndale & Clarke, 2019; NCR, 2021). The results from this study may provide evidence for possible candidate genes/risk loci for OSCC that are unique to SAB populations for further investigation in understanding disease aetiology, and to inform functional studies and pathway analysis. The findings from the GWAS may also provide the evidence needed for the construction of a risk prediction score to stratify and identify at-risk individuals from the general population.

1.5. Study Aim and Objectives

The main aim of the study was to test the hypothesis that genetic variation in the SAB population contributes to the risk of OSCC.

Specific Objectives:

1. To assess whether the quality and quantity of genomic DNA isolated from archival blood samples are suitable for large-scale genomic assays.
2. To test known OSCC genetic risk factors identified in other populations for association in the SAB population.

3. To identify novel and potentially African-specific genetic risk factors for OSCC in the SAB population through a GWAS using the H3A Illumina African SNP array.
4. To perform a trans-ethnic meta-analysis of African OSCC GWAS with Chinese OSCC GWAS to identify shared genetic risk factors for OSCC.

This study achieved the objectives in three manuscripts. Manuscript 1 addressed objective 1, manuscript 2 addressed objective 2 and manuscript 3 addressed objectives 3 and 4.

1.6. References:

Abnet, C.C., Freedman, N.D., Hu, N., et al. (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature Genetics*. 42(9):764–767. DOI: 10.1038/ng.649.

Arnold, M., Soerjomataram, I., Ferlay, J., et al. (2015) Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*. 64(3):381–387. DOI: 10.1136/gutjnl-2014-308124.

Boua, P.R., Brandenburg, J.T., Choudhury, A., et al. (2020) 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*. 10(February). DOI: 10.3389/fgene.2019.01354.

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):855. DOI: 10.1038/s41467-022-28276-x.

Buniello, A., Macarthur, J.A.L., Cerezo, M., et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. 47(D1):D1005–D1012. DOI: 10.1093/nar/gky1120.

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):661–678. DOI: 10.1038/nature05911.

Bye, H., Prescott, N.J., Lewis, C.M., et al. (2012) Distinct genetic association at the *PLCE1* locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis*. 33(11):2155–2161. DOI: 10.1093/carcin/bgs262.

Bye, H., Prescott, N.J., Matejcic, M., et al. (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*. 32(12):1855–1861. DOI: 10.1093/carcin/bgr211.

Chang, J., Zhong, R., Tian, J., et al. (2018) Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics*. 50(3):338–343. DOI: 10.1038/s41588-018-0045-8.

Chelule, P.K., Pegoraro, R.J., Gqaleni, N., et al. (2006) The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. *Disease markers*. 22(5–6):351–4. DOI: 10.1155/2006/980392.

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(March):101701. DOI: 10.1016/j.canep.2020.101701.

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 Comms*. 13(2578). DOI: 10.1038/s41467-022-30098-w.

Cui, R., Kamatani, Y., Takahashi, A., et al. (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*. 137(5):1768–1775. DOI: 10.1053/j.gastro.2009.07.070.

Dandara, C., Ballo, R. & Parker, M.I. (2005) CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer letters*. 225(2):275–82. DOI: 10.1016/j.canlet.2004.11.004.

Dandara, C., Li, D.P., Walther, G. et al. (2006). Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis*. 27(4):791–7. DOI: 10.1093/carcin/bgi257.

Di Pardo, B.J., Bronson, N.W., Diggs, B.S., et al. (2016) The Global Burden of Esophageal Cancer: A Disability-Adjusted Life-Year Approach. *World journal of surgery*. 40(2):395–401. DOI: 10.1007/s00268-015-3356-2.

Dietzsch, E. & Parker, M.I. (2002) Infrequent somatic deletion of the 5' region of the COL1A2 gene in oesophageal squamous cell cancer patients. *Clinical chemistry and laboratory medicine*. 40(9):941–5. DOI: 10.1515/CCLM.2002.165.

Dietzsch, E., Laubscher, R. & Parker, M.I. (2003) Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *International journal of cancer*. 107(1):38–45. DOI: 10.1002/ijc.11314.

Eltahir, H.A., Adam, A.A.M., Yahia, Z.A., et al. (2012) p53 Codon 72 arginine/proline polymorphism and cancer in Sudan. *Molecular biology reports*. 39(12):10833–6. DOI: 10.1007/s11033-012-1978-0.

Enzinger, P.C. & Mayer, R.J. (2003) Esophageal Cancer. *New England Journal of Medicine*. 349(23):2241–2252. DOI: 10.1056/NEJMra035010.

Ferndale, L., Ayeni, O., Chen, W.C., et al. (2021) Factors Influencing Survival Time in Patients Treated for Oesophageal Cancer with Palliative Intent in South Africa. *ResearchSquare-Preprint*. 1–18. DOI: 10.21203/rs.3.rs-1074694/v1.

Ferndale, L., Sartorius, B., Aldous, C., et al. (2019) Oesophageal cancer in Area 2 of KwaZulu-Natal: predictors of late presentation. *South African Journal of Surgery*. 57(2 SE-General Surgery). Available: http://sajs.redbricklibrary.com/index.php/sajs/article/view/2948.

Forman, D., Bray, F., Brewster, D.H., et al. (2013) Cancer Incidence in Five Continents, Vol. X. Lyon: IARC Scientific Publications, No 164.

Gamieldien, W., Victor, T.C., Mugwanya, D., et al. (1998) p53 and p16/CDKN2 gene mutations in esophageal tumors from a high-incidence area in South Africa. *International journal of cancer*. 78(5):544–9. DOI: 10.1002/(sici)1097-0215(19981123)78:5<544::aid-ijc3>3.0.co;2-t.

Halec, G., Schmitt, M., Egger, S., et al. (2016) Mucosal alpha-papillomaviruses are not associated with esophageal squamous cell carcinomas: Lack of mechanistic evidence from South Africa, China and Iran and from a world-wide meta-analysis. *International journal of cancer*. 139(1): 85-98. DOI: 10.1002/ijc.29911.

Isaacson, C., Mothobi, P., Hale, M., et al. (2015) Carcinogenic nitrosamines in traditional beer as the cause of oesophageal squamous cell carcinoma in black South Africans. *South African Medical Journal*. 105(8):656. DOI: 10.7196/SAMJnew.7935.

Kayamba, V., Bateman, A.C., Asombang, A.W., et al. (2015) HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case–control study. *Cancer Medicine*. 4(4):588–595. DOI: 10.1002/cam4.434.

Klein, R.J., Zeiss, C., Chew, E.Y., et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 308(5720):385–389. DOI: 10.1126/science.1109557.

Lagisetty, K.H., Beer, D.G., & Chang, A.C. (2017) Squamous cell carcinomas and adenocarcinomas of the esophagus: One treatment does not rule them all. The *Journal of Thoracic and Cardiovascular Surgery*. 154: 1446-7. DOI: 10.1016/j.jtcvs.2017.04.074.

Li, D., Dandara, C. & Parker, M.I. (2005) Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clinical chemistry and laboratory medicine*. 43(4):370–5. DOI: 10.1515/CCLM.2005.067.

Li, D., Dandara, C. & Parker, M.I. (2010) The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC genetics*. 11:47. DOI: 10.1186/1471-2156-11-47.

Li, D.P., Dandara, C., Walther, G., et al. (2008) Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer. *Clinical chemistry and laboratory medicine*. 46(3):323–8. DOI: 10.1515/CCLM.2008.073.

Li, G., Song, Q., Jiang, Y., et al. (2020) Cumulative Evidence for Associations between Genetic Variants and Risk of Esophageal Cancer. *Cancer Epidemiology Biomarkers & Prevention*. 29(4):838–849. DOI: 10.1158/1055-9965.EPI-19-1281.

Liu, W., Snell, J.M., Jeck, W.R., et al. (2017) Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI insight*. 2(22). DOI: 10.1172/jci.insight.98457.

Loots, E., Ferndale, L. & Clarke, D. (2019) Oesophageal cancer in South Africa: the need for a public health response. *South African Journal of Surgery*. 57(2 SE-Editorial). Available: http://sajs.redbricklibrary.com/index.php/sajs/article/view/3094.

Matejcic, M., Li, D., Prescott, N.J., et al. (2011) Association of a Deletion of GSTT2B with an Altered Risk of Oesophageal Squamous Cell Carcinoma in a South African Population: A Case-Control Study. *PLoS ONE*. 6(12):e29366. DOI: 10.1371/journal.pone.0029366.

Matejcic, M., Vogelsang, M., Wang, Y. et al. (2015) NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC cancer*. 15:658. DOI: 10.1186/s12885-015-1105-4.

Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nature Reviews Genetics*. 9(1):9–14. DOI: 10.1038/nrg2203.

McKay, J.D., Truong, T., Gaborieau, V., et al. (2011) A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. *PLoS Genetics*. 7(3):e1001333. DOI: 10.1371/journal.pgen.1001333.

Menya, D., Kigen, N., Oduor, M., et al. (2019) Traditional and commercial alcohols and esophageal cancer risk in Kenya. *International journal of cancer*. 144(3):459–469. DOI: 10.1002/ijc.31804.

Menya, D., Maina, S.K., Kibosia, C., et al. (2019) Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case-control study and a pan-African perspective. *International journal of cancer*. 145(1):99–109. DOI: 10.1002/ijc.32086.

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

Murphy, G., McCormack, V., Abedi-Ardekani, B., et al. (2017) International cancer seminars: a focus on esophageal squamous cell carcinoma. *Annals of Oncology*. 28(9):2086–2093. DOI: 10.1093/annonc/mdx279.

Naidoo, R., Ramburan, A., Reddi, A. et al. (2005) Aberrations in the mismatch repair genes and the clinical impact on oesophageal squamous carcinomas from a high incidence area in South Africa. *Journal of clinical pathology*. 58(3):281–4. DOI: 10.1136/jcp.2003.014290.

NCR. (2021) South African summary statistics of cancer diagnosed histologically in 2019. https://www.nicd.ac.za/wp-content/uploads/2021/12/NCR_Path_2019_Full_Report_8dec2021.pdf.

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):1751–6. DOI: 10.1038/sj.bjc.6600338.

Patel, K., Mining, S., Wakhisi, J., et al. (2011) TP53 mutations, human papilloma virus DNA and inflammation markers in esophageal squamous cell carcinoma from the Rift Valley, a high-incidence area in Kenya. *BMC research notes*. 4:469. DOI: 10.1186/1756-0500-4-469.

Reichenbach, Z.W., Murray, M.G., Saxena, R., et al. (2019) Clinical and translational advances in esophageal squamous cell carcinoma. *In Advances in Cancer Research*. V. 144. Academic Press Inc. 95–135. DOI: 10.1016/bs.acr.2019.05.004.

Rotimi, C.N., Abayomi, A., Abimiku A., et al. (2014) Research capacity. Enabling the genomic revolution in Africa. *Science*. 344(6190):1346-1348. DOI: 10.1126/science.1251546.

Rotimi, C.N., Bentley, A.R., Doumatey, A.P., et al. (2017) The genomic landscape of African populations in health and disease. *Human Molecular Genetics*. 26(R2):R225–R236. DOI: 10.1093/hmg/ddx253.

Schäfer, G., Kabanda, S., van Rooyen, B., et al. (2013) The role of inflammation in HPV infection of the Oesophagus. *BMC Cancer*. 13(1):185. DOI: 10.1186/1471-2407-13-185.

Segal, I., Reinach, S. & de Beer, M. (1988) Factors associated with oesophageal cancer in Soweto, South Africa. *British Journal of Cancer*. 58(5):681–686. DOI: 10.1038/bjc.1988.286.

Sengupta, D., Choudhury, A., Fortes-Lima, C., et al. (2021) Genetic substructure and complex demographic history of South African Bantu speakers. *Nature communications*. 12(1):2080. DOI: 10.1038/s41467-021-22207-y.

Siewert, J.R. & Ott, K. (2007) Are squamous and adenocarcinomas of the esophagus the same disease? *Seminars in Radiation Oncology*. 17(1): 38-44. DOI: 10.1016/j.semradonc.2006.09.007.

Simba, H., Kuivaniemi, H., Lutje, V., et al. (2019) Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in Genetics*. DOI: 10.3389/fgene.2019.00642.

Strickland, N.J., Matsha, T., Erasmus, R.T., et al. (2012) Molecular analysis of ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International journal of cancer*. 131(3):623–32. DOI: 10.1002/ijc.26418.

Sud, A., Kinnersley, B. & Houlston, R.S. (2017) Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*. 17(11):692–704. DOI: 10.1038/nrc.2017.82.

Sun, J., Shen, L., Shah, M.A., et al. (2021) Pembrolizumab plus chemotherapy versus chemotherapy alone for first-line treatment of advanced oesophageal cancer (KEYNOTE-590): a

randomised, placebo-controlled, phase 3 study. *The Lancet*. 398(10302):759-771. DOI: 10.1016/S0140-6736(21)01234-4.

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):209–249. DOI: 10.3322/caac.21660.

The Cancer Genome Atlas Research Network. (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature*. 541: 169-175. DOI: 10.1038/nature20805.

Thumbs, A., Borgstein, E., Vigna, L., et al. (2012) Self-expanding metal stents (SEMS) for patients with advanced Esophageal cancer in Malawi: An effective palliative treatment. *Journal of Surgical Oncology*. 105(4):410–414. DOI: 10.1002/jso.23003.

Victor, T., Du Toit, R., Jordaan, A.M., et al. (1990) No evidence for point mutations in codons 12, 13, and 61 of the ras gene in a high-incidence area for esophageal and gastric cancers. *Cancer research*. 50(16):4911–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/2199031.

Vogelsang, M., Wang, Y., Veber, N., et al. (2012) The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PLoS One*. 7(5):e36962. DOI: 10.1371/journal.pone.0036962.

Vos, M., Adams, C.H., Victor, T.C., et al. (2003) Polymorphisms and mutations found in the regions flanking exons 5 to 8 of the TP53 gene in a population at high risk for esophageal cancer in South Africa. *Cancer genetics and cytogenetics*. 140(1):23–30. DOI: 10.1016/s0165-4608(02)00638-6.

Wang, L.-D., Zhou, F.-Y., Li, X.-M., et al. (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at *PLCE1*. *Nature genetics*. 42(9):759–763. DOI: 10.1038/ng.648.

White, R.E., Parker, R.K., Fitzwater, J.W., et al. (2009) Stents as sole therapy for oesophageal cancer: a prospective analysis of outcomes after placement. *The Lancet Oncology*. 10(3):240–246. DOI: 10.1016/S1470-2045(09)70004-X.

Wu, C., Hu, Z., He, Z., et al. (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature genetics*. 43(7):679–684. DOI: 10.1038/ng.849.

Wu, C., Kraft, P., Zhai, K., et al. (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nature genetics*. 44(10):1090–1097. DOI: 10.1038/ng.2411.

Wu, C., Wang, Z.Z., Song, X., et al. (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nature genetics*. 46(9):1001–1006. DOI: 10.1038/ng.3064.

Zaahl, M.G., Warnich, L., Victor, T.C., et al. (2005) Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer genetics and cytogenetics*. 159(1):48–52. DOI: 10.1016/j.cancergencyto.2004.09.017.

Zhang, Y.D., Hurson, A.N., Zhang, H, et al. (2020) Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications*. 11, 2253. DOI: 10.1038/s41467-020-16483-3.

2. CHAPTER 2: Summary of Methods

This chapter outlines the approach and summarises the methodology used in each manuscript. Details of the methods used are described in the 3 manuscripts which comprise chapters 3, 4 and 5.

2.1. Overview of study design

The association studies reported in this thesis took the form of case-control studies utilizing retrospectively as well as prospectively recruited patients. The case populations were OSCC cases enrolled from three different sites in South Africa: 1. The Johannesburg Cancer Study (JCS) (n=942); 2. University of Cape Town (UCT) (n=477); 3. Greys Hospital, Pietermaritzburg (n=267). This study utilized ethnically matched population controls who were recruited for cardiometabolic studies and were not known to be affected with cancer, following the principle of shared controls established by the Wellcome Trust Case Control Consortium (Burton et al., 2007). The control populations were enrolled in the AWI-Gen study (n=2731) and the JCS (n=487).

2.2. Objective 1: Assessing the quality of DNA samples

Objective 1 set out to assess whether the quality and quantity of genomic DNA isolated from archival blood samples are suitable for large-scale genomic assays. This objective established the feasibility of using archival blood samples collected and banked by the JCS for large-scale downstream molecular studies.

Patients with OSCC and non-cancer controls from the South African Black population (SAB) were recruited through the JCS in the southern Gauteng province (Chen et al., 2020). The JCS aimed to interview and recruit newly diagnosed Black adult cancer patients attending Johannesburg tertiary public referral hospitals for medical oncology. Trained nurses/interviewers recruited, consented, interviewed and drew blood from cases and non-cancer controls. Questionnaires covered lifestyle, socio-demographics and environmental exposures. Language spoken by the participant and their parents were asked and can act as a proxy for the population group. Pathology reports were reviewed to confirm cancer diagnosis and type. Cases were already recruited patients with

histologically confirmed primary OSCC. The JCS also recruited a series of non-cancer controls which we used for our candidate gene study and GWAS. DNA isolation was performed using the Qiagen DNA FlexiGene kit. The quality and quantity of genomic DNA isolated from blood samples stored for 2 years, up to 19 years were assessed by Nanodrop quantification, Qubit quantification and gel electrophoresis. Quality and quantity metrics were assessed using non-parametric tests.

2.3. Objective 2: Replication of known OSCC genetic risk factors

Objective 2 set out to test previously known OSCC genetic risk factors identified in other populations for replication and association with OSCC in the SAB population. This objective provided insight into the transferability of risk loci from other populations to the SAB population.

Genetic variants with reported genome-wide significant association with OSCC identified in other populations were selected based on the available evidence at the time of the study. These variants were genotyped in cases and controls from the JCS and UCT. Chi-Squared tests and meta-analysis were used to determine whether the previously identified loci are also associated with OSCC in the SAB population and if so, to estimate the size of the effect (odds ratio).

2.4. Objectives 3 and 4: Novel OSCC genetic risk factor discovery and transethnic meta-analysis

Objective 3 was to identify novel, African-specific genetic risk factors for OSCC in the SAB population utilizing a GWAS, and objective 4 was to identify shared genetic risk factors for OSCC in the SAB and Chinese populations by a meta-analysis of GWAS from these two populations.

Currently, the majority of the commercially available SNP-based microarrays are designed for use in European populations. These Euro-centric SNP microarrays are not ideal for the use of GWAS in the SAB populations (and African populations in general) due to the large degree of genomic differences between the populations. African populations are known to be very diverse (Choudhury et al., 2017, 2020, 2021), with a higher degree of genetic

variation and a lower degree of linkage disequilibrium, thus an African SNP-based microarray provides more extensive coverage for GWAS in SAB populations.

The SAB population is made up of multiple linguistic and ethnic groups that share common Bantu-speaking ancestry (Choudhury et al., 2017, 2020, 2021; Sengupta et al., 2021). Modern genetic studies have shown that the major groups such as Zulu, Xhosa and Sotho speakers are closely related genetically (Choudhury et al., 2020). The JCS, first established in 1995, enrolled treatment naïve cancer patients diagnosed within health facilities around Johannesburg, South Africa (Chen et al., 2020). The JCS used the languages spoken by the individuals and their parents as well as their country of birth as proxies to identify the ethnic groups of the participants. With the availability of micro-array data containing millions of genetic markers, our genetic studies can use appropriate advanced statistical methodology to adjust for the effects of population stratification and potentially discover population-specific susceptibility loci for OSCC in the SAB population.

The H3 Africa Consortium, in collaboration with Illumina, has designed a genome-wide SNP microarray specifically for use in African black populations (Mulder et al., 2018). This is a ~2.3M Illumina SNP array, comprising a combination of existing bead pools and custom content based on the H3A design derived from whole-genome sequencing of ±3,000 African genomes, to capture >80% of the common variation (Gurdasani et al., 2015, 2019).

Genotyping of H3A African SNP arrays was done at the IoPNN Genomics & Biomarker Core Facility at King's College London, led By Dr Gerome Breen and Dr Charles Curtis (http://www.kcl.ac.uk/ioppn/depts/mrc/research/The-IoPPN-Genomics--Biomarker-Core-Facility.aspx). They are a CSPro-certified provider of Illumina's platforms and have an excellent track record in high throughput GWAS genotyping. The GWAS genotyping was done as part of a larger ongoing cancer study, the ERICA-SA Study (https://www.samrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa) (SAMRC/Newton fund for Non-communicable diseases). After quality control to remove poorly genotyped SNPs and individuals and to correct for factors such as population stratification and relatedness, SNPs were imputed to whole-genome sequences from the Sanger Imputation Service using the African Genome

Resources Panel containing at least 3,500 SAB genomes.SNP allele frequencies in 1,686 SAB OSCC cases were compared with their frequencies in 3,217 ethnically matched population controls from the H3 Africa AWI-Gen study and the JCS (Mulder et al., 2018; Chen et al., 2020) to identify novel risk loci or African specific loci for OSCC in this population.

Initial power calculation for the estimated sample size (1,471 cases; 1,791 controls) to be used for objective 2 provides 99% power to detect a risk allele with an odds ratio of 1.3 and a minor allele frequency of 0.35 at $\alpha = 0.01$ (Quanto v1.2) (Gauderman & Morrison, 2006). For the GWAS (Objective 3), 1,686 cases and 3,217 controls provide 70% power to detect a risk allele with a minor allele frequency of 0.35 and odds ratio of 1.3 at genome-wide significance ($P < 5x10^{-8}$).

The replication of the African OSCC GWAS was done by a trans-ethnic meta-analysis using published Chinese OSCC GWAS data (Abnet et al., 2010) to identify both distinct and shared risk loci for OSCC. The Chinese GWAS comprised 2,013 Chinese OSCC cases and 2,701 ethnically matched controls, with 4.8 million SNPs overlapping in both GWAS and were used for the meta-analysis. The meta-analysis identified OSCC risk loci common to both African and non-African populations.

2.5. Ethical clearance and certification:

This PhD study was approved by the University of the Witwatersrand, Human Research Ethics Committee (Medical), clearance number: M1706115. Details of all relevant ethical approvals are as follows.

The JCS received ethics approval from the University of the Witwatersrand, Human research ethics committee (Medical), clearance number: M1606103 (previously M981119, M040445, M090361 and M140271) for the collection of peripheral blood samples and epidemiological questionnaires.

Study samples were provided by the University of Cape Town (UCT) through ethical approval from the joint University of Cape Town/Groote Schuur Hospital Research Ethics

Committee and the University of Stellenbosch/Tygerberg Hospital Ethics Committee, certificate number: UCT HREC 040/2005.

Study samples were provided by the Greys Hospital (GH) through ethical approval by the Biomedical Research Ethics Committee of the University of the KwaZulu-Natal (UKZN), certificate number: BF270/15.

Genotype data was provided by the AWI-Gen study under the ethical approval by the Wits HREC (Medical), certificate numbers: M180535 and M121029.

For objective 1: Biobanking of the extracted DNA samples collected through the JCS is approved by the Wits HREC (Medical), clearance number: M120117. The Sydney Brenner Institute for Molecular Biosciences Biobank (SBIMB), University of the Witwatersrand, curated and stored all DNA samples from this study. The SBIMB Biobank has Wits HREC (Medical) approval, certificate number: M200469 and BEC20200401.

For objectives 2,3 and 4: This study forms part of the larger study: Evolving Risk Factors for Cancers in African Populations (ERICA-SA) (https://www.samrc.ac.za/intramural-research-units/evolving-risk-factors-cancers-african-populations-erica-sa). The ERICA-SA study is approved by the Wits HREC (Medical), clearance number: M2111154 (previously M160807). Ethical clearance certificates can be found in Appendix B.

2.6. References:

Abnet, C.C., Freedman, N.D., Hu, N., et al. (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics*. 42(9):764–767. DOI: 10.1038/ng.649.

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):661–678. DOI: 10.1038/nature05911.

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(March):101701. DOI: 10.1016/j.canep.2020.101701.

Choudhury, A., Ramsay, M., Hazelhurst, S., et al. (2017) Whole-genome sequencing for an

enhanced understanding of genetic variation among South Africans. *Nature Communications*. 8(1):1–12. DOI: 10.1038/s41467-017-00663-9.

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

Choudhury, A., Sengupta, D., Ramsay, M., et al. (2021) Bantu-speaker migration and admixture in southern Africa. Human Molecular Genetics. 30(R1):R56–R63. DOI: 10.1093/hmg/ddaa274.

Gauderman, W.J. & Morrison, J.M. (2006) QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., et al. (2015) The African Genome Variation Project shapes medical genetics in Africa. *Nature*. 517(7534):327–332. DOI: 10.1038/nature13997.

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

Mulder, N., Abimiku, A., Adebamowo, S.N., et al. (2018) H3Africa: current perspectives. *Pharmacogenomics and personalized medicine*. 11:59–66. DOI: 10.2147/PGPM.S141546.

Sengupta, D., Choudhury, A., Fortes-Lima, C., et al. (2021) Genetic substructure and complex demographic history of South African Bantu speakers. Nature Communications. 12(1):2080. DOI: 10.1038/s41467-021-22207-y.

3. CHAPTER 3: The Integrity and Yield of Genomic DNA Isolated from Whole Blood following long-term storage at -30°C.

The manuscript "The Integrity and Yield of Genomic DNA Isolated from Whole Blood following long-term storage at -30°C." set out to assess the quality and quantity of genomic DNA isolated from archival peripheral blood samples in long-term storage. High quality and sufficient quantity of genomic DNA is essential for large-scale downstream genetic analysis. This chapter demonstrated that good quality gDNA of sufficient quantity can be isolated from archival blood samples following long-term storage (2 to 19 years), thus unlocking a valuable resource collected by the JCS. The work described in this chapter was published in the peer-reviewed journal, Biopreservation and Biobanking (Appendix C).

Chen, W.C., Kerr, R., May, A., Ndlovu, B., Sobalisa, A., Duze, S.T., Joseph, L., Mathew, C.G., Babb de Villiers, C. (2018) The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term Storage at –30°C. *Biopreservation and Biobanking*. 16(2):106–113. DOI: 10.1089/bio.2017.0050.

The student performed a significant amount of wet laboratory work; performed data cleaning, and data analysis and wrote and reviewed the manuscript.

3.1. Abstract

Long-term storage of whole blood can affect the integrity of DNA if it is not done under optimal conditions. The aim of this study was to determine whether long-term storage (2 - 19 years) of whole blood samples at -30°C had a negative effect on the quality or quantity of genomic DNA that could be recovered at extraction. Genomic DNA was isolated from 2,758 whole blood samples collected in 4 ml EDTA vacutainers from 1997 to 2012. DNA was extracted using the Qiagen® FlexiGene® DNA kit. The average storage duration at -30°C was 12 years. The quality and quantity of the isolated DNA were assessed using spectrophotometry (NanoDrop™), a fluorometric assay for double-stranded DNA (Qubit™) and agarose gel electrophoresis. The mean DNA yield per sample was found to be 114 µg from whole blood volumes which ranged from 0.5 ml to 4 ml. The mean A260/280 ratio and median A260/280 ratios were both 1.8. No correlation was found

between the duration of storage and the total yield or the quality of DNA extracted. These data suggest that high-quality DNA can be extracted from whole blood samples that are stored at -30°C for up to 19 years.

3.2. Introduction

DNA can be isolated from almost all body tissues including bodily fluids (Nederhand et al., 2003). Due to its relative ease of acquisition, whole blood provides one of the most common sources of high quantity and quality DNA for molecular applications, including clinical and epidemiological studies (Miller, Dykes & Polesky, 1988).

Large genetic epidemiological studies can involve the collection of tens of thousands of valuable whole blood samples that can be or have been, stored for many years. The blood volume available for DNA extractions is an obvious indicator of DNA yield (Griffiths & Chacon-Cortes, 2014). However, long-term storage conditions of whole blood can affect the integrity of DNA (Holland et al., 2003). Blood samples stored at 4°C for a short period of time will still yield DNA of acceptable quality provided the correct blood collection tubes were used (Nederhand et al., 2003; Richardson et al., 2006). However, the desired temperature for whole blood samples for DNA isolation is -80°C for long-term storage (Holland et al., 2003). Whole blood samples can also be frozen at -20°C for long-term storage (Steinberg et al., 1997; AlRokayan, 2000; Holland et al., 2003), however, other studies have found lower DNA yields following this approach (AlRokayan, 2000; Nederhand et al., 2003). A study by Di Pietro *et al* (2011) found that whole blood samples stored for long periods at -20°C yielded high-quality DNA for genotyping studies, although the sample size was small (n=82) (Di Pietro et al., 2011).

The current study sought to determine whether the duration of the long-term storage of whole blood samples, using a large sample size (n=2,758), stored at a stable -30°C, has a negative effect on the quality and quantity of the genomic DNA extracted.

3.3. Materials and Methods

Samples

Samples analyzed were sourced from the Johannesburg Cancer Case-Control Study (JCS) at the National Cancer Registry (NCR) of South Africa. The JCS is a large, recently (2016) ended, case-control study established in 1995 (Urban et al., 2012). In brief, the JCS recruited newly diagnosed cancer patients of black African ancestry. These patients were interviewed by trained research nurses and their blood was collected with informed consent for future research as approved by the Research Ethics Committee at the University of the Witwatersrand and described by Urban et al (2012). Blood samples were drawn from the patients in 4 ml EDTA vacutainers and were stored in -30°C freezers until further use. More than 20,000 whole blood samples (representing various cancer types) have been collected since March 1995 for the JCS, of which a subset of 2,758 blood samples was used for this study. DNA extractions were performed on request for specific studies in collaboration with the JCS. At the time of the study being reported here, DNA had been extracted from 2,758 blood samples for four different studies. Ethical approval was obtained from the University of the Witwatersrand Human Research Ethics Committee (Medical), clearance numbers: M140271 and M120117.

Sample Storage

All blood samples collected under the JCS were kept at room temperature until collected for transport. The blood samples were then transported in a cooler box with an ice block from the three separate study sites to a centralized storage facility. On the rare occasion that there was an issue with the transportation, samples were frozen at the collection site and collected at the next available time. The average transit time was ≤ 30 minutes. All samples were frozen in their collection tubes in an upright position at -30°C within a maximum of 24 hours of collection. Miele™ -30°C general freezers were used for sample storage. The freezers' temperatures were monitored with an alarm-based system and backup freezers were available. As part of our risk management strategy, when a freezer failed, the samples were transferred to another freezer prior to thawing. All the blood samples were frozen in an upright position and thereafter stored on their sides. Although the blood was collected in a 4 ml EDTA tube, not all were filled to capacity; therefore, the volume of the blood sample available for extraction was noted prior to DNA extraction.

Sample information was managed using Microsoft Excel and then incorporated into a STATA database.

DNA Isolation

Whole blood samples collected between 1997 and 2012 (n=2,758) were extracted. DNA extraction procedures occurred in batches between May 2013 and June 2016. Six different technicians were involved over the four-year period. The technicians were trained on the extraction protocol, blinded to the storage duration of the blood samples and each technician extracted samples collected over the whole time period. Total genomic DNA was isolated using the FlexiGene® DNA Kit (Qiagen®). This kit was chosen initially for its ability to handle a diverse range of input blood volume and then used subsequently for consistency. Extractions were undertaken following the manufacturer's protocol, with the following minor modifications: the use of a dry heat incubator instead of a hot water bath, and resuspension of the DNA pellet in 1x TE buffer (Invitrogen™) instead of the supplied FG3 buffer. To ensure no wastage and optimum yield, the entire blood volume in a tube was used for DNA extraction. Two fixed protocols, one for samples with <2 ml of blood and the other for >2 ml, were optimized and used per the manufacturer's procedures. The most noticeable difference between protocols was the use of 50 ml Nunc tubes for samples with blood volumes of >2 ml, rather than the 15 ml Nunc tubes used for samples with blood volumes of <2 ml. Reagent volumes were adjusted per the sample volume (per manufacturer's protocol). The centrifugation conditions were unchanged. No RNAse digestion was incorporated into the DNA isolation protocol.

DNA Quantity and Quality

DNA quantity and purity were assessed spectrophotometrically using the NanoDrop™ 1000 (Thermo Scientific™) directly after extraction and prior to storage at -30°C. DNA concentration was measured and DNA purity was calculated through the standard A260/280 and A260/230 ratios. This is a fast and accurate technique for determining the DNA concentration of pure samples (Desjardins & Conklin, 2010). DNA/RNA and proteins have their maximum absorbance at 260 and 280 nm, respectively. An A260/280 ratio of ~1.8 is generally accepted as a reflection of a pure DNA sample and a ratio of ~2.0 is generally accepted as pure for RNA. A secondary measurement of the A260/230 ratio is used to determine the purity of DNA against other contaminants. A higher A260/230 is

expected for pure DNA (~2.0-2.2). Low A260/230 ratios indicate the presence of contamination or proteins that absorb light at the 230 nm wavelength (Thermo Scientific™, 2011).

Genomic DNA integrity assessment was done, at a later stage, using gel electrophoresis for 218 samples selected at random (4 – 19 years storage, mean = 12 years) (Figure 3.1). Genomic DNA was resolved on 0.8% agarose gel, stained with GelRed™, at 8 V/cm for 45 minutes against the GeneRuler 1kb Plus DNA Ladder (Thermo Scientific™).

DNA quantitation was additionally assessed using a second methodology. Qubit fluorometry was performed using the Qubit™ 3.0 Fluorometer (Thermo Scientific™). This method uses the double-stranded DNA (dsDNA) BR (Broad Range) assay to quantify dsDNA, as opposed to single-stranded DNA (ssDNA), present in a sample. Only a subset of samples (n=200) were quantified using this chemistry owing to cost restraints. The 200 samples were randomly selected for the year of collection (2 - 19 years of storage), with concentrations representative of the overall DNA concentration range. This measures dsDNA concentration only, which provides a more useful reading as contaminants, including degraded DNA and RNA, are not measured. The %dsDNA was calculated by dividing the Qubit™ concentration by the Nanodrop™ concentration (which measures all DNA) and multiplying by 100.

Analysis

Descriptive statistics were compiled by years of storage. Means and medians were calculated for DNA yields, DNA concentrations, blood volume available per sample, DNA yield per ml of blood, A260/280 and A260/230. Both means and medians were compiled given the non-normally distributed nature of the data which may limit the mean accuracy. Spearman's ranked correlations were performed to determine whether the duration of storage (years) of the samples, blood volume available for extraction, and the age of participants correlated with the total DNA yields. Samples (n=564) without blood volume data were excluded.

Kruskal-Wallis H tests, with multiple group comparisons, were performed to determine whether DNA yields, A260/280 and A260/230 ratios, differed between the different years

of storage duration. Kruskal-Wallis H tests were also used to determine differences in mean DNA yield per ml of blood across different storage durations as well as to determine differences in mean total DNA yields measured by Qubit™ and differences in mean percentage of dsDNA (%dsDNA) across different storage durations. A nonparametric test for trend was performed to evaluate the relationship between %dsDNA across the storage duration of the blood samples. Quantile regression estimating the conditional median was performed to model the effect of storage duration of the blood samples, blood volume available for DNA extraction, age and gender of the participants on the %dsDNA.

A Two-sample t-test with equal variances was performed to assess the difference in mean DNA yield per 1 ml of blood between the two different DNA extraction protocols as well as mean DNA yield per 1 ml of blood between cancer samples and non-cancer samples and between HIV-positive and HIV negative samples.

3.4. Results

Total genomic DNA was isolated from a total of 2,758 samples collected between 1997 and 2012 for the JCS (Table 3.1). NanoDrop™ data were used to calculate DNA yield and DNA concentration, and to determine A260/280 and A260/230 ratios. Blood samples were stored on average for 11.9 years. The longest storage period of samples before DNA extraction was 19 years (n=125) and the shortest was 2 years (n=31) (Table 3.1). Information on the volume of blood collected was available for 80% (n=2,206) of all samples (Table 3.1). Freezers were monitored closely so that no thawing of samples (resulting from freezer failures) happened during this study.

Table 3.1. Descriptive statistics for the samples used for genomic DNA isolation per year of storage.

Storage Duration (yr)	Sample Size	Mean Blood Volume# (ml)	A260/280				A260/230				Mean DNA Conc. (ng/µl) per Sample$	Median DNA Yield (µg) per Sample$	Mean DNA Yield/ml of Blood (µg)$	Mean DNA Yield (µg) per Sample$
			Median	Range	SD	Mean	Median	Range	SD	Mean				
2	31	-	1.81	1.68-1.88	0.03	1.81	1.69	1.24-2.04	0.18	1.69	305.97	110.88	-	122.39
3	25	-	1.80	1.70-1.90	0.05	1.82	1.60	0.29-2.15	0.45	1.56	266.66	98.46	-	106.66
4	54	2.1	1.74	1.52-1.90	0.08	1.73	1.70	0.76-2.13	0.30	1.63	217.16	62.95	28.81	86.86
5	84	2.6	1.79	1.24-1.91	0.08	1.76	1.80	0.36-2.31	0.39	1.69	224.90	59.10	30.70	89.96
6	102	2.6	1.79	1.51-1.95	0.06	1.78	1.88	0.91-2.26	0.34	1.80	195.36	61.60	30.05	78.14
7	212	3.0	1.80	1.51-2.25	0.06	1.81	1.79	0.29-2.58	0.37	1.73	266.64	96.05	37.20	106.66
8	185	3.3	1.81	1.27-1.97	0.06	1.81	1.66	0.26-2.78	0.39	1.68	278.78	102.72	35.40	111.51
9	198	3.1	1.81	1.58-2.18	0.07	1.81	1.66	0.81-2.56	0.42	1.68	285.70	97.68	41.60	114.28
10	218	3.1	1.80	1,62-2,05	0.06	1.81	1.65	0.57-2.41	0.40	1.66	293.74	97.21	40.62	117.50
11	203	3.1	1.80	1.48-2.15	0.07	1.80	1.69	0.30-2.62	0.46	1.68	289.54	103.96	38.67	115.82
12	194	3.1	1.80	1.26-2.31	0.09	1.80	1.64	0.35-2.34	0.41	1.60	291.38	96.57	37.60	116.55
13	172	3.1	1.80	0.71-2.09	0.11	1.78	1.63	0.19-2.60	0.44	1.62	316.77	100.73	40.87	126.71
14	210	2.9	1.80	0.84-2.18	0.10	1.79	1.69	0.23-2.43	0.42	1.61	294.15	101.16	37.04	117.66
15	113	2.9	1.80	0.32-1.89	0.16	1.77	1.95	0.22-2.43	0.43	1.85	281.76	88.59	38.87	112.71
16	339	3.1	1.79	1.02-3.30	0.11	1.79	1.78	0.22-2.44	0.41	1.71	353.03	109.14	45.55	141.21
17	103	3.0	1.80	0.24-1.91	0.16	1.78	2.07	0.51-2.44	0.33	1.98	252.63	78.42	33.68	101.05
18	192	3.1	1.80	1.10-1.96	0.06	1.80	2.08	0.81-2.51	0.25	2.04	273.24	91.02	35.26	109.30
19	124	2.9	1.80	0.56-1.96	0.15	1.77	1.81	0.08-2.49	0.55	1.67	235.57	75.86	32.49	94.23

Mean blood volume per sample
$ Mean DNA concentrations, median and mean DNA yields are calculated using NanoDrop data.

The quality (A260/280 and A260/230 ratios), concentration (ng/µl) and the quantity of isolated DNA (µg), and the mean and median concentration per year of storage for the samples are shown in Table 3.1. The mean A260/280 ratio across all samples was 1.79 and the median A260/280 ratio was 1.8, suggesting good quality DNA was isolated. The mean A260/230 ratio across all samples was 1.71 and the median A260/230 ratio was 1.78, less than 2.0, suggesting the presence of protein, contaminants or RNA residues that absorb at 230 nm or less.

Genomic DNA from 218 samples was subjected to agarose gel electrophoresis to assess possible degradation. High molecular weight staining was observed for all 218 samples. Figure 3.1 shows a representation of the gel electrophoresis result for a representative subset of samples that had been stored for different periods of time. A consistent low level of partial degradation (low intensity of smearing from the genomic band down to 3 kb) was noted across most samples analyzed on gel electrophoresis. The consistency of this smearing pattern suggests that this is likely an artefact from the DNA isolation protocol that caused DNA smearing rather than storage-related DNA degradation.

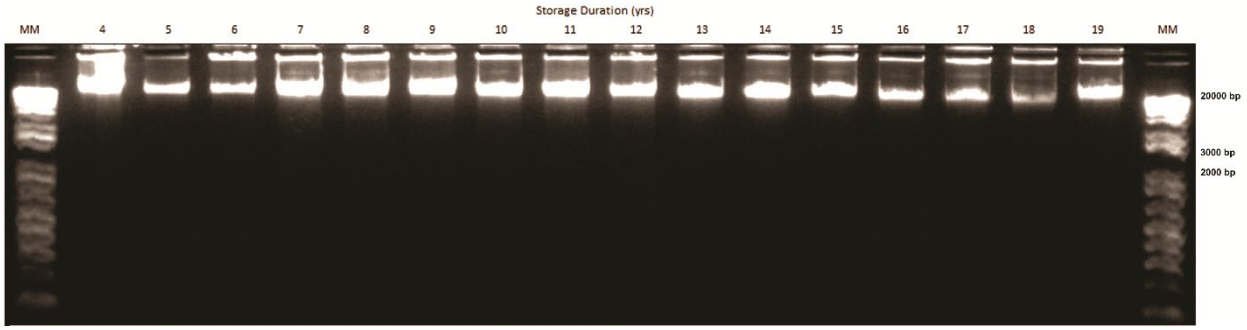


Figure 3.1. Gel image showing the integrity of genomic DNA from 4 years to 19 years of blood sample storage.
One representative sample was used from each year of storage (4 years to 19 years). 2 µl of genomic DNA was resolved on 1.0% agarose gel with GelRed™ staining. GeneRuler™ 1kb Plus DNA Ladder molecular marker (MM) was used for sizing reference. Electrophoresis was done at 8 V/cm for 45 minutes.

The DNA yields of samples per year of storage clustered tightly together, except for a few outlier samples with very high DNA yields or extremely low DNA yields (Figure 3.3). The presence of the outliers illustrates the value of the median as a more informative reporting tool than the mean.

The average yield for all samples was 108.95 µg as measured by NanoDrop™, whereas the average yield for the 200 samples measured by Qubit™ was 86.19 µg per sample. This indicates that approximately 79% of total DNA measured by spectrophotometry was dsDNA, with the remaining proportion presumed to be ssDNA and/or denatured or degraded DNA. A plot of the dsDNA percentage against the storage duration is shown in Figure 3.2.

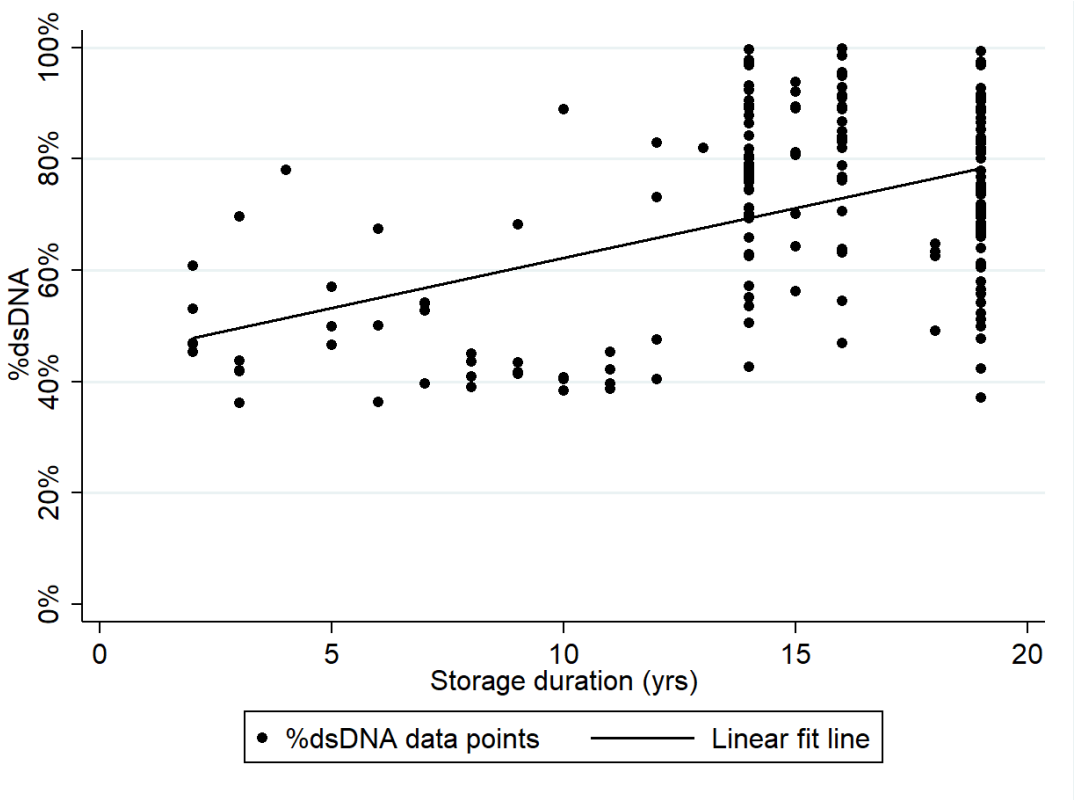


Figure 3.2. %dsDNA vs storage duration.
The linear fit line indicates a positive trend exists between %dsDNA and the increase of storage duration.

A Kruskal-Wallis H test showed no statistically significant differences exist between the mean total DNA yield by Qubit™ measurements and the number of years of storage (P = 0.46) and no statistically significant differences exist between the mean percentage of dsDNA and the number of years of storage (P = 0.10). Nonparametric trend analysis for %dsDNA, quantified by Qubit™ across the storage durations indicated a positive trend (z = 6.24, P < 0.001), with the %dsDNA increasing with increased storage duration of the blood samples. However, the storage duration of the blood samples had a negligible effect on %dsDNA after adjusting for blood volume used for extraction, age and gender (beta = 0.01, P = 0.045).

Spearman's ranked correlation was performed to assess the relationship between total DNA yields, storage duration of blood samples, blood volumes available for extractions, and age of participants at recruitment using 2,194 samples. After Bonferroni correction, there was a strong positive correlation observed between the total blood volumes available for extractions and the total DNA yield, which was statistically significant, r_s = 0.2258, P < 0.0001. So, the more total volume of blood available the more total DNA yield there was. There was no evidence of a correlation between storage duration of blood

samples or age of participants at recruitment with total DNA yields. There was no evidence of a correlation between total blood volume with A260/280 or A260/230 ratios.

A Kruskal-Wallis H test, after Bonferroni correction and correcting for technician variability, showed no statistically significant differences exist between the DNA yields, A260/280 or, A260/230 ratios and the number of years of storage. As expected, total blood volume is a direct indicator of the amount of total DNA available (r_s = 0.2258, P < 0.0001). Similarly, no statistically significant differences were observed between the mean DNA yield per ml of blood and the storage duration of the blood sample (P = 0.25).

The FlexiGene® DNA Kit (Qiagen®) gave an expected DNA yield of 37.5 µg per 1 ml of blood assuming a normal white cell count of 7 x 106 cells/ml of blood (Qiagen®, 2010). JCS sample yield was between 25.83 µg and 46.54 µg (mean = 37.77 µg) of DNA per 1 ml of blood.

The effect of the optimized protocols for different sample blood volumes was assessed using a two-sample t-test with equal variances. Samples with blood volumes of <2 ml or less (n=403) yielded more DNA per 1 ml of blood (44.12 µg, 95% CI: 39.73 - 48.51 µg) when compared to DNA yielded per 1 ml of blood for samples with blood volumes of more than 2 ml (n=1,792) (36.79 µg, 95% CI: 35.42 - 38.16 µg) (P = 0.0001). The typical DNA yield of the FlexiGene® kit for a 2 ml volume would be 33-45 ug/ml. No difference in mean DNA yield per 1 ml of blood was observed between cancer samples (n=1,482) and non-cancer samples (n=707) (P = 0.50). A difference in mean DNA yield per 1 ml of blood was observed between HIV positive (n=1,851) and HIV negative samples (n=313), with HIV negative samples (38.70 µg, 95% CI: 37.23 - 40.17 µg) yielding ~5 µg DNA more per ml than HIV positive samples (33.36 µg, 95% CI: 29.39 - 37.32 µg) (P = 0.0076).

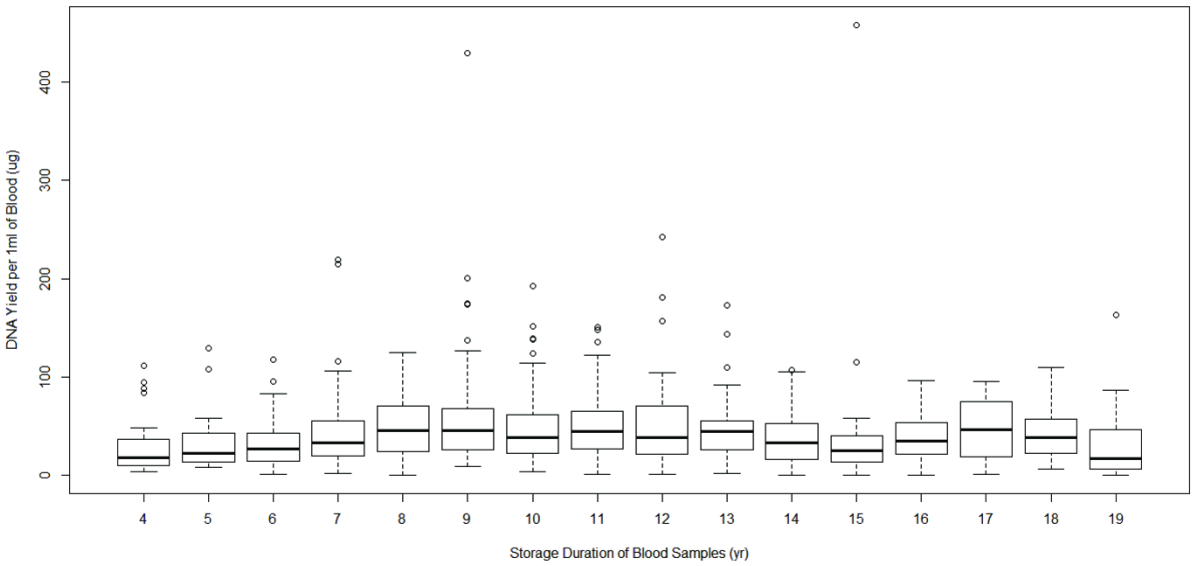


Figure 3.3. Range and distribution of DNA yields per 1ml of blood, extracted from frozen whole blood stored over 4 to 19 years.
Blood volume data was not available for 2 and 3 years of storage. No statistically significant difference in yield was observed across the duration of blood storage. DNA yields were measured by NanoDrop.

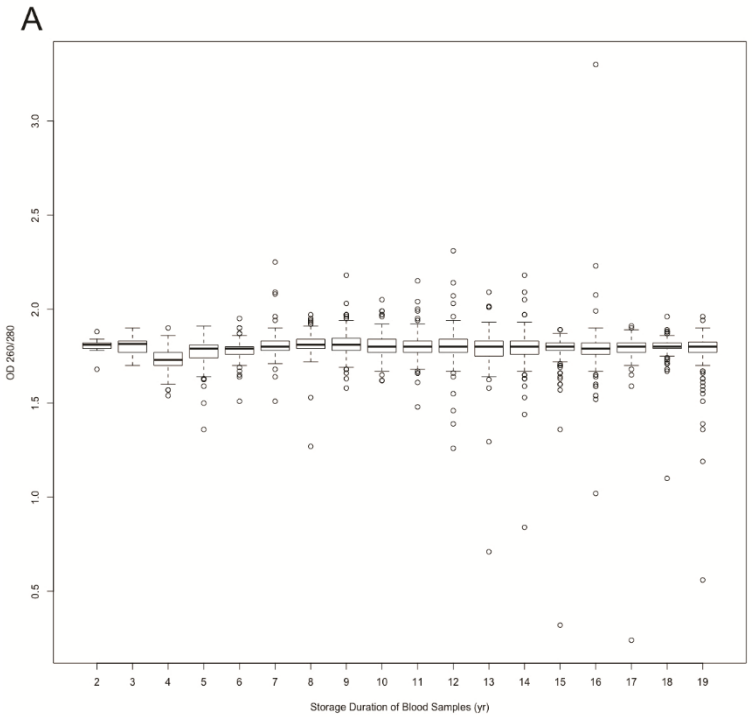


Figure 3.4. A260/280 ratios distribution of all samples across the duration of blood sample storage.
A260/280 ratios were measured by NanoDrop.

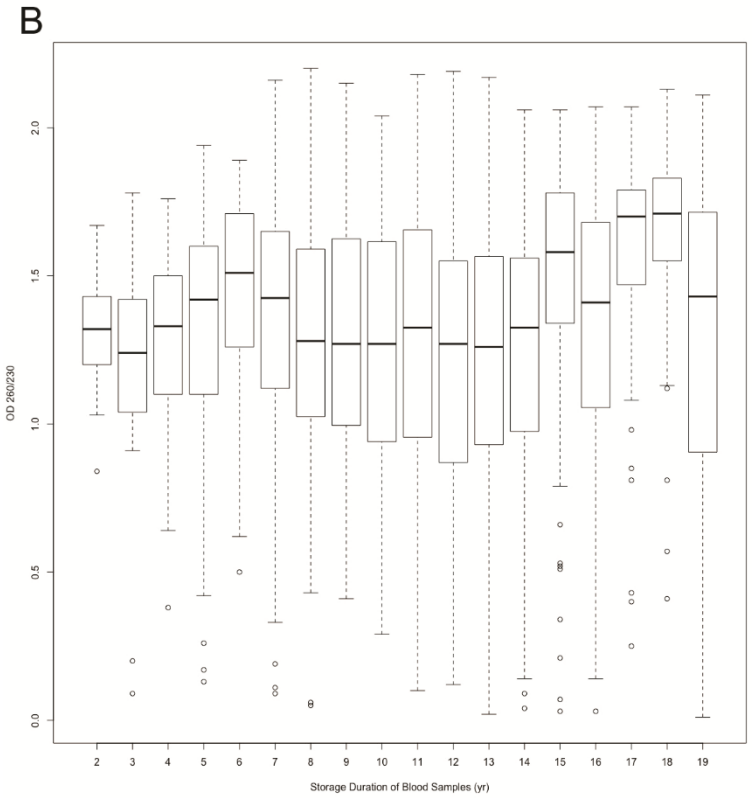


Figure 3.5. A260/230 ratios distribution of all samples across the duration of blood sample storage. A260/230 ratios were measured by NanoDrop.

3.5. Discussion

This study was undertaken to answer questions about the expected quality and quantity of DNA isolated from whole blood samples that have been stored at -30°C for between 2 and 19 years. The concern was that long-term storage of whole blood at -30°C may have a negative effect on the quality and/or quantity of the DNA (AlRokayan, 2000). Genomic DNA was isolated from 2,758 blood samples collected between 1997 and 2012 using the Qiagen FlexiGene® DNA kit. To our knowledge, this is the largest study of its kind assessing the quality and quantity of DNA isolated from frozen blood samples stored for long periods.

The results from our study are unique from previous findings (AlRokayan, 2000; Nederhand et al., 2003) as we used a different DNA extraction method (kit versus salting out or phenol/chloroform), stored samples at -30°C temperature (compared to room temperature and 4°C) and used a much larger sample size. The large sample size of our study affords greater statistical power to evaluate the relationship between DNA yields and the storage duration of the blood samples. A recent publication looking at various

storage conditions on forensic examinations of blood samples and blood stains (n=6) concluded that to prevent DNA degradation during long-term (20 years) storage, they be stored at below -20°C (Hara et al., 2016), such as our samples have been. We have a much larger sample size with varying lengths of storage to assess and have found that the quantity of DNA is not affected and are fairly confident that the same can be said about the quality.

This study found that long-term storage of whole blood samples at -30°C does not have a negative effect on the quality and quantity of the DNA extracted. The average duration of storage was approximately 12 years. The mean A260/280 ratios were found to be approximately 1.8 across all the years of collection, suggesting high purity of the DNA regardless of the duration of blood storage (Figure 3.4). The lower than expected A260/230 ratios indicate the presence of other organic residues (Figure 3.5). The source of the organic residues is likely artefacts resulting from the inadequate removal of organic chemicals used in the extraction chemistry.

The DNA yield was expected to decrease for samples stored for longer periods. Alrokayan reported a 30 to 40% decrease in total DNA yield for whole blood samples frozen at -20°C (AlRokayan, 2000). A decrease in DNA yield was not observed in the JCS samples extracted, as no correlation between storage duration and total DNA yields was observed. Assuming the blood samples are stored within the desired temperature range at stable conditions, DNA can be extracted from these samples after prolonged storage. DNA yields of samples per year of collection clustered tightly together for the 2 to 19 years, except for a few outliers (Figure 3.3).

High-quality dsDNA is preferred for downstream applications such as PCR, microarray analysis or next-generation sequencing analysis. Qubit™ is well known to be a more sensitive method to quantify DNA, as compared to NanoDrop™ (Nakayama et al., 2016; Robin et al., 2016). Qubit™ readings are lower than NanoDrop™ readings as they do not measure single-stranded, degraded DNA (Nakayama et al., 2016; Robin et al., 2016). Our results showed that, on average, 79% of total DNA measured by NanoDrop™ was dsDNA. From the subset of samples that were analysed on Qubit™, the %dsDNA did not decrease with increased years of storage ($P = 0.10$); instead, a positive trend was seen with

increasing storage duration but the effect size was minimal (beta = 0.01, P = 0.045). This supports our hypothesis that the storage duration of the blood samples does not influence DNA quality. It is worth noting that although DNA was extracted from these archived blood samples, no assessments were done for RNA, micro RNA, cell-free circulating DNA or proteins.

The DNA yields are in line with expected yields cited by the manufacturers. With every additional 1 ml of blood volume available for extraction, the DNA yield would be expected to increase by 20.94 µg (P < 0.01). The DNA yield per ml of blood does not increase when more blood was processed. Rather optimization of the DNA extraction protocol used for different volumes of blood affects the DNA yield per ml of blood. Two fixed protocols were used in this study, the protocol for sample volumes of <2 ml of blood yielded ~7 µg (P = 0.0001) more DNA per ml of blood than the protocol used for samples of >2 ml of blood. This is possibly due to the extra time and care given per extraction for the small volumes of these samples because of limited volume. It is possible that the potential yield could be improved by optimizing the extraction protocol to the exact blood volume available. This would however require extra laboratory time and staff.

The patients' age may also influence the number of white cells available (Richardson et al., 2006; Mychaleckyj et al., 2011). In older patients, there is often a decline in the quantity of DNA due to a reduction in the number of leukocytes and lymphocytes with age progression (Erkeller-Yuksel et al., 1992). However, this correlation was not observed in this study. As anticipated, samples from individuals that were HIV positive at the time of sampling had a lower DNA yield as a likely consequence of fewer peripheral blood mononuclear cells due to the infection. However, no difference in mean yield was observed between cancer samples and non-cancer samples.

One limitation of this study was that gel electrophoresis and fluorometry were not done for all 2,758 samples to assess the integrity of the extracted DNA at the beginning of DNA storage. Hence the data lack robustness to definitively conclude that long-term storage of blood samples does not negatively affect DNA integrity. However, from the representative subset of samples (n=200) that were assessed by gel electrophoresis, the authors are confident that the DNA integrity of DNA extracted from long-term stored

blood samples is sound. A comprehensive analysis would include doing a quantitative PCR or a long-range PCR and we have applied for funding to do this. We have used these samples in other genotyping studies and they have produced adequate genotyping call rates (unpublished).

From the results of this large study, the authors are confident that long-term storage (average 11.9 years) of whole blood samples stored in 4 ml EDTA vacutainers at a stable -30°C does not have a negative effect on the quality and quantity of the DNA. The authors are also confident that the DNA integrity assessment done on a representative sample is sufficient to represent the larger cohort. Despite the limitations of this study, we found no evidence that DNA integrity worsened over a prolonged period of blood sample storage. Archived whole blood samples can still be utilized in genetic epidemiology studies provided the blood samples are stored correctly with constant temperature monitoring to prevent freeze-thaw episodes.

3.6. Acknowledgements

The authors wish to acknowledge Sisters Gloria Mokwatle, Patricia Rapoho and Pheladi Kale who carried out the interviews and collected blood specimens from the patients included in this analysis, Mrs Lettie Bester who prepared the specimens for testing and storage, and Mrs Margaret Urban and Mr Tonicah Maphanga for maintaining the JCS database. We thank Dr Mazvita Sengayi for their statistical analysis support. We thank Ms Cassandra Soo and Ms Natalie Smyth for fluorescent DNA quantification. We thank the oncology clinicians and the administration at Charlotte Maxeke Johannesburg Academic Hospital (formerly Johannesburg General Hospital) for assistance and access to patients in their care, and the patients who gave freely of their time. Research reported in this publication and the ongoing extraction of DNA from JCS blood samples is supported by the Grants, Innovation and Product Development (GIPD) Unit of the South African Medical Research Council with funds received from the South African National Department of Health and the UK Medical Research Council from the UK Government's Newton Fund. We thank the DST/NRF internship programme for supporting BN, AS, STD and LJ while interns at CERG. We thank the DAAD/NRF in-country scholarship for a fellowship to WCC, and the Cancer Association of South Africa (CANSA) for research cost

support to CGM. The contents are solely the responsibility of the authors and do not necessarily reflect the views of the funding bodies.

3.7. Author contributions

WCC substantial contributions to conception and design, acquisition of data (DNA extraction), analysis and interpretation of data, drafting of the manuscript and final approval of the version to be published.
RK interpretation of data, revising the manuscript critically for important intellectual content and final approval of the version to be published.
AM analysis and interpretation of data, revising the manuscript critically for important intellectual content, and final approval of the version to be published.
BN, AS, STD and **LJ** acquisition of data (DNA extraction), maintenance of the database and summary report (base of the manuscript), and final approval of the version to be published.
CGM interpretation of data and revising the manuscript critically for important intellectual content and final approval of the version to be published.
CBdV conceived the project, substantial contributions to conception and design, interpretation of data, revising the manuscript critically for important intellectual content and final approval of the version to be published.

3.8. Author Disclosure Statement

No conflicting financial interests exist.

3.9. References

AlRokayan, S.A.H. (2000) Effect of Storage Temperature on the Quality and Quantity of DNA Extracted from Blood. Pakistan Journal of Biological Sciences. 3(3):392–394. DOI: 10.3923/pjbs.2000.392.394.

Desjardins, P. & Conklin, D. (2010) NanoDrop microvolume quantitation of nucleic acids. *Journal of visualized experiments :* JoVE. (45). DOI: 10.3791/2565.

Di Pietro, F., Ortenzi, F., Tilio, M., et al. (2011) Genomic DNA extraction from whole blood stored from 15- to 30-years at -20 °C by rapid phenol-chloroform protocol: a useful tool for genetic epidemiology studies. Molecular and cellular probes. 25(1):44–8. DOI: 10.1016/j.mcp.2010.10.003.

Erkeller-Yuksel, F.M., Deneys, V., Yuksel, B., et al. (1992) Age-related changes in human blood lymphocyte subpopulations. The Journal of pediatrics. 120(2 Pt 1):216–22. DOI: 10.1016/s0022-3476(05)80430-5.

Griffiths, L. & Chacon-Cortes, D. (2014) Methods for extracting genomic DNA from whole blood samples: current perspectives. Journal of Biorepository Science for Applied Medicine. (May):1. DOI: 10.2147/BSAM.S46573.

Hara, M., Nakanishi, H., Yoneyama, K., et al. (2016) Effects of storage conditions on forensic examinations of blood samples and bloodstains stored for 20 years. Legal medicine (Tokyo, Japan). 18:81–4. DOI: 10.1016/j.legalmed.2016.01.003.

Holland, N.T., Smith, M.T., Eskenazi, B. et al. (2003) Biological sample collection and processing for molecular epidemiological studies. Mutation research. 543(3):217–34. DOI: 10.1016/s1383-5742(02)00090-x.

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

Mychaleckyj, J.C., Farber, E.A., Chmielewski, J., et al. (2011) Buffy coat specimens remain viable as a DNA source for highly multiplexed genome-wide genetic tests after long term storage. Journal of translational medicine. 9:91. DOI: 10.1186/1479-5876-9-91.

Nakayama, Y., Yamaguchi, H., Einaga, N., et al. (2016) Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. PloS one. 11(3):e0150528. DOI: 10.1371/journal.pone.0150528.

Nederhand, R.J., Droog, S., Kluft, C., et al. (2003) Logistics and quality control for DNA sampling in large multicenter studies. Journal of thrombosis and haemostasis : JTH. 1(5):987–91. DOI: 10.1046/j.1538-7836.2003.00216.x.

Qiagen®. (2010) FlexiGene® DNA Handbook. Available: http://www.qiagen.com/.

Richardson, A.J., Narendran, N., Guymer, R.H., et al. (2006) Blood storage at 4 degrees C-factors involved in DNA yield and quality. *The Journal of laboratory and clinical medicine*. 147(6):290–4. DOI: 10.1016/j.lab.2006.01.005.

Robin, J.D., Ludlow, A.T., LaRanger, R., et al. (2016) Comparison of DNA Quantification Methods for Next Generation Sequencing. *Scientific Reports*. 6(1):24067. DOI: 10.1038/srep24067.

Thermo Scientific™. (2011) NanoDrop: Assessment of Nucleic Acid Purity. Protoc Prod Manuals. 1–2.

Steinberg, K.K., Sanderlin, K.C., Ou, C.Y., et al. (1997) DNA banking in epidemiologic studies. Epidemiologic reviews. 19(1):156–62. DOI: 10.1093/oxfordjournals.epirev.a017938.

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):e1001182. DOI: 10.1371/journal.pmed.1001182.

4. CHAPTER 4: Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population.

The aim of this chapter was to determine whether genes and loci previously shown to be associated with OSCC in non-African populations were also associated with OSCC in the South African Black population. Seven risk loci from a joint analysis of three GWAS from Chinese populations (Wu et al, 2014) were prioritized for testing in two study populations from South Africa. Bye et al (2011, 2012) genotyped top SNPs from the risk loci in OSCC cases and controls from the Western Cape. Chen et al (2019) genotyped the same SNPs, using the Agena MassARRAY iPLEX genotyping assay, in OSCC cases and controls from the JCS, thereafter, performed a meta-analysis of the genotyped SNPs common to both studies to increase the power of the combined study to identify robust associations. The findings from the chapter were published in Carcinogenesis. gDNA isolated from the JCS (in Chapter 2) were of good quality and quantity for use in various genetic assays. The published manuscript can be found in Appendix C.

Chen, W.C., Bye, H., Matejcic, M., Amar, A., Govender, D., Khew, Y.W., Beynon, V., Kerr, R., Singh, E., Prescott, N.J., Lewis, C.M., Babb de Villiers, C., Parker, M.I., Mathew, C.G. (2019) Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. Carcinogenesis. 40(4):513–520. DOI: 10.1093/carcin/bgz026.

The student performed the data cleaning, and data analysis and wrote and reviewed the manuscript.

4.1. Abstract

Oesophageal squamous cell carcinoma (OSCC) has a high incidence in southern Africa and a poor prognosis. Limited information is available on the contribution of genetic variants in susceptibility to OSCC in this region. However recent genome-wide association studies have identified multiple susceptibility loci in Asian and European populations. In this study, we investigated genetic variants from 7 OSCC risk loci identified in non-African populations for association with OSCC in the South African Black population. We performed association studies in a total of 1,471 cases and 1,791 controls from two study

sample groups, which included 591 cases and 852 controls from the Western Cape and 880 cases and 939 controls from the Johannesburg region in the Gauteng province. Thereafter we performed a meta-analysis for 11 variants that had been genotyped in both studies. A single nucleotide polymorphism in the *CHEK2* gene, rs1033667, was significantly associated with OSCC (P = 0.002; OR = 1.176; 95% CI: 1.06 - 1.30). However, single nucleotide polymorphisms in the *CASP8/ALS2CR12*, *TMEM173*, *PLCE1*, *ALDH2*, *ATP1B2/TP53* and *RUNX1* loci were not associated with the disease (P > 0.05). The lack of association of six of these loci with OSCC in South African populations may reflect different genetic risk factors in non-African and African populations or differences in the genetic architecture of African genomes. The association at *CHEK2*, a gene with key roles in cell cycle regulation and DNA repair, in an African population provides further support for the contribution of common genetic variants at this locus to the risk of oesophageal cancer.

4.2. Introduction

Oesophageal cancer (OC) is the eighth most common cancer worldwide (Ferlay et al., 2015), with oesophageal squamous cell carcinoma (OSCC) accounting for nearly 90% of all cases of OC worldwide (Arnold et al., 2015). Particularly high incidence rates are observed in regions such as southern Africa, China and Japan. In South Africa, OC was the third most common cancer in males of African ancestry and the fourth most common cancer in females of African ancestry in 2014, with age-standardized incidence rates of 4.79 and 3.16 per 100,000, respectively (NCR, 2018). In addition to the recognized risk factors for OSCC of alcohol consumption and tobacco use, polycyclic aromatic hydrocarbons exposure, high food temperatures and other factors such as diet and consumption of Fusarium-contaminated maize may also be involved in the development of the disease (reviewed in Matejcic & Iqbal Parker, 2015; Matejcic, Gunter & Ferrari, 2017; Murphy, McCormack et al., 2017; Abnet, Arnold & Wei, 2018).

Over the past 9 years, several genome-wide association studies (GWAS) have been performed in non-African populations to determine the genetic risk factors involved in the development of OSCC. In 2009, a GWAS for OSCC in the Japanese population found that SNPs at *ADH1B* (rs1229984) and *ALDH2* (rs671) were associated with the disease (Cui et al., 2009), and a European GWAS of upper aerodigestive tract cancers detected

genome-wide significant association with OSCC at *ALDH2* (rs4767364) (McKay et al., 2011). Since 2010, four OSCC GWAS and a meta-analysis have been performed in the Chinese population, which identified multiple additional risk loci (Abnet et al., 2010; Wang et al., 2010; Wu et al., 2011, 2012, 2014). These included SNPs at *PLCE1*, *HEATR3*, *HAP1*, *CHEK2/XBP1*, *ST6GAL1*, *SMG6*, *PTPN2*, *SLC52A3 (C20orf54)*, *PDE4D*, *RUNX1*, *UNC5CL* and *CYP26B1*. However, in a subsequent analysis the associations at *HAP1*, *SMG6*, *SLC52A3 (C20orf54)* and *PTPN2* were not confirmed (Wu et al., 2014).

We previously tested several of these variants for association with OSCC in the South African Black (SAB) population recruited in the Western Cape region (Bye et al., 2011, 2012). Coding SNPs in *ADH1B* and *ALDH2* were absent in the SAB population (Bye et al., 2011), and none of the lead SNPs from the Chinese GWAS studies was associated with OSCC (*PLCE1* rs2274223, *SLC52A3/C20orf54* rs13042395, *PDE4D* rs10052657, *RUNX1* rs2014300, *UNC5CL* rs10484761). However, further analysis of the *PLCE1* locus identified a significant association with the SNP Arg548Leu (rs17417407).

The Johannesburg Cancer Study (JCS) was established in 1995, with the aim of recruiting Black, treatment-naïve, patients diagnosed with cancer from the greater Johannesburg area in Gauteng province (Urban et al., 2012). The JCS recruited all cancer cases, including OSCC, as well as non-cancer, ethnically matched controls from three tertiary academic hospitals in the greater Johannesburg region. The University of Cape Town cancer study (UCT) was established to recruit OSCC patients from the Western Cape region of South Africa through tertiary hospitals (Matejcic et al., 2011; Bye et al., 2012). Here we investigate the association of 7 main-effect susceptibility loci (*CASP8/ALS2CR12, TMEM173, PLCE1, ALDH2, ATP1B2/TP53, RUNX1, CHEK2/XBP1*) identified by GWAS (Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012) with OSCC in SAB populations recruited in these two regions of South Africa.

4.3. Materials and Methods

Study subjects

The UCT study sample consisted of 591 SAB OSCC cases and 852 ethnically matched non-cancer controls that were available for genotyping. The cases and controls were mainly

Xhosa-speakers (97% and 95% respectively) from the Western or Eastern Cape provinces of South Africa. The JCS samples consisted of 880 SAB OSCC cases and 939 ethnically matched non-cancer controls available for genotyping. The JCS cases and controls were recruited from the greater Johannesburg area in Gauteng province, with self-reported African ancestry. The major linguistic groups in the JCS in both cases and controls were the Nguni (Zulu and Xhosa speakers, 37.8%) and Sotho-Tswana (Southern Sotho, Tswana and Northern Sotho speakers, 39.2%). Control individuals for both study samples had no history of cancer, lived in the same residential areas, and had a similar socioeconomic status to the cases. The JCS controls were recruited mainly from the cardiovascular unit at the Charlotte Maxeke Academic Hospital in Johannesburg, a high proportion of whom consumed alcohol.

All patients had a histologically confirmed primary invasive OSCC and were recruited between 1995 and 2016. Smoking status was subdivided into ever-smokers (those who had smoked at some point in their lives) or never-smokers. Drinkers were defined as subjects who consumed alcohol at least once every week or non-drinkers. Whole blood samples were collected, with informed consent, from all subjects and DNA was extracted at the University of Cape Town and the University of the Witwatersrand as previously described (Bye et al., 2011; Chen et al., 2018). Ethical approval for the study was obtained from the joint University of Cape Town/Groote Schuur Hospital Research Ethics Committee and the University of Stellenbosch/Tygerberg Hospital Ethics Committee (UCT HREC 040/2005), and the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (M140271, M160807).

SNP selection and genotyping

For the UCT samples, 12 SNPs were selected from the 7 OSCC susceptibility loci described above and genotyped. The list of these SNPs and loci is shown in Table 4.2, the majority of which were the lead SNPs from published GWAS (Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014). The index SNP for the *TMEM173* locus on chromosome 5q31.2, rs7447927 (Wu et al., 2014) failed TaqMan assay design; a proxy for this SNP, rs13153461, was genotyped which is in strong linkage disequilibrium (LD) with rs7447927 in both Chinese ($r^2 = 0.98$) and African ($r^2 = 0.69$ in Yoruban) populations in the 1000 Genome Project Phase 3. The functional SNP in the *TP53* gene (rs1800371, p.P47S),

which has only been detected in African populations (Jennis et al., 2016), was also genotyped. For the JCS samples, the lead SNPs from the 7 OSCC susceptibility loci and one or more tagging SNPs in LD with the lead SNP were genotyped to provide additional coverage in case of genotype assay failure. Tag SNP selection was based on allele frequencies and LD in the Yoruban (YRI) and Chinese (CHB) populations from the 1000 Genomes Consortium (Auton et al., 2015), using LDLink 3.0 (Machiela & Chanock, 2015), which resulted in the selection of 18 SNPs for genotyping (Table 4.2). An overview of the study design is shown in Figure 4.1.

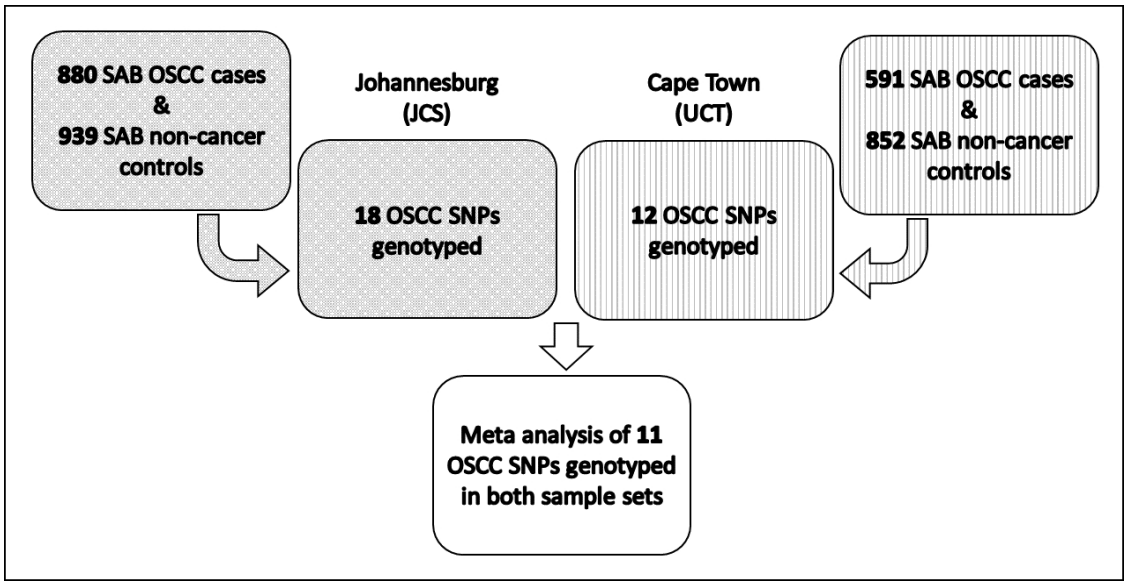


Figure 4.1. Schematic of the overall study design.

The Agena MassARRAY iPLEX genotyping assay (Agena Bioscience) was used to genotype 18 SNPs in the JCS samples. The assay was performed using the iPLEX assay protocol (Gabriel, Ziaugra & Tabbaa, 2009). Oligonucleotide primers were designed using the Agena Assay Design Suite (Agena Bioscience). Reactions for the PCR assays were carried out in 5.0 µl volumes in 96-well plates. Each reaction contained 25 ng DNA, 0.5 µl 10X PCR Buffer, 2 mM $MgCl_2$, 500µM dNTP mix, 100 nM Primer mix, 1.0 U Taq polymerase and distilled water (dH_20) made up the remaining volume. Reactions for the shrimp alkaline phosphatase (SAP) post-PCR clean-up were carried out by adding 2.0 µl SAP mix, each containing 0.17 µl 10X SAP buffer, 0.51 U SAP enzyme and dH_20 to each reaction well. Reactions for the iPLEX single-base extension were carried out by adding 4.0 µl iPLEX mix, each containing 0.2 µl 10X iPLEX Buffer Plus, 0.2 µl iPLEX termination mix, 8 µM primer mix, 0.041 µl iPLEX enzyme and dH_20 to each reaction well. Primer extension products

were spotted onto the SpectroCHIPs and detection of the primer extension products by mass spectrometry was done on the Agena MassARRAY Compact mass spectrometer. All but one of 18 SNPs (rs1642764) produced discrete genotype clusters in the iPLEX assay.

The 12 SNPs for the UCT samples were genotyped using TaqMan SNP assays (Life Technologies). A TaqMan SNP assay was also used to genotype the SNP (rs1642764) in the JCS samples that had failed in the iPLEX assay. Reactions for the TaqMan SNP assays were carried out in 2.5 µl volumes in 96-well plates. Each reaction contained 20ng DNA, 1.25 µl ABsolute QPCR ROX mix (Thermo Scientific) and 0.03 µl 40X TaqMan SNP assay mix (Life Technologies), with the PCR performed on a PTC-0225 DNA Engine (MJ Research) according to manufacturer's protocol. Fluorescent levels at the PCR endpoint were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems), with additional manual checks for discrete genotype clusters. High levels of concordance of genotyping results between MassARRAY chemistry and TaqMan chemistry have been described previously, including for clinical implementation (Syrmis et al., 2011; Goh et al., 2017).

Pearson's chi-squared (χ^2) test was used to determine deviations from the Hardy–Weinberg equilibrium in controls only, using a cut-off of $P < 0.001$. Call rates for all other SNPs genotyped by either method were > 95%.

Statistical analysis

In order to test for association with OSCC in the UCT and JCS study samples, allele frequencies in cases and controls were compared using Pearson's chi-squared (χ^2) test using Plink (version 1.9, https://www.cog-genomics.org/plink/1.9/). Allelic odds ratios (OR) and 95% confidence intervals (CI) were calculated using the common allele as the reference. In view of the differences in the ethnolinguistic composition of the subjects in the UCT and JCS samples and the different genotyping methodology used, a meta-analysis of the two data sets for fixed-effects and random-effects was performed using Plink (version 1.9, https://www.cog-genomics.org/plink/1.9/). A Bonferroni-corrected P value of < 0.0045 (0.05/11) was used as a significance threshold to account for the testing of 11 variants in the meta-analysis. The proportion of SNPs showing the same direction of effect in the meta-analysis as compared to the direction of effect observed in non-African

studies was tested using the exact binomial test for a difference from 0.5. As secondary analyses, dominant and recessive models were also tested for association with each SNP.

The power of the association tests was determined using Quanto (v1.2.4, http://hydra.usc.edu/gxe/). The average effect size of the loci tested was 1.26 and a range of minor allele frequency (MAF) for the SNPs was used for the power calculation. The power of the UCT study sample to detect an effect size of 1.26 for MAFs of 0.1 to 0.5 at alpha 0.05 was 48% to 86%. The power of the JCS samples to detect an effect size of 1.26 for MAFs of 0.05 to 0.45 at alpha 0.05 was 36% to 94%. The power of the combined study samples (1,471 cases and 1,791 controls) in the meta-analysis to detect an effect size of 1.26 was calculated for the range of MAFs observed in the combined study sample (0.17 – 0.46) and was 78% – 96% at alpha 0.0045.

Gene x environmental (GxE) interactions with tobacco smoking were investigated for SNPs with significant evidence of allelic association ($P < 0.0045$) by testing for association in cases and controls in the combined UCT/JCS study samples stratified by smoking status (ever vs never) in a χ^2 analysis using Stata MP 15.1 (StataCorp, USA). GxE interactions for alcohol consumption were investigated in a case-only analysis of the combined UCT/JCS study samples. Case-control interactions were not tested owing to the high level of alcohol use in the JCS controls.

Bioinformatic analysis of associated variants to look for potential functional effects was carried out using established tools and resources such as HaploReg (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php), GTEx (https://gtexportal.org), RegulomeDB (http://www.regulomedb.org), ENCODE (https://www.encodeproject.org) and FuncPred (https://snpinfo.niehs.nih.gov/snpinfo/snpfunc.html).

4.4. Results

The characteristics of the cases and controls in the JCS and UCT sample sets are shown in Table 4.1.

Table 4.1. Characteristics of OSCC cases and controls in the SAB populations.

Study Samples		JCS		UCT	
		Case	Control	Case	Control
Total Number		880	939	591	852
Age, mean years (SD)		58.2 (10.2)	50.0 (15.5)	60.2(11.3)	48.9 (16.8)
Sex, n (%)	Male	545 (61.9)	240 (25.6)	284 (48.1)	342 (40.1)
	Female	332 (37.7)	698 (74.3)	307 (51.9)	507 (59.5)
	Unknown	3 (0.3)	1 (0.1)	0 (0)	3 (0.4)
Smoking status, n (%)	Ever-smoker	598 (68.0)	333 (35.5)	364 (61.6)	338 (39.7)
	Never-smoker	276 (31.4)	584 (62.2)	227 (38.4)	504 (59.1)
	Unknown	6 (0.7)	22 (2.3)	0 (0)	10 (1.2)
Alcohol consumption, n (%)	Drinker	473 (53.8)	633 (67.4)[#]	370 (62.6)	458 (53.7)
	Non-drinker	404 (45.9)	305 (32.5)[#]	219 (37.1)	390 (45.8)
	Unknown	3 (0.3)	1 (0.1)[#]	2 (0.3)	4 (0.5)

[#]Ascertainment bias of individuals with higher baseline alcohol consumption levels.

The mean age of diagnosis was similar in the JCS and UCT groups (58.2 and 60.2 years respectively). The male-to-female ratio was 1.64 in the JCS cases but 0.93 in the UCT sample. Smoking rates were high (> 60%) in both groups of cases. The proportion of cases reporting alcohol consumption was higher in the UCT group.

Three of the 12 variants tested, under the allelic additive model, in the SAB UCT study sample were significantly associated with OSCC after Bonferroni multiple testing correction ($P < 0.0042$) (Table 4.2). These were *XBP1* rs2239815, *CHEK2* rs4822983 and *CHEK2* rs1033667. The other 9 SNPs in *CASP8*, *ALS2CR12*, *TMEM173*, *PLCE1*, *ATP1B2/TP53*, *TP53* and *RUNX1* were not significantly associated with the disease, although rs17417407 at *PLCE1* showed nominal evidence of association. In the secondary analysis of the SAB UCT study sample, results under a dominant and a recessive model of inheritance were similar to the additive model (Supplementary Table 4.1), and all SNPs showing association with OSSC in the dominant or recessive models were also associated in the additive model. None of the 18 variants tested in the SAB JCS study sample showed evidence of association with OSCC ($P > 0.05$) under the allelic additive model (Table 4.2), or the dominant and recessive models (Supplementary Table 4.1).

Table 4.2. Case-control association study analysis for SNPs genotyped in 7 known risk loci for OSCC in the SAB populations.

				UCT				JCS			
Chr	**Gene**	**SNP**	**Alleles**	**MAF Case**	**MAF Control**	**P-value**	**OR (95% CI)**	**MAF Case**	**MAF Control**	**P-value**	**OR (95% CI)**
2	*CASP8*	rs10931936	C/T	0.19	0.20	0.391	0.90 (0.71-1.15)	0.22	0.20	0.271	1.09 (0.93-1.29)
2	*ALS2CR12*	rs13016963#	G/A	0.35	0.35	0.994	1.00 (0.82-1.22)	0.39	0.38	0.515	1.05 (0.91-1.20)
2	*ALS2CR12*	rs10201587#	A/G	-	-	-	-	0.38	0.39	0.783	0.98 (0.86-1.12)
5	*TMEM173*	rs13181561	A/G	-	-	-	-	0.48	0.49	0.275	0.93 (0.82-1.06)
5	*TMEM173*	rs13153461	G/A	0.04	0.05	0.357	0.83 (0.56-1.23)	-	-	-	-
10	*PLCE1*	rs17417407	G/T	0.17	0.21	0.014*	0.76 (0.60-0.95)	0.19	0.19	0.881	0.99 (0.84-1.17)
10	*PLCE1*	rs7084339	G/A	-	-	-	-	0.48	0.46	0.178	1.09 (0.96-1.25)
10	*PLCE1*	rs3765524#	T/C	0.47	0.47	0.854	1.02 (0.83-1.26)	0.48	0.46	0.186	1.09 (0.96-1.24)
10	*PLCE1*	rs2274223#	A/G	0.42	0.40	0.508	1.06 (0.89-1.26)	0.41	0.43	0.334	0.94 (0.82-1.07)
10	*PLCE1*	rs11187850	A/G	-	-	-	-	0.21	0.19	0.204	1.11 (0.94-1.31)
12	*ALDH2*	rs4767364#	A/G	-	-	-	-	0.12	0.11	0.138	1.17 (0.95-1.44)
17	*ATP1B2/TP53*	rs1642764#	C/T	0.21	0.20	0.527	1.07 (0.86-1.33)	0.18	0.18	0.982	1.00 (0.85-1.19)
17	*ATP1B2/TP53*	rs1641511	A/G	-	-	-	-	0.39	0.42	0.081	0.89 (0.78-1.02)
17	*TP53*	rs1800371#	G/A	0.02	0.03	0.145	0.67 (0.38-1.16)	0.03	0.02	0.117	1.38 (0.92-2.07)
21	*RUNX1*	rs2014300#	A/G	0.38	0.40	0.376	0.92 (0.77-1.10)	0.36	0.36	0.795	1.02 (0.89-1.17)
21	*RUNX1*	rs2834718	T/A	-	-	-	-	0.33	0.33	0.667	0.97 (0.84-1.11)
22	*CHEK2*	rs4822983#	C/T	0.46	0.39	0.001***	1.32 (1.12-1.56)	0.43	0.42	0.836	1.01 (0.89-1.16)
22	*CHEK2*	rs1033667#	C/T	0.44	0.38	0.002**	1.30 (1.10-1.53)	0.42	0.39	0.145	1.10 (0.97-1.26)
22	*XBP1*	rs2239815#	C/T	0.21	0.16	0.001***	1.41 (1.15-1.74)	0.16	0.18	0.162	0.88 (0.74-1.05)

MAF – Minor allele frequency; C/T – Major allele/Minor allele; OR – Odds ratio, for each minor allele carried; 95% CI – 95% confidence interval; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001. # - Previously published lead SNPs (9 – 14); other SNPs tag the lead SNPs.

A fixed-effect meta-analysis was performed for the 11 SNPs genotyped in both study samples under the allelic additive model (Table 4.3). One SNP, *CHEK2* rs1033667 (P = 0.002, OR = 1.18 (1.06 – 1.30)), was significantly associated with OSCC under this model. This was also the only SNP with evidence of association when analyzed under dominant or recessive models (Supplementary tables 4.2 and 4.3). A random-effect meta-analysis was also performed to investigate potential heterogeneity between the two studies (Table 4.3). The only marker with any evidence of association with OSCC under the random-effect model was rs1033667 (P(R) = 0.032), and the direction of the effect was

the same in both studies. Three variants showed evidence of heterogeneity (*TP53* rs1800371, *CHEK2* rs4822983, and *XBP1* rs2239815), but none of these was associated with OSCC under the random-effect model. Of the eleven variants included in the meta-analysis, 9 were genotyped in the Chinese genome-wide scans and 5 had effects in the same direction as in the original Chinese studies (P = 0.556).

Table 4.3. Meta-analysis of 11 variants genotyped in the SAB populations.

Chr	Gene	SNP	Major Allele	Minor Allele	P	P(R)	OR (95% CI)	Q	I^2
2	*CASP8*	rs10931936	C	T	0.663	0.886	1.030 (0.902-1.177)	0.186	42.91
2	*ALS2CR12*	rs13016963	G	A	0.588	0.588	1.031 (0.923-1.152)	0.726	0.00
10	*PLCE1*	rs17417407	G	T	0.116	0.310	0.898 (0.786-1.027)	0.061	71.57
10	*PLCE1*	rs3765524	T	C	0.221	0.221	1.072 (0.959-1.197)	0.588	0.00
10	*PLCE1*	rs2274223	A	G	0.708	0.780	0.980 (0.882-1.089)	0.267	18.70
17	*ATP1B2/TP53*	rs1642764	C	T	0.682	0.682	1.028 (0.899-1.176)	0.627	0.00
17	*TP53*	rs1800371	G	A	0.685	0.962	1.070 (0.772-1.483)	0.037*	77.10
21	*RUNX1*	rs2014300	A	G	0.740	0.740	0.982 (0.882-1.094)	0.390	0.00
22	*CHEK2*	rs4822983	C	T	0.028*	0.285	1.123 (1.013-1.245)	0.015*	83.03
22	*CHEK2*	rs1033667	C	T	0.002**	0.032*	1.176 (1.060-1.304)	0.137	54.85
22	*XBP1*	rs2239815	C	T	0.313	0.648	1.071 (0.938-1.758)	0.001***	91.30

OR – Odds ratio, for each minor allele carried; 95% CI – 95% confidence interval; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001. P – P-value for the fixed-effect meta-analysis. P(R) – P values for the random effect meta-analysis. Q – Cochrane's Q statistic P values. I^2 – I^2 heterogeneity index.

A comparison of allele frequencies of the 19 SNPs of interest in the UCT and JCS controls with their allele frequencies in the 1000 Genomes project phase 3 populations, Yoruban (YRI), Han Chinese (CHB) and Northern European (CEU), is shown in Figure 4.2 and Supplementary Table 4.4. The majority of the SNPs were common in all these populations, but allele frequencies were very different for SNPs such as *TMEM173* rs13153461, which had low frequencies in African populations. In such cases, there would be low power to detect an association of this marker with OSCC in Africans. Also, for SNPs such as *ALDH2* rs4767364 and XBP1 rs2239815, the rarer allele in one population was the common allele in another. Notably, the SNP with the strongest evidence of association with OSCC, *CHEK2* rs1033667, had similar frequencies in all these populations.

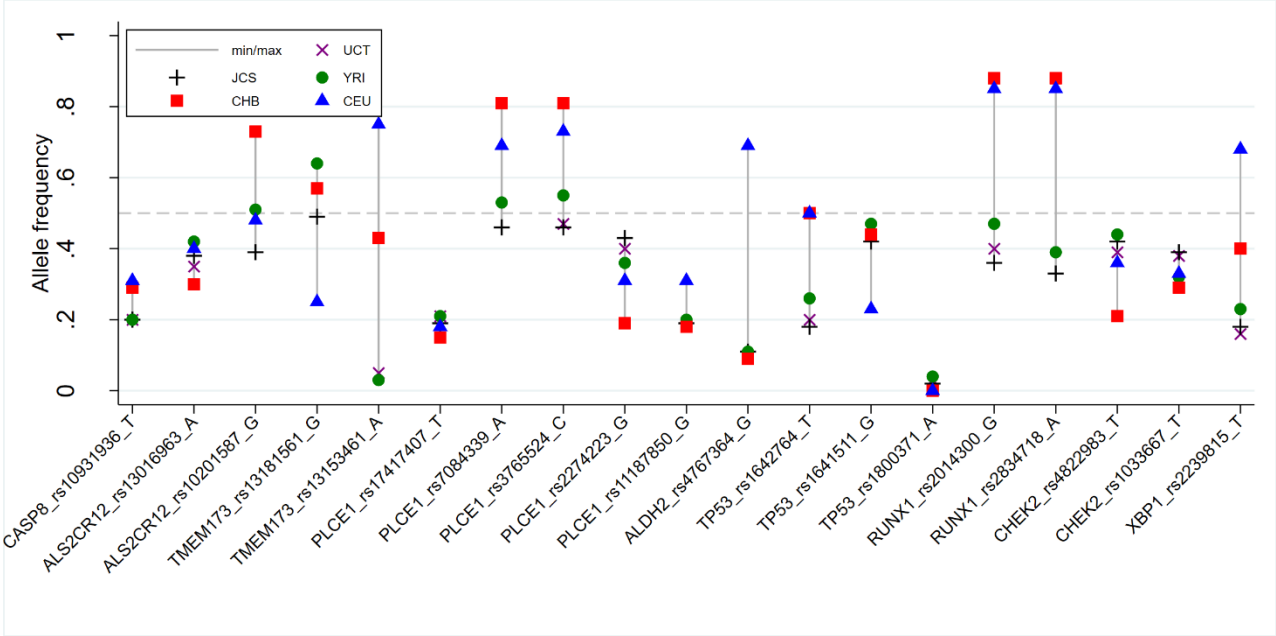


Figure 4.2. Comparison of the allele frequencies of the 19 SNPs of interest by population.
JCS - Johannesburg cancer study, CHB - Han Chinese, UCT - University of Cape Town, YRI - Yorabun, CEU - Northern European.

The rs1033667 variant was investigated by bioinformatic analysis for evidence of a functional effect. The SNP is located in an intron of *CHEK2*, 90 bp downstream of a donor splice site. Analysis of this variant using a variety of predictive tools (see Methods) did not provide any direct evidence for a functional effect. However, expression quantitative trait locus (eQTL) data show that it is associated with altered expression of several genes in this region in multiple tissue types, including increased expression of *CHEK2* in liver and *HSCB* in oesophageal mucosa, and reduced expression of *TTC28* in transformed fibroblasts (Westra et al., 2013; GTEx Consortium, 2015). The only SNP is currently known to be in strong linkage disequilibrium ($r^2 > 0.8$) with rs1033667 in African populations is rs2078555, which is also an intronic SNP in *CHEK2*.

Bioinformatic analysis of this variant using the same tools also failed to reveal any direct evidence for a functional effect and showed similar eQTLs to rs1033667 for *CHEK2*, *HSCB* and *TTC28* both in terms of the direction of effect and tissue type.

The *CHEK2* variant rs1033667 was tested for gene x environmental interactions with tobacco smoking in a case-control analysis. Tobacco smoking was an independent risk factor for OSCC, OR = 3.04 (2.36 - 3.90) for ever-smokers ($P < 0.001$) compared to non-smokers. No significant interaction with rs1033667 was observed for tobacco smoking (P

= 0.60). A case-only analysis was carried out for interaction of rs1033667 with alcohol use (see Methods), but no significant interaction was observed (P = 0.13). Although the *ALDH2* SNP rs4767364 was not associated with OSCC in our study, this variant was tested for interaction with alcohol because of strong evidence for interaction between variants in this gene and alcohol use in other populations (Cui et al., 2009; Wu et al., 2012), but no interaction was detected (P = 0.30).

4.5. Discussion

This study investigated 7 main-effect susceptibility loci that were previously reported to be associated with OSCC in non-African populations (Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014), for association with this disease in the SAB population. The meta-analysis included 1,471 cases and 1,791 controls from the Western Cape and Gauteng provinces, which is more than double the sample size of any previous African study of oesophageal cancer genetics. Only one SNP (*CHEK2* rs1033667) was significantly associated with OSCC in the SAB population in the meta-analysis of the two study samples (P = 0.002, OR = 1.18 (1.06 – 1.30)). This finding suggests that shared aetiology of the disease may exist between the Chinese populations and the SAB populations involving the *CHEK2* gene. None of the other loci tested showed evidence of transference to the SAB population, even though this study was well powered to detect these associations except for two low-frequency SNPs.

CHK2 (encoded by *CHEK2*) is a serine-threonine kinase with a key role in the DNA damage response pathway. Upon activation by genotoxic stress, it can phosphorylate downstream proteins to lead to the activation of DNA repair, cell-cycle arrest, senescence or apoptosis (Zannini, Delia & Buscemi, 2014). Rare germline mutations in *CHEK2* have been implicated in susceptibility to multiple cancers, including strong evidence of association with breast and prostate cancer (Southey et al., 2016). The *CHEK2* gene lies within a 100kb region of chromosome 22 which also contains other genes of interest in relation to cancer, *XBP1*, *HSCB* and *TTC28*. *XPB1* encodes an X-box binding protein 1 which is involved in the unfolded protein response in the endoplasmic reticulum. The stress caused by the accumulation of unfolded proteins activates this response which leads to the restoration of normal protein folding, or in severe cases, apoptosis (Hetz, Chevet &

Harding, 2013). *XBP1* has been shown to have an important role in the development and progression of triple-negative breast cancer through its control of the HIF1alpha pathway (Chen et al., 2014). *HSCB* encodes a member of the heat shock cognate B (HscB) family of proteins involved in the synthesis of iron-sulphur clusters and redox reactions of mitochondrial electron transport. Expression of HSCB is elevated in breast cancer tissue, and downregulation of this gene has recently been shown to reduce cell proliferation (Lee et al., 2018). *TTC28* encodes a tetratricopeptide repeat-containing protein which may be involved in spindle formation in mitosis. Whole-genome sequencing of OSCC tumour tissue has revealed frequent structural alterations in *TTC28*, the significance of which is not yet clear (Chang et al., 2017). The *CHEK2* SNP rs1033667 is associated with altered expression of CHEK2 itself, and also of *HSCB* and *TTC28*. A detailed fine map of this region in large sample sizes will be required to address the genes and variants which are driving this association in more detail.

The variants from the other 6 loci of interest were not significantly associated with OSCC in either of our sample sets or in the meta-analysis. These included a low-frequency functional SNP in *TP53*, Pro47Ser, which is unique to African populations (Jennis et al., 2016). However, the SNP has an allele frequency of only 2-3% so it would require a very large sample size to resolve a possible association with this or any other cancers in African populations. Interestingly, this variant has recently been reported to be associated with pre-menopausal breast cancer in African-American women (Murphy, Liu, et al., 2017). In our previous study of the SAB population of the Western Cape, we reported evidence of a possible association of a coding SNP in *PLCE1*, rs17417407, with OSCC (Bye et al., 2012). However, this was not replicated in the JCS sample and the association was not significant in the meta-analysis. It is possible that the larger sample size is a more robust reflection of the effect size for this variant, or there may be inter-African population differences that will only be resolved by the analysis of very large sample sizes in these populations (Gurdasani et al., 2015). As discussed previously there are several possible reasons for a lack of association in our studies (Bye et al., 2011, 2012). Firstly, there may be insufficient power to detect the effect sizes observed in the original GWAS studies. However, we had 78 - 96% power to detect an effect (OR) of 1.26 for the GWAS-derived SNPs at alpha 0.0046 in the meta-analysis, so detection of an association signal should have been possible at most of these loci.

Another possibility is that the actual causal variants in the Chinese population arose after the migration of humans out of Africa and are therefore not present in African populations. A recent exome-wide association study of OSCC reported 6 new susceptibility loci in the Chinese population (Chang et al., 2018). However, an inspection of the 1000 Genome project phase 3 data for West and East African populations showed that two of the three low-frequency variants (rs117353193 and rs17848945) are monomorphic and one (rs138478634) has a MAF of 0.001, thus these 3 loci are unlikely to make a significant contribution to OSCC risk in African populations. The 3 common associated variants from chromosome 6p21.3 in the exome study will be tested as part of an ongoing GWAS in the SAB population.

An important challenge in the investigation of transference of genetic factors between African and non-African populations is that linkage disequilibrium (LD) is generally lower across African genomes as compared to Asian and European genomes, with shorter haplotype blocks (Teo, Small & Kwiatkowski, 2010). Thus, the genotyped SNPs may be in high LD with the causal variant in the Chinese population but in low LD in the SAB populations, and hence the genotyped SNP would not 'tag' the causal SNP in an African population well and a disease association might not be observed. Finally, it is possible that differences in population structure between cases and controls could contribute to a weakening of genuine differences in allele frequencies of SNPs between cases and controls (Teo, Small & Kwiatkowski, 2010). This seems unlikely in the UCT sample since cases and controls were predominantly from the same ethnolinguistic group, but the JCS study sample included a broader ethnic mix of South African Black participants. In general, one would expect differences in the population structure of cases and controls to increase false-positive associations rather than reducing genuine disease associations, but this could only be resolved by high throughput genotyping of very large numbers of SNPs with appropriate statistical correction for any differences observed.

Tobacco smoking is a well-established risk factor for OSCC (reviewed in Abnet, Arnold and Wei, 2018). We, therefore, investigated the only locus in our study associated with OSCC for interaction with smoking using a case-control analysis. We confirmed tobacco smoking as an independent risk factor for OSCC in this sample but did not find evidence

for an interaction between the *CHEK2* variant rs1033667 and smoking. A case-only analysis, rather than a case-control analysis, was done to investigate a gene-environment interaction for the rs1033667 variant with alcohol consumption owing to the very high alcohol consumption rate in the JCS control samples. No differences in allele frequencies of rs1033667 were observed when comparing drinkers with non-drinkers, which is consistent with a previous report that associations with OSCC at the chromosome 22q12 locus did not differ significantly between subgroups with different alcohol drinking statuses (Wu et al., 2012). Although the *ALDH2* SNP rs4767364 was not associated with OSSC in our study, this variant was tested for interaction with alcohol use because of strong evidence for interaction between variants in this gene and alcohol in other populations (McKay et al., 2011; Wu et al., 2012), but no interaction was detected.

Encouragingly, our current study has identified an SNP in the *CHEK2* locus that is associated with OSCC in the SAB population. This provides support for the hypothesis that at least some risk variants for OSCC will prove to be shared across multiple populations, and for the prior evidence of association at *CHEK2* in GWAS and DNA repair pathway analysis of OSCC in Chinese populations (Wu et al., 2012, 2014; Li et al., 2013). Although most risk variants from other populations do not replicate in the South African Black population, an African GWAS is needed to provide deep coverage of the known OSCC risk loci and to identify potential novel risk variants in African populations.

4.6. Funding

This work was supported by the Cancer Association of South Africa (CGM, CBdV, CML), the University of the Witwatersrand Research Council (CGM), the National Research Foundation of South Africa (CGM: Incentive grant to a rated researcher; WCC: German Academic Exchange Service-National Research Foundation Joint In-country Scholarship Programme, WCC: the National Research Foundation and Department of Science and Technology Thuthuka fund); Boehringer Ingelheim Fonds travel grant (WCC); The Grants, Innovation and Product Development Unit (GIPD) of the South African Medical Research Council (SAMRC) and the South African National Department of Health (ES), King's Health Partner's Research and Development Challenge Fund, King's College London (CGM, HB), the South African Medical Research Council Newton fund #045 to the ERICA-SA project

(CGM, CBdV, CML, ES) and the Wellcome Trust (grant 094491/Z/10/Z, NJP). In addition, this paper represents independent research part-funded by the National Institute for Health Research Biomedical Research Centre at South London and Maudsley National Health Service Foundation Trust and King's College London, and the National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' National Health Service Foundation Trust and King's College London, United Kingdom. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care. MIP, MM and CGM were supported by a grant from the South African Medical Research Council, with funds received from the South African National Department of Health, the Medical Research Council (United Kingdom) and GlaxoSmithKline via the Newton Fund grant #046. MM was the recipient of an International Centre for Genetic Engineering and Biotechnology Postdoctoral Fellowship.

4.7. Acknowledgements

We thank Sisters Gloria Mokwatle, Patricia Rapoho and Pheladi Kale for assistance with sample collection, and Phillip Tombleson (National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London), Margaret Urban and Tonicah Maphanga (National Cancer Registry, National Health Laboratory Service, South Africa) for assistance in the preparation and management of the sample database; Cassandra Soo and Natalie Smyth (Sydney Brenner Institute for Molecular Bioscience Biobank) and Ariella Rowe (International Centre for Genetic Engineering and Biotechnology) for DNA extractions; Dr Aron Abera and Felicity Azubuike (Inqaba Biotechnical Industries (Pty) Ltd. Southern Africa) for MassARRAY genotyping.

4.8. Supplementary

Supplementary Table 4.1. Case-control association study analysis for SNPs genotyped in 7 known risk loci for OSCC in the SAB populations for dominant inheritance model and recessive inheritance model.

				UCT				JCS			
				Dominant Model		Recessive Model		Dominant Model		Recessive Model	
Chr	Gene	SNP	Alleles	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)
2	*CASP8*	rs10931936	C/T	0.2850	0.86 (0.65-1.14)	0.8760	1.06(0.53-2.12)	0.4436	1.08 (0.89-1.31)	0.2439	1.28 (0.85-1.92)
2	*ALS2CR12*	rs13016963#	G/A	0.5837	0.93 (0.71-1.22)	0.3891	1.20(0.79-1.83)	0.9060	0.99 (0.82-1.20)	0.1513	1.21 (0.93-1.57)
2	*ALS2CR12*	rs10201587#	A/G	-	-	-	-	0.2184	0.89 (0.73-1.07)	0.2525	1.16 (0.90-1.50)
5	*TMEM173*	rs13181561	A/G	-	-	-	-	0.1402	0.86 (0.69-1.05)	0.7806	0.97 (0.78-1.20)
5	*TMEM173*	rs13153461	G/A	0.0665	0.66 (0.43-1.03)	NA	NA	-	-	-	-
10	*PLCE1*	rs17417407	G/T	0.0027**	0.67 (0.51-0.87)	0.7504	1.11(0.60-2.05)	0.8794	0.99 (0.81-1.20)	0.9548	0.99 (0.62-1.56)
10	*PLCE1*	rs7084339	G/A	-	-	-	-	0.1904	1.15 (0.94-1.41)	0.3717	1.11 (0.89-1.39)
10	*PLCE1*	rs3765524#	T/C	0.8439	1.03 (0.74-1.44)	0.9172	1.02(0.71-1.47)	0.2043	1.14 (0.93-1.41)	0.3695	1.11 (0.89-1.39)
10	*PLCE1*	rs2274223#	A/G	0.5536	1.08 (0.84-1.40)	0.6293	1.08(0.78-1.51)	0.4666	0.93 (0.77-1.13)	0.3775	0.90 (0.70-1.15)
10	*PLCE1*	rs11187850	A/G	-	-	-	-	0.3699	1.09 (0.90-1.32)	0.1365	1.44 (0.89-2.34)
12	*ALDH2*	rs4767364#	A/G	-	-	-	-	0.1286	1.19 (0.95-1.49)	0.7154	1.16 (0.53-2.55)
17	*ATP1B2/TP53*	rs1642764#	C/T	0.5070	1.09 (0.84-1.41)	0.8447	1.06(0.58-1.95)	0.6443	1.05 (0.86-1.28)	0.2367	0.72 (0.42-1.24)
17	*ATP1B2/TP53*	rs1641511	A/G	-	-	-	-	0.0517	0.83 (0.68-1.00)	0.4329	0.90 (0.70-1.17)
17	*TP53*	rs1800371#	G/A	0.1684	0.67 (0.39-1.18)	NA	NA	0.1665	1.34 (0.89-2.02)	NA	NA
21	*RUNX1*	rs2014300#	A/G	0.8974	1.02 (0.79-1.31)	0.0657	0.72(0.51-1.02)	0.6542	1.04 (0.87-1.26)	0.8975	0.98 (0.75-1.28)
21	*RUNX1*	rs2834718	T/A	-	-	-	-	0.5467	0.94 (0.79-1.14)	0.9634	1.01 (0.75-1.35)
22	*CHEK2*	rs4822983#	C/T	0.0219*	1.34 (1.04-1.71)	0.0016**	1.62(1.20-2.18)	0.7717	1.03 (0.85-1.25)	0.9855	1.00 (0.79-1.27)
22	*CHEK2*	rs1033667#	C/T	0.0161*	1.35 (1.06-1.72)	0.0079**	1.52(1.12-2.06)	0.3374	1.10 (0.91-1.33)	0.1390	1.21 (0.94-1.54)
22	*XBP1*	rs2239815#	C/T	0.0086**	1.39 (1.09-1.77)	0.0050**	2.34(1.29-4.25)	0.1332	0.86 (0.70-1.05)	0.7773	0.93 (0.55-1.57)

MAF – Minor allele frequency; C/T – Major allele/Minor allele; OR – Odds ratio, for each minor allele carried; 95% CI – 95% confidence interval; NA = not analysed in recessive model for rare variants with <5 homozygotes in cases/controls; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001. # - Previously published lead SNPs (9 – 14); other SNPs tag the lead SNPs.

Supplementary Table 4.2. Meta-analysis of 11 variants genotyped in the SAB populations under the dominant inheritance model.

Chr	Gene	SNP	Major Allele	Minor Allele	P	P(R)	OR (95% CI)	Q	I^2
2	*CASP8*	rs10931936	C	T	0.973	0.893	1.003 (0.856-1.175)	0.188	42.26
2	*ALS2CR12*	rs13016963	G	A	0.681	0.681	0.968 (0.832-1.134)	0.703	0.00
10	*PLCE1*	rs17417407	G	T	0.056	0.305	0.859 (0.729-0.991)	0.020*	81.40
10	*PLCE1*	rs3765524	T	C	0.237	0.237	1.112 (0.933-1.326)	0.614	0.00
10	*PLCE1*	rs2274223	A	G	0.823	0.823	0.982 (0.852-1.157)	0.362	0.00
17	*ATP1B2/TP53*	rs1642764	C	T	0.443	0.443	1.063 (0.897-1.222)	0.803	0.00
17	*TP53*	rs1800371	G	A	0.766	0.945	1.052 (0.762-1.451)	0.054	73.17
21	*RUNX1*	rs2014300	A	G	0.659	0.659	1.034 (0.863-1.161)	0.870	0.00
22	*CHEK2*	rs4822983	C	T	0.101	0.257	1.137 (0.976-1.322)	0.104	62.23
22	*CHEK2*	rs1033667	C	T	0.025*	0.072	1.189 (1.023-1.379)	0.195	40.34
22	*XBP1*	rs2239815	C	T	0.617	0.733	1.040 (0.884-1.201)	0.003**	88.78

OR – Odds ratio, for each minor allele carried; 95% CI – 95% confidence interval; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001. P – P-value for the fixed-effect meta-analysis. P(R) – P values for the random effect meta-analysis. Q – Cochrane's Q statistic P values. I^2 – I^2 heterogeneity index.

Supplementary Table 4.3. Meta-analysis of 11 variants genotyped in the SAB populations under recessive inheritance model.

Chr	Gene	SNP	Major Allele	Minor Allele	P	P(R)	OR (95% CI)	Q	I^2
2	*CASP8*	rs10931936	C	T	0.278	0.278	1.215 (0.858-1.733)	0.649	0.00
2	*ALS2CR12*	rs13016963	G	A	0.095	0.095	1.209 (0.972-1.514)	0.976	0.00
10	*PLCE1*	rs17417407	G	T	0.886	0.886	1.027 (0.722-1.493)	0.773	0.00
10	*PLCE1*	rs3765524	T	C	0.411	0.411	1.083 (0.898-1.313)	0.704	0.00
10	*PLCE1*	rs2274223	A	G	0.677	0.677	0.959 (0.795-1.175)	0.360	0.00
17	*ATP1B2/TP53*	rs1642764	C	T	0.455	0.455	0.856 (0.545-1.229)	0.348	0.00
17	*TP53*	rs1800371	G	A	NA	NA	NA	NA	NA
21	*RUNX1*	rs2014300	A	G	0.224	0.326	0.876 (0.694-1.051)	0.166	48.00
22	*CHEK2*	rs4822983	C	T	0.050*	0.332	1.204 (1.011-1.462)	0.014*	83.57
22	*CHEK2*	rs1033667	C	T	0.005**	0.013*	1.319 (1.091-1.601)	0.253	23.35
22	*XBP1*	rs2239815	C	T	0.101	0.416	1.390 (0.921-2.015)	0.022*	80.94

OR – Odds ratio, for each minor allele carried; 95% CI – 95% confidence interval; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001. P – P-value for the fixed-effect meta-analysis. P(R) – P values for the random effect meta-analysis. Q – Cochrane's Q statistic P values. I^2 – I^2 heterogeneity index.

Supplementary Table 4.4. Comparison of minor allele frequencies of the 19 SNPs across the UCT, JCS study cohort and the YRI, CHB and CEU populations from the 1000 Genome Project Phase 3 data.

Chr	Gene	SNP	Alleles	UCT MAF Control	JCS MAF Control	YRI MAF#	CHB MAF#	CEU MAF#
2	*CASP8*	rs10931936	C/T	0.20	0.20	0.20	0.29	0.31
2	*ALS2CR12*	rs13016963	G/A	0.35	0.38	0.42	0.30	0.40
2	*ALS2CR12*	rs10201587	A/G	-	0.39	0.51	0.73	0.48
5	*TMEM173*	rs13181561	A/G	-	0.49	0.64	0.57	0.25
5	*TMEM173*	rs13153461	G/A	0.05	-	0.03	0.43	0.75
10	*PLCE1*	rs17417407	G/T	0.21	0.19	0.21	0.15	0.18
10	*PLCE1*	rs7084339	G/A	-	0.46	0.53	0.81	0.69
10	*PLCE1*	rs3765524	T/C	0.47	0.46	0.55	0.81	0.73
10	*PLCE1*	rs2274223	A/G	0.40	0.43	0.36	0.19	0.31
10	*PLCE1*	rs11187850	A/G	-	0.19	0.20	0.18	0.31
12	*ALDH2*	rs4767364	A/G	-	0.11	0.11	0.09	0.69
17	*ATP1B2/ TP53*	rs1642764	C/T	0.20	0.18	0.26	0.50	0.50
17	*ATP1B2/ TP53*	rs1641511	A/G	-	0.42	0.47	0.44	0.23
17	*TP53*	rs1800371	G/A	0.03	0.02	0.04	0.00	0.00
21	*RUNX1*	rs2014300	A/G	0.40	0.36	0.47	0.88	0.85
21	*RUNX1*	rs2834718	T/A	-	0.33	0.39	0.88	0.85
22	*CHEK2*	rs4822983	C/T	0.39	0.42	0.44	0.21	0.36
22	*CHEK2*	rs1033667	C/T	0.38	0.39	0.32	0.29	0.33
22	*XBP1*	rs2239815	C/T	0.16	0.18	0.23	0.40	0.68

MAF – Minor allele frequency; C/T – First allele/Second (minor) allele; # – 1000 Genome Phase 3 allele frequencies. Frequencies of the second alleles are presented.

4.9. References

Abnet, C.C., Freedman, N.D., Hu, N., et al. (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics*. 42(9):764–767. DOI: 10.1038/ng.649.

Abnet, C.C., Arnold, M. & Wei, W.-Q. (2018) Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology*. 154(2):360–373. DOI: 10.1053/j.gastro.2017.08.023.

Arnold, M., Soerjomataram, I., Ferlay, J., et al. (2015) Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*. 64(3):381–387. DOI: 10.1136/gutjnl-2014-308124.

Auton, A., Abecasis, G.R., Altshuler, D.M., et al. (2015) A global reference for human genetic variation. *Nature*. 526(7571):68–74. DOI: 10.1038/nature15393.

Bye, H., Prescott, N.J., Matejcic, M., et al. (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*. 32(12):1855–1861. DOI: 10.1093/carcin/bgr211.

Bye, H., Prescott, N.J., Lewis, C.M., et al. (2012) Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis*. 33(11):2155–2161. DOI: 10.1093/carcin/bgs262.

Chang, J., Tan, W., Ling, Z., et al. (2017) Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nature communications*. 8:15290. DOI: 10.1038/ncomms15290.

Chang, J., Zhong, R., Tian, J., et al. (2018) Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics*. 50(3):338–343. DOI: 10.1038/s41588-018-0045-8.

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):106–113. DOI: 10.1089/bio.2017.0050.

Chen, X., Iliopoulos, D., Zhang, Q., et al. (2014) *XBP1* promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature*. 508(7494):103–107. DOI: 10.1038/nature13119.

Cui, R., Kamatani, Y., Takahashi, A., et al. (2009) Functional variants in *ADH1B* and *ALDH2* coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*. 137(5):1768–1775. DOI: 10.1053/j.gastro.2009.07.070.

Ferlay, J., Soerjomataram, I., Dikshit, R., et al. (2015) Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 136(5):E359–E386. DOI: 10.1002/ijc.29210.

Gabriel, S., Ziaugra, L. & Tabbaa, D. (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics*. Chapter 2:Unit 2.12. DOI: 10.1002/0471142905.hg0212s60.

Goh, L.L., Lim, C.W., Sim, W.C., et al. (2017) Analysis of Genetic Variation in CYP450 Genes for Clinical Implementation. *PloS one*. 12(1):e0169233. DOI: 10.1371/journal.pone.0169233.

GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* (New York, N.Y.). 348(6235):648–60. DOI: 10.1126/science.1262110.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., et al. (2015) The African Genome Variation Project

shapes medical genetics in Africa. *Nature*. 517(7534):327–332. DOI: 10.1038/nature13997.

Hetz, C., Chevet, E. & Harding, H.P. (2013) Targeting the unfolded protein response in disease. Nature reviews. *Drug discovery*. 12(9):703–19. DOI: 10.1038/nrd3976.

Jennis, M., Kung, C.-P., Basu, S., et al. (2016) An African-specific polymorphism in the *TP53* gene impairs p53 tumor suppressor function in a mouse model. *Genes & development*. 30(8):918–30. DOI: 10.1101/gad.275891.115.

Lee, S., Park, S., Lee, H., et al. (2018) ChIP-seq analysis reveals alteration of H3K4 trimethylation occupancy in cancer-related genes by cold atmospheric plasma. *Free radical biology & medicine*. 126:133–141. DOI: 10.1016/j.freeradbiomed.2018.08.004.

Li, W.-Q., Hu, N., Hyland, P.L., et al. (2013) Genetic variants in DNA repair pathway genes and risk of esophageal squamous cell carcinoma and gastric adenocarcinoma in a Chinese population. *Carcinogenesis*. 34(7):1536–1542. DOI: 10.1093/carcin/bgt094.

Machiela, M.J. & Chanock, S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* (Oxford, England). 31(21):3555–7. DOI: 10.1093/bioinformatics/btv402.

Matejcic, M. & Iqbal Parker, M. (2015) Gene-environment interactions in esophageal cancer. *Critical reviews in clinical laboratory sciences*. 52(5):211–31. DOI: 10.3109/10408363.2015.1020358.

Matejcic, M., Li, D., Prescott, N.J., et al. (2011) Association of a Deletion of *GSTT2B* with an Altered Risk of Oesophageal Squamous Cell Carcinoma in a South African Population: A Case-Control Study. *PLoS ONE*. 6(12):e29366. DOI: 10.1371/journal.pone.0029366.

Matejcic, M., Gunter, M.J. & Ferrari, P. (2017) Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. *Carcinogenesis*. 38(9):859–872. DOI: 10.1093/carcin/bgx067.

McKay, J.D., Truong, T., Gaborieau, V., et al. (2011) A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. *PLoS Genetics*. 7(3):e1001333. DOI: 10.1371/journal.pgen.1001333.

Murphy, G., McCormack, V., Abedi-Ardekani, B., et al. (2017) International cancer seminars: a focus on esophageal squamous cell carcinoma. *Annals of Oncology*. 28(9):2086–2093. DOI: 10.1093/annonc/mdx279.

Murphy, M.E., Liu, S., Yao, S., et al. (2017) A functionally significant SNP in TP53 and breast cancer risk in African-American women. *NPJ breast cancer*. 3:5. DOI: 10.1038/s41523-017-0007-9.

NCR. (2018) South African summary statistics of cancer diagnosed histologically in 2014. http://www.nicd.ac.za/wp-content/uploads/2017/03/2014-NCR-tables-1.pdf.

Southey, M.C., Goldgar, D.E., Winqvist, R., et al. (2016) *PALB2*, *CHEK2* and *ATM* rare variants and

cancer risk: data from COGS. *Journal of medical genetics*. 53(12):800–811. DOI: 10.1136/jmedgenet-2016-103839.

Syrmis, M.W., Moser, R.J., Whiley, D.M., et al. (2011) Comparison of a multiplexed MassARRAY system with real-time allele-specific PCR technology for genotyping of methicillin-resistant Staphylococcus aureus. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. 17(12):1804–10. DOI: 10.1111/j.1469-0691.2011.03521.x.

Teo, Y.-Y., Small, K.S. & Kwiatkowski, D.P. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics*. 11(2):149–160. DOI: 10.1038/nrg2731.

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):e1001182. DOI: 10.1371/journal.pmed.1001182.

Wang, L.-D., Zhou, F.-Y., Li, X.-M., et al. (2010). Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at PLCE1. *Nature Genetics*. 42(9):759–763. DOI: 10.1038/ng.648.

Westra, H.-J., Peters, M.J., Esko, T., et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*. 45(10):1238–1243. DOI: 10.1038/ng.2756.

Wu, C., Hu, Z., He, Z., et al. (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature genetics*. 43(7):679–684. DOI: 10.1038/ng.849.

Wu, C., Kraft, P., Zhai, K., et al. (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nature Genetics*. 44(10):1090–1097. DOI: 10.1038/ng.2411.

Wu, C., Wang, Z.Z., Song, X., et al. (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nature Genetics*. 46(9):1001–1006. DOI: 10.1038/ng.3064.

Zannini, L., Delia, D. & Buscemi, G. (2014) CHK2 kinase in the DNA damage response and beyond. *Journal of molecular cell biology*. 6(6):442–57. DOI: 10.1093/jmcb/mju045.

5. CHAPTER 5: Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations.

As discussed in the Introduction to this thesis (section 1.2, p.7) the candidate gene approach to the identification of genetic risk factors in complex disorders has significant limitations regarding the scope of such studies and the lack of the means to adequately adjust for population substructure in cases and controls. In addition, there may be substantial differences in the genes and genetic variants which confer susceptibility to OSCC in African populations compared to non-African populations. We have therefore carried out a GWAS for OSCC in the SAB population in order to scan all regions of the genome for evidence of association with this disorder and to use the genome-wide data to adjust for potential population substructure between cases and controls. In this chapter, a GWAS was conducted for 1,686 SAB OSCC cases and 3,217 population controls. This is the first African OSCC GWAS to date. After applying appropriate population substructure correction, our GWAS identified novel genetic risk loci for African OSCC on chromosome 9 and chromosome 2. The trans-ethnic meta-analysis of the African and Chinese OSCC GWAS further identified shared OSCC risk loci in populations of diverse ancestry. The work in this chapter has been submitted for peer-reviewed publication.

Chen, W.C., Brandenburg, J-T., Choudhury, A., Hayat, M., Sengupta, D., Swiel, Y., Babb de Villiers, C., Ferndale, L., Aldous, C., Soo, C., Lee, S., Curtis, C., Newton, R., Waterboer, T., Sitas, F., Bradshaw, D., Abnet, C.C., Ramsay, M., Parker, M.I., Singh, E., Lewis, C.M., Mathew, C.G. (2022) Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations. Submitted to *Nature Communications*.

The student performed genotype calling, data cleaning, and data analysis and wrote and reviewed the manuscript.

5.1. Abstract

Oesophageal squamous cell carcinoma (OSCC) has a high disease burden in sub-Saharan Africa. A genome-wide association study was performed comprising 1,686 African OSCC patients and 3,217 controls to investigate its genetic aetiology. One novel, genome-wide

significant risk locus was identified on chromosome 9 upstream of the *FAM120A* gene (rs12379660, P = $4.58x10^{-8}$, odds ratio = 1.28, 95% confidence interval = 1.22-1.34) and one African-specific risk locus was identified on chromosome 2 (rs142741123, P = $5.49x10^{-8}$) within the *MYO1B* gene. A meta-analysis was performed for the African OSCC GWAS with a Chinese OSCC GWAS (2,013 OSCC patients, 2,701 controls) and identified three genome-wide significant loci on chromosome 9 at *FAM120A* (rs12379660, $P_{meta} = 9.36x10^{-10}$), chromosome 10 at *PLCE1* (rs7099485, $P_{meta} = 1.48x10^{-8}$) and chromosome 22 at *CHEK2* (rs1033667, $P_{meta} = 1.47x10^{-9}$). This is the first GWAS of OSCC conducted in an African population and confirms genetic contributions to OSCC risk. The trans-ethnic meta-analysis identified both shared and distinct risk loci for OSCC.

5.2. Introduction

Oesophageal Cancer (OC) is the eighth most common cancer in the world and can be classified as either adenocarcinoma (ADC) or squamous cell carcinoma (SCC). ADC is primarily associated with gastric reflux and Barret's oesophagus and occurs mainly in industrialized countries (Enzinger & Mayer, 2003). Oesophageal SCC (OSCC) is more common in the developing world, which accounts for 88% of its global incidence (Arnold et al., 2015). High-risk areas for OSSC include Cixian in China, northern Iran, and sub-Saharan Africa. The age-standardized incidence rate (ASR) in Southern and Eastern Africa ranges from 6.8 to 12.7 cases per 100,000 (Sung et al., 2021) whereas in rural populations of the Eastern Cape in South Africa the ASR is 32 per 100,000 in males and 19.6 per 100,000 in females (Forman et al., 2013).

The aetiology of OSCC is complex; potential risk factors in Africa include low socioeconomic status, poor nutrition, tobacco smoking, alcohol use, smoke inhalation from cooking fires, consumption of fumonisin-contaminated maize, dental fluorosis and the presence of human papillomavirus (HPV) and human immunodeficiency virus (HIV) (Murphy et al., 2017; Menya, Kigen, et al., 2019; Menya, Maina, et al., 2019), but no single major cause has been identified. In South Africa, case-control studies in urban Soweto confirmed the association with smoking and alcohol, particularly smoking pipe tobacco and consuming home-brewed beer (Segal, Reinach & de Beer, 1988; Isaacson et al., 2015). The Johannesburg Cancer Study (JCS) reported that the combination of smoking and regular alcohol consumption increased the risk further, with an odds ratio of 4.4 (95%

confidence interval (CI): 3.2-6.1) (Pacella-Norman et al., 2002; Chen et al., 2020). Infection with HPV may contribute to OSCC risk (Schäfer et al., 2013), but extensive analysis of tumour tissues from patients seropositive for HPV antibodies in high-incidence regions did not find the consistent presence of HPV DNA, HPV mRNA or p16(INK4a) upregulation (Sitas et al., 2012). HIV infection was reported to confer an increase in risk for OSCC in Zambia, with an odds ratio of 2.3 (95%CI: 1.0-5.1), and no risk associated with HPV infection (Kayamba et al., 2015). Clinically, the development of OSCC is largely asymptomatic, resulting in late presentation with advanced dysphagia and a very poor prognosis. In Africa, treatment is largely palliative, including insertion of a stent to enable swallowing and radio or chemotherapy. Patients treated with stents in Kenya and Malawi, for example, had a median survival of only 250 and 210 days respectively after treatment (White et al., 2009; Thumbs et al., 2012). The high incidence and dismal prognosis for OC constitute a major ongoing public health problem in South and East Africa.

Genome-wide association studies (GWAS) of many diseases have yielded a rich harvest of robust associations, which provided substantial insights into their pathogenesis (reviewed in Mathew, 2008). In OSCC, GWAS from Japanese, Chinese and European populations have reported associations at multiple loci, including *ALDH2*, *ADH1B*, *PLCE1*, *PDE4D*, *TMEM173*, *HEATR3*, *TP53*, *MTMR3*, *CASP8*, *RUNX1*, *XPB1/CHEK2*, *CCHCR1*, *TCN2*, *TNXB*, *LTA*, *CYP26B1* and *FASN* (Cui et al., 2009; Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014; Chang et al., 2018). Previous studies tested single nucleotide polymorphisms (SNPs) that had been associated with OSCC in GWAS from other populations for association in Xhosa speakers from the South African Black (SAB) population (Bye et al., 2011, 2012; Matejcic et al., 2011; Chen, Bye, et al., 2019). Most SNPs showed no evidence of association with OSCC, although preliminary evidence was found for an association of a missense variant Arg548Leu in the *PLCE1* gene (Bye et al., 2012), and of an SNP rs1033667 in *CHEK2* (Chen, Bye, et al., 2019). These results suggest there may be substantial differences in the genetic determinants of susceptibility to OSCC in the SAB population as compared to other populations. A substantially larger and broader association study that includes cases and controls from other major SAB population groups is needed to investigate genetic susceptibility to OSCC in the Black South African population. Given the limited extent of knowledge of the genetic contribution to OSCC in African populations, we set out to investigate and identify population-specific genetic

susceptibility loci for OSCC using the genome-wide association approach in the first African OSCC GWAS analysis, as well as identifying shared genetic susceptibility loci for OSCC using trans-ethnic meta-analysis.

5.3. Results

Study participants and dataset. After data quality control (see Methods), our final study population consisted of 1,686 OSCC patients and 3,217 population controls (Supplementary Table 5.1), all of SAB ancestry. Our final genotyped dataset of 1,697,520 SNPs was submitted for imputation using the Sanger Imputation Service with the African Genome Resource Panel. Post imputation, SNPs with an Impute2 score of 0.6 or more (Supplementary Table 5.2) were retained, returning 14,472,257 SNPs for the final imputed dataset.

Population substructure in African cases and controls. Complex degrees of population structure were present within our case and control sample set enrolled from multiple sites within South Africa (Supplementary Table 5.1). Despite all participants self-reporting black African ancestry, additional methodological steps were taken to adjust for the sub-structure observed within the study population. A combination of four methods was used: **1.** ADMIXTURE analysis, **2.** Iterative random sub-sampling testing of controls, **3.** Eigen decomposition for principal component analysis and **4.** Kin-ship matrix (see Methods). ADMIXTURE analysis using the 1000Genome project (1KG) phase 3 reference data (Auton et al., 2015) was used to exclude individuals with high (>= 20%) European (CEU) or East-Asian (CHB) ancestry from our study population (Figure 1A). K = 6 was found to have the lowest cross-validation error (CV error = 0.51286). A total of 92 (1.84%) participants were removed due to significant non-SAB ancestry. Through the 30 rounds of iterative random sub-sampling testing of controls, large population substructure differences were observed between the male and female controls recruited from the Soweto region. A total of 1,213,862 SNPs were identified and removed from the final OSCC linear-mixed model (LMM) summary statistics (see Methods). Thereafter, principal component (PC) analysis through Eigen decompositions was performed. PCs 1 to 5 were selected (based on visual inspection of the eigenvalue curve, as well as generalized linear modelling (GLM) of the PCs with the binary phenotype) and incorporated into the LMM as covariates (Supplementary Figure 5.1). PC1 and PC2 accounted for the majority of the

variance observed in our dataset, with variance explained decreasing sharply with increasing numbers of PCs modelled. Overall, our cases and controls were well matched on PC (Figure 5.1B). When we compared the PC by study sites, samples from Soweto (both cases and controls) were homogeneous, with samples from other study sites clustering together (Figure 5.1C). Lastly, the inclusion of a kinship matrix to account for participants' relatedness using ~1 million markers in the LMM completed the population sub-structure correction methodology.

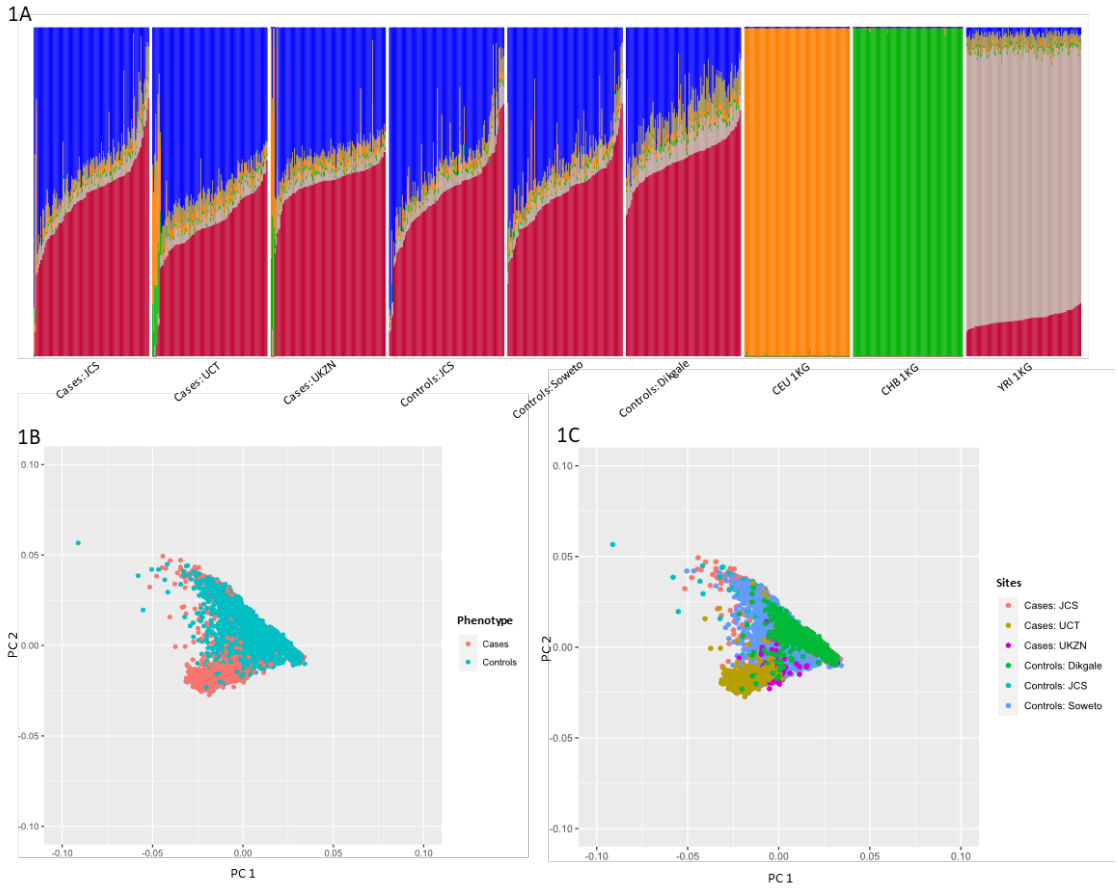


Figure 5.1. Admixture and Principal component plots of OSCC cases and controls.
A. Admixture plot of OSCC cases and controls from each study site (K=6). B. Principal component (PC) plot of study participants by case-control status. C. PC plot of study participants by study site. Legend: JCS – Johannesburg Cancer Study; UCT – University of Cape Town; UKZN – University of KwaZulu-Natal.

GWAS analysis of African OSCC cases and controls. The GWAS tested for association of the SNPs with OSCC. The LMM method was chosen for its effectiveness in correcting genetic inflation from polygenic backgrounds (Wen et al., 2018). The genomic inflation factor Lambda for our LMM is 1.01 (Figure 5.2A). We identified one genome-wide significant signal on chromosome 9 (rs12379660, $P_{LMM} = 4.58x10^{-8}$; odds ratio(OR) = 1.28, 95% confidence interval (95%CI): 1.22-1.34, Effect-allele frequency$_{cases}$ (EAF_{cases}) = 0.251) near the *FAM120A* gene, and one signal nearing genome-wide significance on

chromosome 2 (rs142741123, $P_{LMM} = 5.49x10^{-8}$, OR = 2.66, 95%CI: 2.43-2.89, $EAF_{cases} = 0.017$) in the *MYO1B* gene (Figure 5.2B & Table 5.1). Both signals have support from regional associations with SNPs in high linkage disequilibrium (LD) and with highly significant permutation P values (Chr9:rs10992729 – $P_{perm} = 9.71x10^{-8}$; Chr2:rs141499985 – $P_{perm} = 5.57x10^{-8}$). The lead SNP on chromosome 9, rs12379660, was genotyped rather than imputed and showed discrete genotype clusters (Supplementary Figure 5.2). SNPs in LD with this SNP span the entire *FAM120A* gene and the adjacent *FAM120AOS* gene (Figure 5.3A). The chromosome 2 locus has two independent signals within a 400KB window (rs142741123 & rs113702517, $r^2 = 0.0007$). The lead signal in this window (rs142741123) is located within the *MYO1B* gene (Figure 5.3B) and the second signal (rs113702517) is located within the *STAT4* gene ($P_{LMM} = 9.37x10^{-7}$, OR = 0.68, 95%CI: 0.57-0.78, $EAF_{cases} = 0.05$) (Figure 5.3C). Both signals are flanked by recombination hotspots. Top SNPs in both chromosome 2 signals have strong support from SNPs in LD and are likely to be African-specific since they are monomorphic in European and Asian populations (Supplementary Table 5.3). In addition, both these signals had permutation P values of $< 1x10^{-7}$ after 700 million adaptive permutation tests using the Plink permutation algorithm (Table 5.1) (see Methods). We also found a further 60 SNPs from 29 independent loci with suggestive association to OSCC ($P_{LMM} < 5x10^{-6}$) (Supplementary Table 5.4).

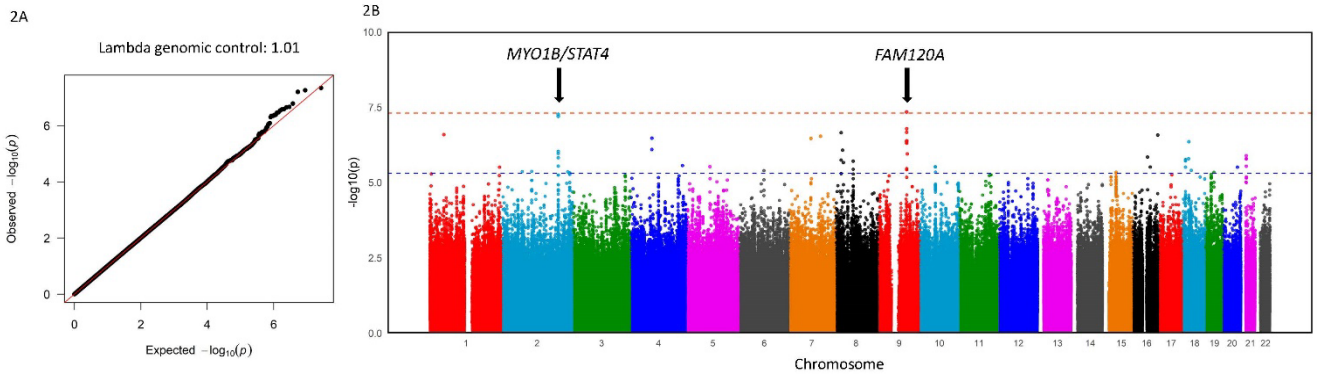


Figure 5.2. Quantile-Quantile and Manhattan plot of the African OSCC GWAS.
A. Quantile-Quantile plot of the OSCC GWAS, lambda for genomic control is 1.01. B. Manhattan plot of the OSCC GWAS, top hits on Chromosome 9 and Chromosome 2. The blue dash-line represents the 5×10^{-6} genome-wide suggestive threshold. The red dash-line represents the 5×10^{-8} genome-wide significant threshold.

Genetic heritability estimation. Using the generalized restricted maximum likelihood estimation implemented in LDAK (Speed et al., 2017) and the GLOBOCAN 2020's OC incidence (for South Africa: 6.8/100 000) as a proxy for the disease prevalence (Sung et

al., 2021), the estimated genetic heritability of OSCC (h2g) using 1,097,541 genotyped SNPs was 32.67% (SD: 1.98%) on the liability scale.

Replication of known OSCC risk loci. A total of 33 SNPs and 23 loci were assessed for association in our African GWAS based on prior evidence of association in genome-wide studies with at least suggestive evidence of association at $P < 5x10^{-6}$ (Supplementary Table 5.5) (Abnet et al., 2010; Bye et al., 2011, 2012; Wu et al., 2012, 2014; Hyland et al., 2016; Chang et al., 2018). Of the 33 SNPs, 24 SNPs had the same directional effect in the African OSCC GWAS. None of these known OSCC associated SNPs and loci reached genome-wide or suggestive significance thresholds in our study, although 3 loci showed modest association: rs1614972 at *ADH1C* (P = 0.011), rs130079 at *CCHCR1* (P = 0.037) and *rs1033667* at *CHEK2* (P = 0.005). Although the reported top signal at the *RUNX1* locus on chromosome 21, rs2014300, was not associated in our study (P = 0.099), a regional plot of this locus SNP located just upstream of the *RUNX1* gene, rs2834762, showed evidence of association ($P = 6.64x10^{-5}$, OR = 1.19, 95%CI: 1.13-1.24) (Figure 5.3D).

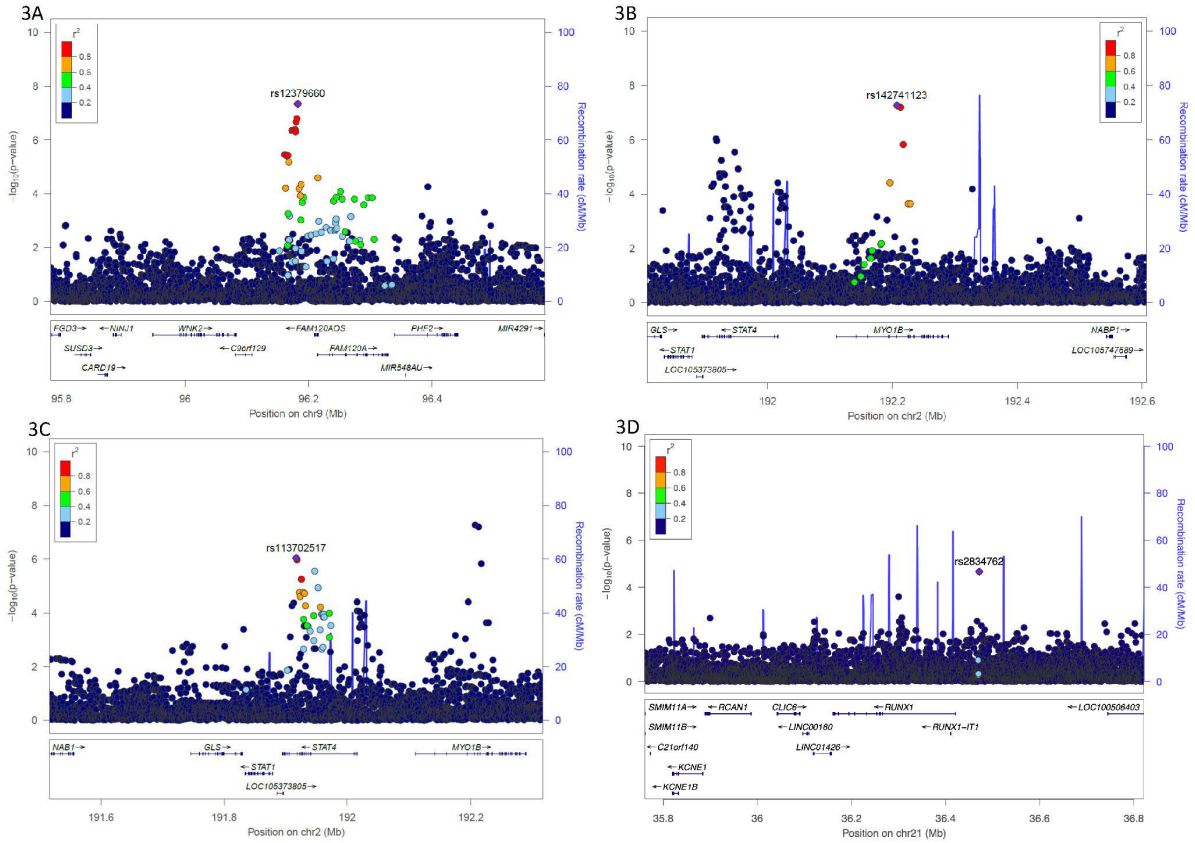


Figure 5.3. Locuszoom plots of the African OSCC GWAS hits.
A. Locuszoom plot of rs12379660/FAM120A on Chromosome 9. B. Locuszoom plot of rs142741123/MYO1B on Chromosome 2. C. Locuszoom plot of rs113702517/STAT4 on Chromosome 2. D. Locuszoom plot of rs2834763/RUNX1 on Chromosome 21.

Trans-ethnic meta-analysis of African and Chinese OSCC GWAS. We carried out a meta-analysis of our African GWAS with an OSCC GWAS from the Chinese population to increase the power of the study to identify risk loci and to look for evidence of transferability of risk loci across different ethnicities. Summary statistics from Abnet et al (Abnet et al., 2010), comprising 2,013 Chinese OSCC cases and 2,701 ethnically matched controls, were used for the meta-analysis. The primary Chinese GWAS was modelled with sex and PCs as covariates. Our African OSCC GWAS was remodelled to include sex as a covariate to match the modelling conditions of the Chinese GWAS. A total of 4,850,898 SNPs from the Abnet et al (Abnet et al., 2010) study overlapped with our study and were used for the meta-analysis using METAL (Willer, Li & Abecasis, 2010). A total of 12 SNPs from three loci reached the genome-wide significant threshold (Table 5.2). This includes *FAM120A* on chromosome 9 led by rs12379660 ($P_{meta} = 9.36x10^{-10}$); *CHEK2* on chromosome 22 led by rs1033667 ($P_{meta} = 1.47x10^{-9}$) and *PLCE1* on chromosome 10 led by rs7099485 ($P_{meta} = 1.48x10^{-8}$) (Figure 5.4). The chromosome 9 locus is novel and was identified in our African OSCC GWAS. A moderate level of association was present in the Chinese OSCC GWAS, but this locus did not reach genome-wide significance thresholds in previous OSCC GWAS (Figure 5.5A). The chromosome 22 locus at *CHEK2* was previously identified in the Chinese OSCC GWAS, with moderate levels of support in the African OSCC GWAS (Figure 5.5B). The chromosome 10 locus at *PLCE1,* previously identified in the Chinese OSCC GWAS, shows genome-wide significance in the meta-analysis, but the signal is primarily driven by the Chinese OSCC GWAS (Figure 5.5). Lastly, a signal on chromosome 2 nearing genome-wide significance was found near the *SLC16A14* gene led by rs34115901 ($P_{meta} = 1.39x10^{-7}$). A moderate level of association with this SNP was present in both GWAS but did not reach genome-wide significance thresholds, whereas the increase in power of the meta-analysis provided strengthened evidence of association for this locus (Figure 5.5D). Strong additional support is evident from nearby SNPs within this locus.

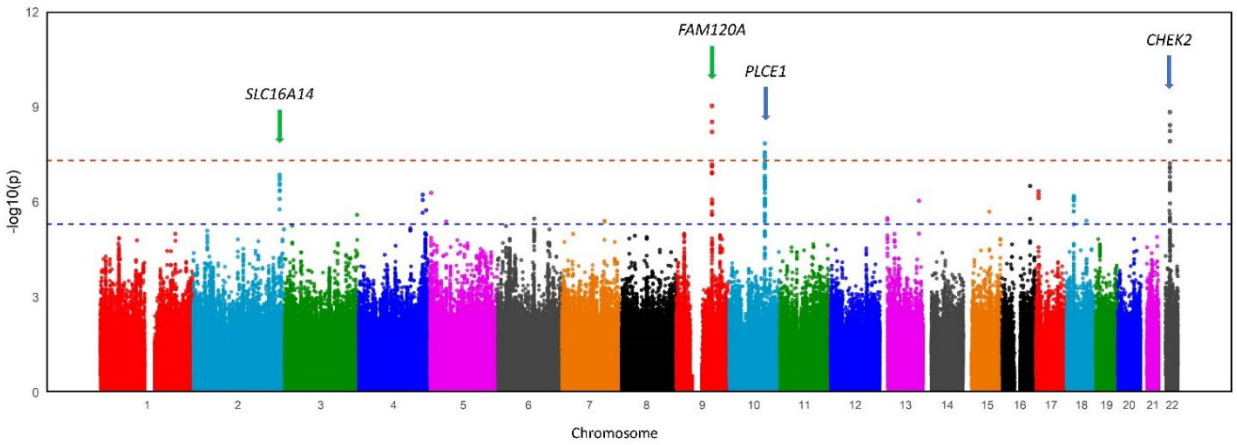


Figure 5.4. Meta-analysis Manhattan plot for African OSCC and Chinese OSCC.
With top hits on Chromosomes 9, 10 and 22. The blue dash-line represents the 5 x 10^{-6} genome-wide suggestive threshold. The red dash-line represents the 5 x 10^{-8} genome-wide significant threshold.

Functional analysis using bioinformatics tools. The two most strongly associated SNPs at the *FAM120A* locus (rs12379660 and rs10992729) are among 7 highly significant eQTLs for *FAM120AOS* in oesophageal mucosa and the 2 eQTLs for *FAM120AOS* in oesophageal muscularis (Supplementary Table 5.6) using GTEx version 8 (GTEx Consortium, 2013). No eQTL data is available for the rare African-specific SNPs in the *MYO1B/STAT4* region. The top associated SNPs at both *PLCE1* and *CHEK2* are strong eQTLs in oesophageal mucosa (Supplementary Table 5.7). Using Reactome (Fabregat et al., 2017; Griss et al., 2020; Jassal et al., 2020), FAM120A was found to interact with YBX1, a DNA- and RNA-binding protein involved in various processes, such as translational repression, RNA stabilization, mRNA splicing, DNA repair and transcriptional regulation (Horwitz, Maloney & Ley, 1994; Chen et al., 2000; Capowski et al., 2001; Gaudreault, Guay & Lebel, 2004; Chattopadhyay et al., 2008; Chen, Li, et al., 2019).

Table 5.1. Top SNPs associated with African oesophageal squamous cell carcinoma from Chromosome 9 and Chromosome 2.

rsID	Chr	Position (hg19)	Gene	Alleles (Effect/Non-Effect)	EAF cases	EAF controls	OR (95% CI)	P_{LMM} value	P_{perm} value[$]	Impute2 Score
[%]rs12379660	9	96182487	*FAM120A*	A/G	0.251	0.207	1.28 (1.22-1.34)	$4.58x10^{-8}$	$3.71x10^{-8}$	#
rs10992729	9	96181075	*FAM120A*	T/C	0.254	0.212	1.26 (1.21-1.32)	$1.64x10^{-7}$	$9.71x10^{-8}$	0.995
[%]rs142741123	2	192207115	*MYO1B*	C/T	0.017	0.007	2.66 (2.43-2.89)	$5.49x10^{-8}$	$3.86x10^{-8}$	0.935
rs141499985	2	192212930	*MYO1B*	T/G	0.017	0.007	2.64 (2.41-2.87)	$6.33x10^{-8}$	$5.57x10^{-8}$	0.929

% - Lead SNP at each locus, with their strongest supporting SNP

EAF - Effect-allele frequency

OR - Odds ratio

P_{LMM} - P-value from the linear mixed model

$ - P-value from 700 million adaptive permutations - see Methods

- Genotyped SNP

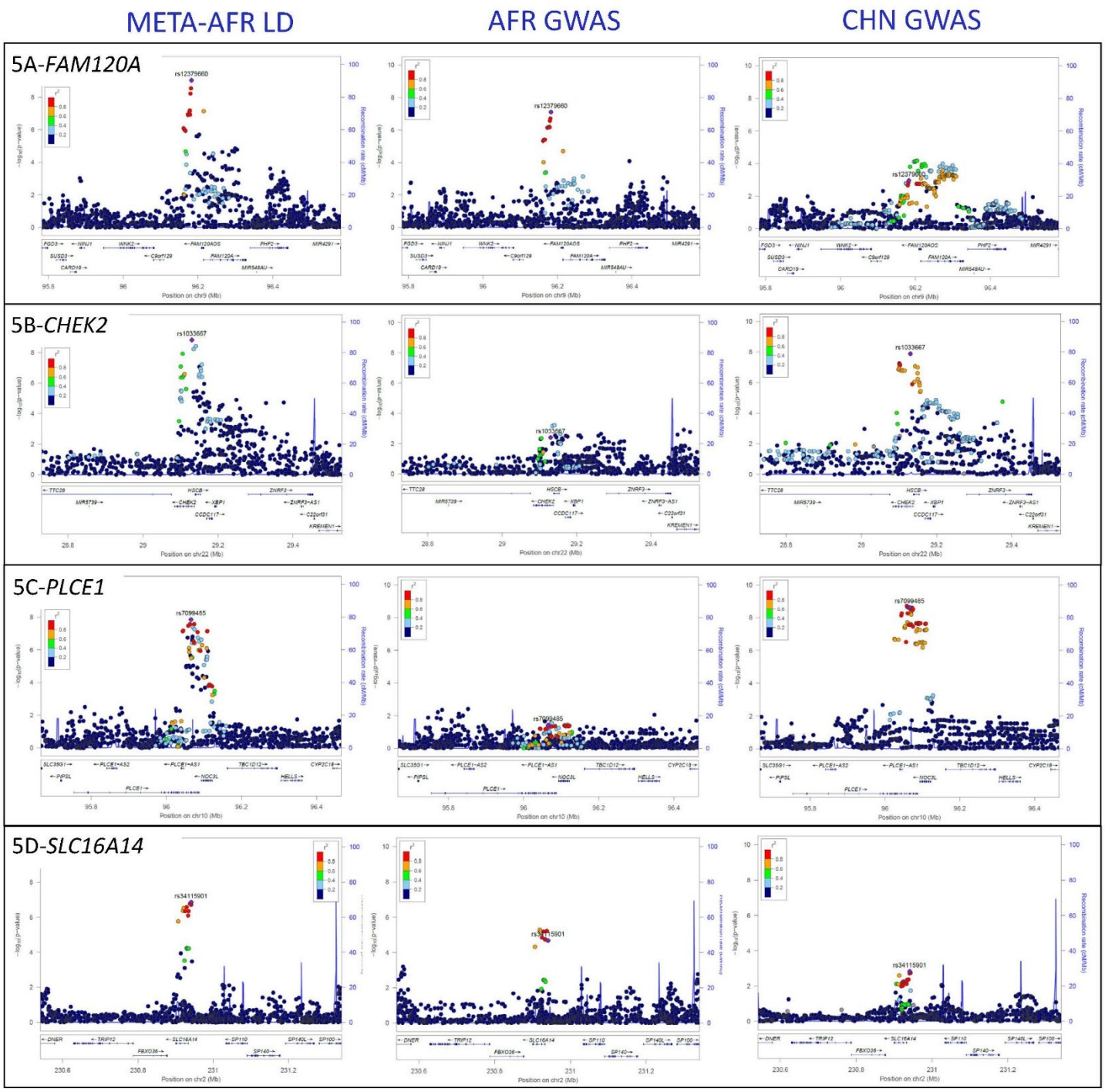


Figure 5.5. Meta-analysis Locus zoom plots.
A. Chromosome 9, FAM120A/FAM120AOS. B. Chromosome 22, CHEK2. C. Chromosome 10, PLCE. D. Chromosome 2, SLC16A14. The first column contains Locus zoom plots of the meta-analysis (African GWAS and Chinese GWAS), using the African GWAS LD data as a reference. The second column contains Locus zoom plots of the African GWAS, using African LD data as a reference. The third column contains the Chinese GWAS plots, using Chinese LD data as a reference.

Table 5.2. Top signals from African-Chinese meta-analysis for oesophageal squamous cell carcinoma.

Chr	Position (hg19)	rsID	Allele[#]	Effect allele freq (CHN)	Effect allele freq (AFR)	P value (CHN)	P value (AFR)[$]	P value meta-analysis	Odds ratio	95% Confidence interval	Heterogeneity P value[@]
2	230941048	[%]rs34115901	A/G	0.228	0.029	$1.48x10^{-3}$	$2.06x10^{-5}$	$1.39x10^{-7}$	1.86	1.72-2.01	0.4764
2	230937668	rs34590391	A/G	0.249	0.028	$4.27x10^{-3}$	$5.90x10^{-6}$	$1.65x10^{-7}$	1.81	1.67-1.96	0.2580
2	230940750	rs35396967	A/G	0.246	0.030	$1.94x10^{-3}$	$2.17x10^{-5}$	$1.97x10^{-7}$	1.81	1.68-1.96	0.4472
9	96182487	[%]rs12379660	A/G	0.406	0.251	$1.09x10^{-3}$	$7.89x10^{-8}$	$9.36x10^{-10}$	1.51	1.42-1.60	0.1533
9	96181075	rs10992729	T/C	0.388	0.254	$1.51x10^{-3}$	$2.01x10^{-7}$	$2.97x10^{-9}$	1.50	1.41-1.60	0.1694
9	96179785	rs1351117	A/G	0.408	0.254	$2.20x10^{-3}$	$2.75x10^{-7}$	$6.12x10^{-9}$	1.48	1.39-1.58	0.1579
10	96065694	[%]rs7099485	C/T	0.231	0.456	$2.26x10^{-9}$	$3.74x10^{-2}$	$1.42x10^{-8}$	1.94	1.86-2.03	0.0049
10	96074157	rs7903902	C/T	0.231	0.456	$3.38x10^{-9}$	$4.70x10^{-2}$	$2.74x10^{-8}$	1.93	1.85-2.01	0.0046
10	96063279	rs3781265	A/T	0.231	0.459	$2.26x10^{-9}$	$5.48x10^{-2}$	$2.76x10^{-8}$	1.94	1.86-2.02	0.0035
10	96058298	rs3765524	T/C	0.230	0.465	$1.88x10^{-9}$	$6.62x10^{-2}$	$3.42x10^{-8}$	1.94	1.86-2.03	0.0027
10	96070375	rs3781264	G/A	0.170	0.234	$6.50x10^{-9}$	$4.55x10^{-2}$	4.00x10-8	2.05	1.96-2.14	0.0061
22	29130300	[%]rs1033667	T/C	0.279	0.442	$1.28x10^{-8}$	$3.82x10^{-3}$	$1.47x10^{-9}$	1.84	1.76-1.92	0.0417
22	29140823	rs5752783	C/A	0.731	0.506	$8.80x10^{-7}$	$5.97x10^{-4}$	$3.78x10^{-9}$	0.58	0.41-0.82	0.2682
22	29135121	rs9613668	C/G	0.275	0.506	$1.23x10^{-6}$	$6.78x10^{-4}$	$5.79x10^{-9}$	1.73	1.65-1.81	0.2781
22	29105415	rs5752773	C/G	0.779	0.419	$1.62x10^{-7}$	$4.42x10^{-3}$	$1.20x10^{-8}$	0.55	0.34-0.87	0.0807

SLC16A gene at Chromosome 2, *FAM120A* gene at Chromosome 9, *PLCE1* gene at Chromosome 10 and *CHEK2* gene at Chromosome 22.
% - Lead SNP at each locus, with their supporting SNP.
# - Effect allele/Non-effect allele
$ - P-value (AFR) differs from the main LMM as the AFR GWAS was remodelled to match the conditions of the CHN GWAS.
@ - Heterogeneity P-value threshold is $P < 0.05$.

5.4. Discussion

The genetic contribution to oesophageal squamous cell carcinoma disease risk is not well understood. The unique geographical distribution of OSCC across two hot spots (north-central China to the Caspian Sea and eastern to southern Africa) suggests possible shared disease aetiology among populations located far apart (Sung et al., 2021). The common epidemiological risk factors described for OSCC to date do not account for the global burden of the disease (Kamangar et al., 2009; Abnet, Arnold & Wei, 2018). Previous studies have provided strong evidence for genetic risk factors for OSCC; however, most large-scale genetic studies were conducted in Asian or European populations (Wang et al., 2010; Wu et al., 2011, 2012, 2014), with smaller candidate gene studies conducted in African populations (Simba et al., 2019). Our study identified population-specific risk loci for OSCC in a South African Black (SAB) population and provide evidence for a substantial contribution of genetic factors to the risk of OSCC. We also identified trans-ethnic, shared genetic risk alleles between the Chinese population and the SAB population.

Given the complex genetic diversity and continental and regional population substructure present within the African and SAB populations (Choudhury et al., 2020, 2021; Sengupta et al., 2021), the ability to control for this complexity in GWAS is critical to the success of the GWAS. In this study, we used a combination of four methodologies to control for the complex population substructure and demonstrated the robustness of this combined approach. First, we screened our study sample by admixture analysis and removed individuals with a substantial degree of European or Asian ancestry; although all participants self-identified as African, nearly 2% were excluded from the final analysis by this process. Secondly, using PCs as covariates and incorporating the kinship matrix in the linear-mixed modelling, we were able to capture the majority of residual genetic variance within our dataset; The iterative random sub-sampling permutation testing of controls allowed us to identify and remove SNPs that produced false-positive association signals is the OSCC GWAS due subpopulation-specific differences. Although unexpected, the genetic differences between male and female controls recruited from Soweto may be accounted for by the labour migration history of South Africa, since the Gauteng province has long been a hotspot for migrant male labourers seeking employment (Statistics South Africa, 2006, 2010). By randomly assigning controls to "case-control" groups over 30 permutations, we were able to identify and exclude most of the signals driven by

population substructure differences rather than by the disease phenotype. This combined approach, in addition to a rigorous data quality control (see Methods) workflow provided a firm foundation for the OSCC GWAS.

We detected robust signals for African OSCC on Chromosome 9 and Chromosome 2. Several lines of evidence support our lead SNP on Chromosome 9, rs12379660, upstream of the *FAM120A* gene. This association is supported by multiple SNPs in LD across the *FAM120A* gene locus, and the top SNPs remained significant after 700 million adaptive permutation tests. Also, this was the strongest signal in the meta-analysis, with the Chinese data increasing the significance of the association at rs12379660 to a combined P value of 9.36×10^{-10}. This SNP is common in both populations of interest (EAF_{case}: 40% in CHN vs 25% in AFR), and the evidence is indicative of shared risk alleles for OSCC in the Chinese and African populations at *FAM120A*. The potential functional significance of the top SNPs at this locus is suggested by their co-localisation with highly significant cis-eQTLs in *FAM120AOS* in both oesophageal mucosa and oesophageal muscularis tissue. The *FAM120A* gene encodes the protein FAM120A, also known as the constitutive co-activator of PPAR-gamma-like protein 1, which is a key component of the oxidative stress-induced survival signalling pathway. It activates SRC family kinases and enables SRC family kinases to phosphorylate and activate PI3-kinase (GTEx Consortium, 2013). *FAM120A* has been reported to be overexpressed in several cancers, including colorectal cancer and head and neck squamous cell carcinoma (Cerami et al., 2012).

The signal at *MYO1B* on chromosome 2 has not been reported before in non-African populations. The top SNP has an allele frequency of 1.7% in the SAB population and is monomorphic in other populations (Supplementary Table 5.3). The gene of interest in this window, *MYO1B*, encodes the myosin 1b protein which has been linked to cellular proliferation and migration in head and neck squamous cell carcinomas and cervical cancer with HPV co-infection (Chapman et al., 2015; Ohmura et al., 2015; Zhang et al., 2018). A second, independent locus, *STAT4* is present within 400kb of *MYO1B*. The lead SNPs in both loci are LD independent of each other. *STAT4* is involved in cytokine signalling in the immune system by binding to IL12RB2 in the interleukin-12 receptor complex (Naeger et al., 1999).

The trans-ethnic meta-analysis confirmed our previous findings that *CHEK2* is a risk factor for OSCC in the SAB population, although at lower risks when compared to the Chinese population (OR_{SAB}(95%CI): 1.11 (1.06-1.16) vs OR_{CHN}(95%CI): 1.33 (1.20-1.46)). However, *PLCE1*, which is very strongly associated with OSCC in the Chinese population, shows only nominal evidence of association in the SAB population with the lowest P-value of 0.037. There are several possible explanations for this. One is that the causal variant at this locus has not been genotyped and is in low LD with genotyped SNPs at this locus in African populations, hence generating a very weak signal. Other explanations are that the causal variant may have arisen after the Out-of-Africa migration (López, van Dorp & Hellenthal, 2015), or there may be gene x environmental interactions involved which are specific to the Chinese population. Even though few of the known OSCC genetic risk factors were replicated in our study, genetics plays a sizable role in the aetiology of African OSCC.

Our genetic heritability analysis (on the liability scale) estimates the genetic contribution for OSCC is approximately 33% in our African study, using global OC incidence data as a more robust estimate for the disease prevalence. This finding is comparable with previous findings of 38% genetic heritability for OSCC estimated in Asian populations (Sampson et al., 2015).

Several reasons may account for the differences in the heritability estimates by Sampson et al and our study. The number of SNPs used to calculate the h2g by Sampson et al was 406,167 SNPs. The number of SNPs used by our study for the h2g estimation was 1,097,541 SNPs. The inclusion of more SNPs may have provided a more accurate h2g estimation. The h2g model by Sampson et al included correction for sex, sub-studies and 20 principal components, whereas our model used LD reduction and a kinship matrix in the calculation. Variations in the modelling parameters may contribute to the differences in variance observed. Despite both studies using the restricted maximum likelihood estimation (REML) approach, we have observed a variation in its implementation in different software packages. Sampson et al used GCTA, whereas our study used LDAK. The choice to use LDAK was made based on its superior performance and previous known limitations of GCTA producing higher heritability estimates (e.g., Kumar et al, 2015). As prevalence data were not available for OSCC in the South African or Southern African populations, we used incidence data as a proxy for prevalence. Incidence is the

number of new cases diagnosed in a specific population for a specific year, and prevalence is the number of people diagnosed with cancer and is still alive at a given time point, the true prevalence of OSCC in our population is likely higher than the incidence. Thus, if prevalence data were available, it would have yielded a higher h2g for our population. Lastly, it is plausible that the percentage of genetic heritability may be different between the Chinese and African, particularly the South African black population, despite shared genetic risk factors.

Although our GWAS is the largest genetics study of African OSCC, a much larger sample size will be required to detect a larger proportion of potential genetic loci associated with OSCC in African populations. A GWAS of OSCC in African populations is being carried out by the African Esophageal Cancer Consortium (van Loon et al., 2018) in OSCC cases and ethnically matched controls from east Africa. This will allow future replication and meta-analysis studies of African OSCC.

In conclusion, this study is the first GWAS for OSCC conducted in an African population. Our study for the first time provides evidence of substantial genetic contributions to OSCC risk in an African population and has identified novel and shared risk loci for this important African cancer. It also provides an example of the power of trans-ethnic meta-analysis to identify common and distinct risk loci in populations of diverse ancestry.

5.5. Methods

Study design. This case-control 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). The overall schematic and workflows for the study are outlined in Supplementary Figure 5.3.

Study participants. Patients with histologically confirmed oesophageal squamous cell carcinoma (OSCC) were recruited from three study sites across South Africa: the Johannesburg Cancer Study (JCS) from Soweto, Gauteng Province (Chen et al., 2020), the University of the Cape Town (UCT), Western Cape Province (Bye et al., 2011, 2012), and Greys Hospital, Pietermaritzburg, KwaZulu-Nata Province (Chen, Bye, et al., 2019). Ethnically matched population controls samples were provided by the H3Africa AWI-Gen

study (from Soweto, Gauteng Province and Dikgale, Limpopo Province) (Ramsay et al., 2016; Ali et al., 2018) and the JCS (Soweto, Gauteng Province) (Chen et al., 2020) (Supplementary Table 5.1).

Bio-sampling. Genomic DNA (gDNA) was isolated from peripheral blood samples collected from all study participants. DNA isolation methodologies were previously described (Bye et al., 2011, 2012; Chen et al., 2018). In summary, gDNA was isolated using either a kit-based DNA extraction with the Qiagen DNA FlexiGene kit or the salting-out method (Miller, Dykes & Polesky, 1988). Isolated gDNA were resuspended in low Tris-EDTA buffer and stored at -80C within a University of the Witwatersrand Human Research Ethics Committee (Medical) approved biobank until use (Soo et al., 2017).

Genotyping. Genotyping of all gDNA samples (cases and controls) was done using the H3Africa Custom African ~2.3 million SNP Array (Illumina) (Mulder et al., 2018) (https://www.h3abionet.org/h3africa-chip). Genotyping of all OSCC cases and the JCS population controls were performed at the Genomics Core Facility, Social Genetics & Development Psychiatry Centre, King's College London as a part of the ERICA-SA study. Genotyping of the AWI-Gen population control samples were performed using the Illumina FastTrack Sequencing Service (https://www.illumina.com/services/sequencing-services.html). Raw intensity data files were used for data analysis. Genotype clustering and calling were done for all cases and controls using the Illumina Array Analysis Platform Genotyping orchestrated command-line workflow, using the Illumina GenCall algorithm, predefined cluster file and manifest file provided by Illumina (https://emea.support.illumina.com/downloads/iaap-genotyping-orchestrated-workflow.html#:~:text=Support%20Center%3A,GTC%20format%20and%20PED%20Files). The Plink software version 1.9 was used for genotype data management (Chang et al., 2015). A modified version of the H3ABioNet/H3Agwas Pipeline Version 3 (https://github.com/h3abionet/h3agwas) was used for data formatting and data quality control. Parameters for the data quality control are outlined in Supplementary Table 5.8. Post genotyping, the raw data achieved a total genotyping rate of 97.83%. The genotyping rate improved to 99.92% post-data QC.

Imputation. The Sanger Imputation Service was used for genotype imputation (https://imputation.sanger.ac.uk/). Pre-imputation QC was applied to the dataset before imputation (Supplementary Table 8). Imputation was done using the African Genome Resource panel with pre-phasing done with EAGLE2. An Impute2 score of 0.6 was used as the post-imputation cut-off. Post imputation, 14,472,257 SNPs were retained for the final imputed dataset, of which 12,843,994 SNPs were imputed SNPs and 1,628,263 were genotyped SNPs.

Population sub-structure control. Given the complexity of the genetic architecture and the presence of extensive population substructure present in the Black South African populations (Choudhury et al., 2020, 2021; Sengupta et al., 2021), our study took four approaches to account and control for the population sub-structure within our sample cohort: **1.** ADMIXTURE analysis using CEU (n=99, European), CHB (n=103, Han Chinese, East Asian) and YRI (n=108, Yoruba, Nigeria) individuals from the 1000 Genomes Project (1KG) as the reference population. For this analysis, 275,209 LD independent SNPs were present in both the 1KG dataset as well as the African OSCC GWAS dataset and were used. ADMIXTURE analysis was performed for K = 2 to K = 9 using ADMIXTURE v1.3 (Alexander, Novembre & Lange, 2009) and K = 6 was found to have the lowest cross-validation error (Supplementary Figure 5.4). Study individuals with >= 20% CEU or CHB genetic contributions were excluded. Visualization of the ADMIXTURE results was done using Stata 17.0 (Stata Corp). **2.** Iterative random sub-sampling testing of controls was done to identify population outlier SNPs. Under this iterative testing, we tested 5 different models: a) a random sub-sampling of Soweto controls (n=800) tested against a random sub-sampling of Dikgale controls (n=800); b) random sampling of controls from the combined pool of Soweto and Dikgale controls (n=800) tested against a random sub-sampling of Soweto controls (n=800); c) a random sub-sampling of Soweto male controls (n=800) tested against a random sub-sampling of Soweto female controls (n=800); d) a random sub-sampling of Soweto pooled controls (n=800) tested against a random sub-sampling of Soweto pooled controls (n=800); e) a random sub-sampling of Dikgale pooled controls (n=500) tested against a random sub-sampling of Dikgale pooled controls (n=500). Each model was subjected to 30 permutations and modelled using linear-mixed model (LMM) methods (Gemma 0.98.1) (Zhou & Stephens, 2012). A total of 1,213,862 SNPs with nominal p-values of less than 0.05 and appearing 10 times or more out of the

30 permutations under each model were flagged and removed from the final GWAS LMM summary statistics. **3.** Eigen decomposition for PC analysis using LD independent SNPs (100kb window, 20 SNPs within each window, at an r^2 of 0.2). Eigen decomposition was performed using Plink v.1.9 (Chang et al., 2015). After LD pruning of the genotyped dataset, 428,379 (out of 1,699,678 genotyped SNPs) LD independent SNPs were retained for PC analysis. The results were visualized in R (https://www.R-project.org/) using ggplot2 (Wickham, 2016). **4.** A genetic relationship/kinship matrix (GRM) was constructed using ~1 million genotyped SNPs. The GRM is incorporated into the LMM for the final GWAS modelling.

SNP-based heritability estimation (h2g). We calculated SNP-based heritability estimates (h2g) for OSCC using restricted maximum likelihood estimations implemented in LDAK (Speed et al., 2017). LDAK weighting was performed to account for LD using a correlation squared threshold of 0.98. This allowed for the thinning of predictors from 1,699,678 SNPs to 1,097,541 SNPs. A kinship matrix was computed on the thinned predictors and used for the h2g estimation. The h2g for OSCC was estimated on the liability scale using Globocan 2020's oesophageal cancer incidence in South Africa as a proxy for prevalence (Sung et al., 2021).

GWAS LMM modelling. For the LMM, the binary case-control phenotype, together with PCs1-5 as covariates and the kinship matrix was used for the modelling. LMM was done using Gemma v.0.98.1 (Zhou & Stephens, 2012). Odds ratios were approximated using the beta and case-control ratios. Data visualization was done in Stata 17.0 (Stata Corp) and R's ggplot2 (https://www.R-project.org/) (Wickham, 2016).

Permutation testing. 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 linearized phenotype derived by residualizing Binary phenotype and PC1-5 under an additive model. Seven hundred million adaptive permutations were the limit of the AWS EC2 FPGA deployment and sufficient for our application.

Meta-analysis. A meta-analysis of SNPs common to both our study and the GWA study by Abnet et al (2010) was performed using METAL (Willer, Li & Abecasis, 2010). Both GWAS were modelled using the same conditions (Sex and PCs used as covariates). The meta-analysis was performed using P values and direction of effect, weighted by sample size (1,686 African OSCC cases, 3,217 ethnically matched populations and 2,013 Chinese OSCC cases, 2,701 ethnically matched controls), with missing SNP calls from a study would decrease the weighting. 4,850,898/7,599,678 SNPs from Abnet et al overlapped with our dataset.

Fine mapping & Functional analysis of associated variants. We performed fine mapping using LocusZoom v1.4 (Pruim et al., 2010) for all top hits with $p < 5x10^{-6}$, with a 400kb flanking nucleotide window and using precomputed LD information from our study population. The FUMA online application was used (P_{LMM} threshold = $5x10^{-6}$) to annotate and interpret associated GWAS variants (Watanabe et al., 2017), and to annotate co-localized eQTLs in the oesophagus tissues of interest from GTEx version 8 (GTEx Consortium, 2013). REACTOME pathway database was utilized to analyse pathway knowledge of top GWAS hits (Fabregat et al., 2017; Griss et al., 2020; Jassal et al., 2020).

5.6. Data Availability

The African OSCC data used in this study are available to interested researchers through the European Genome-Phenome Archive (EGA), subject to controlled access review by the Data and Biospecimen Access Committee of the University of the Witwatersrand. African OSCC genotype dataset accession number: *Pending*. The AWI-Gen data used in this study are available to interested researchers through EGA, subject to controlled access review by the Data and Biospecimen Access Committee of the H3Africa Consortium. AWI-Gen genotype dataset accession number: EGAD00010001996. GWAS Catalog (https://www.ebi.ac.uk/gwas/). Summary statistics reported in the paper are accessible on GWAS Catalog at the accession numbers: *Pending*.

5.7. Acknowledgements

This work was supported by the South African Medical Research Council, the National Health Laboratory Service (to E.S.) and the Cancer Association of South Africa (to C.G.M., C.B.dV. and C.M.L.). The Evolving Risk Factors for Cancers in African Populations (ERICA-

SA) Study and the Johannesburg Cancer Study have been 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) (MRCRFA-SHIP 01-2015). 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. German Academic Exchange Service-National Research Foundation Joint In-country Scholarship Programme (to W.C.C.) and the National Research Foundation and Department of Science and Technology Thuthuka fund (to W.C.C.); Boehringer Ingelheim Fonds travel grant (to W.C.C.); The Grants, Innovation and Product Development Unit (GIPD) of the South African Medical Research Council (SAMRC) and the South African National Department of Health (to E.S.), King's Health Partner's Research and Development Challenge Fund, King's College London (to C.G.M). We acknowledge the NIH-funded H3Africa Consortium Collaborative Centre, AWI-Gen, for sharing data for the population controls used in our study (PI: M.R.).

We thank sisters Gloria Mokwatle, Patricia Rapoho and Pheladi Kale for assistance with sample collection, Ariella Amar and Phillip Tombleson (National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London), Margaret Urban and Tonicah Maphanga (National Cancer Registry, National Health Laboratory Service, South Africa) for assistance in the preparation and management of the sample database; Scott Hazelhurst, Shaun Aron, Phelelani Mpangase and Natalie Smyth (Sydney Brenner Institute for Molecular Bioscience Biobank) and Ariella Rowe (International Centre for Genetic Engineering and Biotechnology) for bioinformatics support and sample processing.

5.8. Author contributions

C.G.M., W.C.C., C.M.L., D.B., F.S., E.S. designed the study.

C.G.M., D.B., F.S., E.S., T.W., R.N., C.B.dV., C.M.L. acquired the funding.

W.C.C., C.B.dV., L.F., C.S., L.S., C.J.C., F.S., M.R., C.A., M.I.P., E.S. sample acquisition, processing and management and genotyping.

W.C.C., J.T.B., M.H., Y.S. performed the analysis.

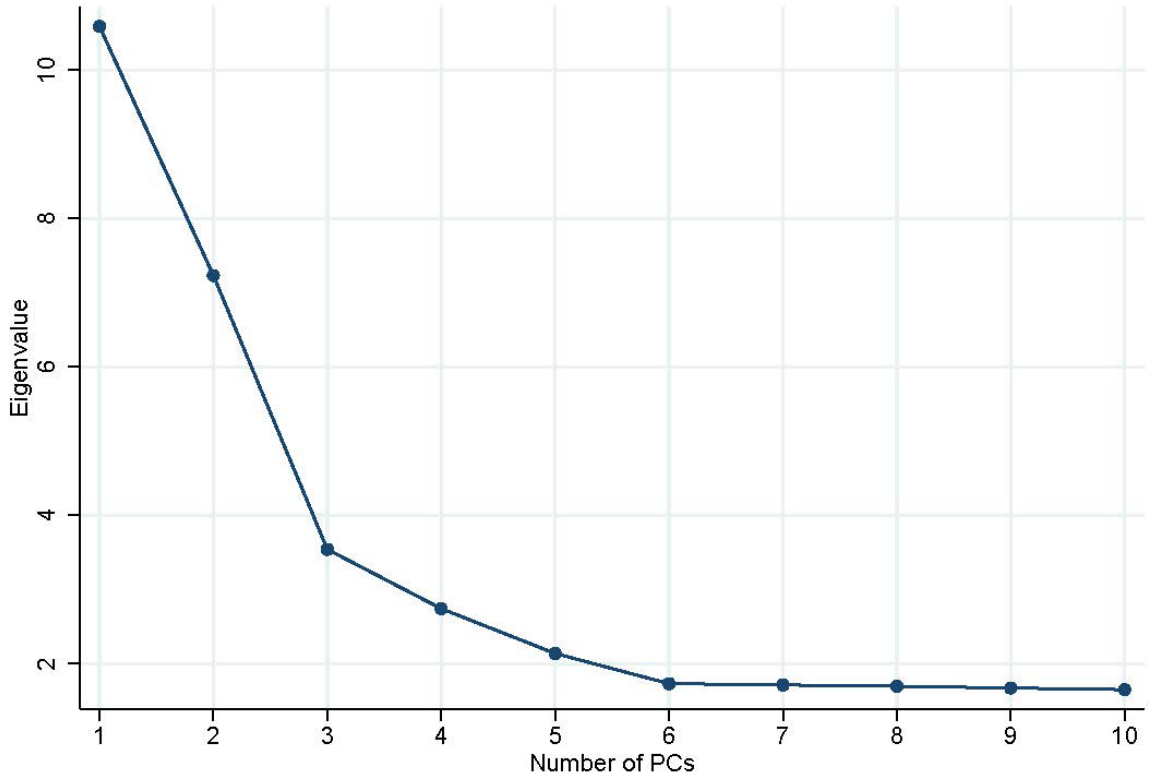
W.C.C., C.G.M. wrote the manuscript.

All authors critically reviewed and approved the manuscript.

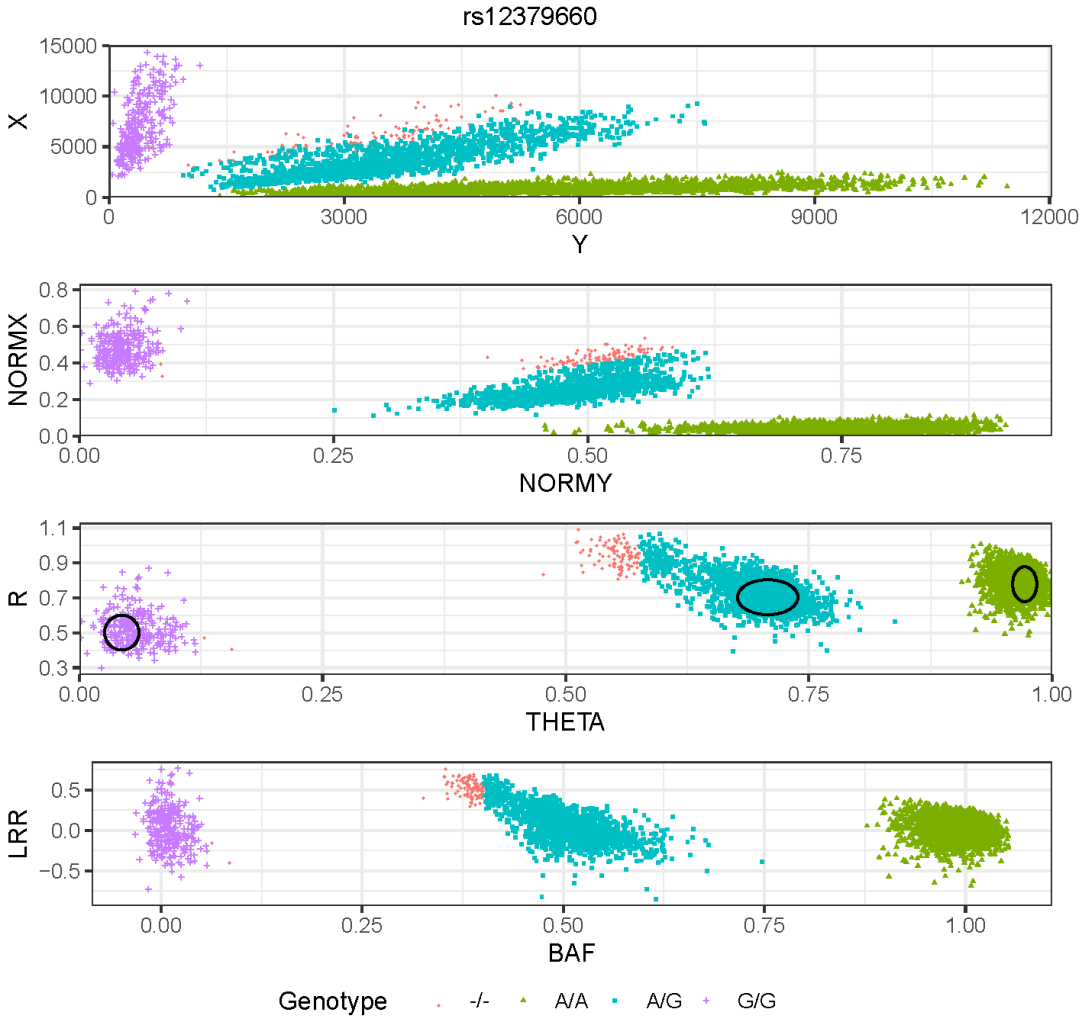
5.8. Competing interests

The authors declare no competing interests.

5.9. Supplementary



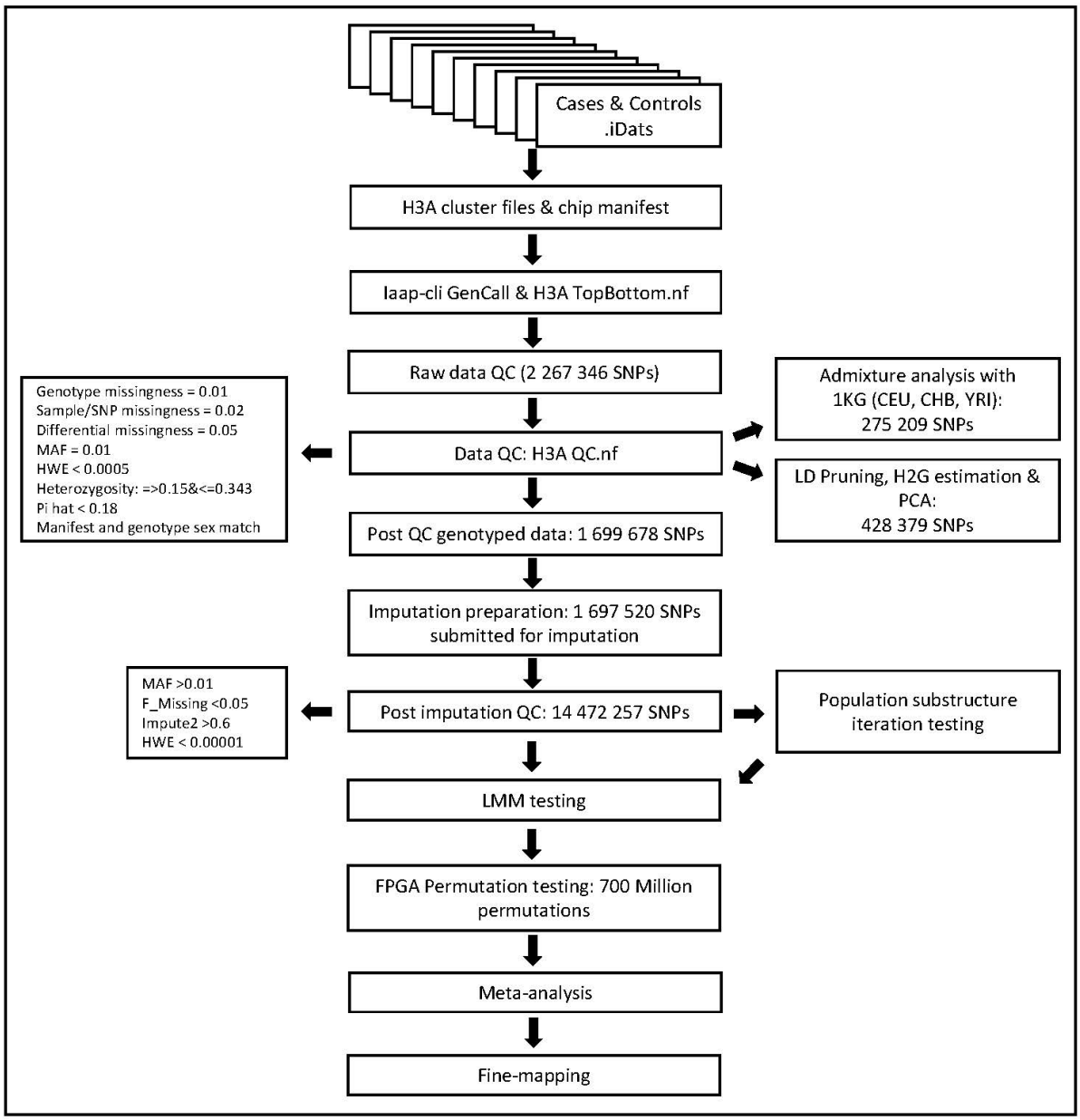
Supplementary figure 5.1. The eigenvalues for Principal component analysis of OSCC GWAS.



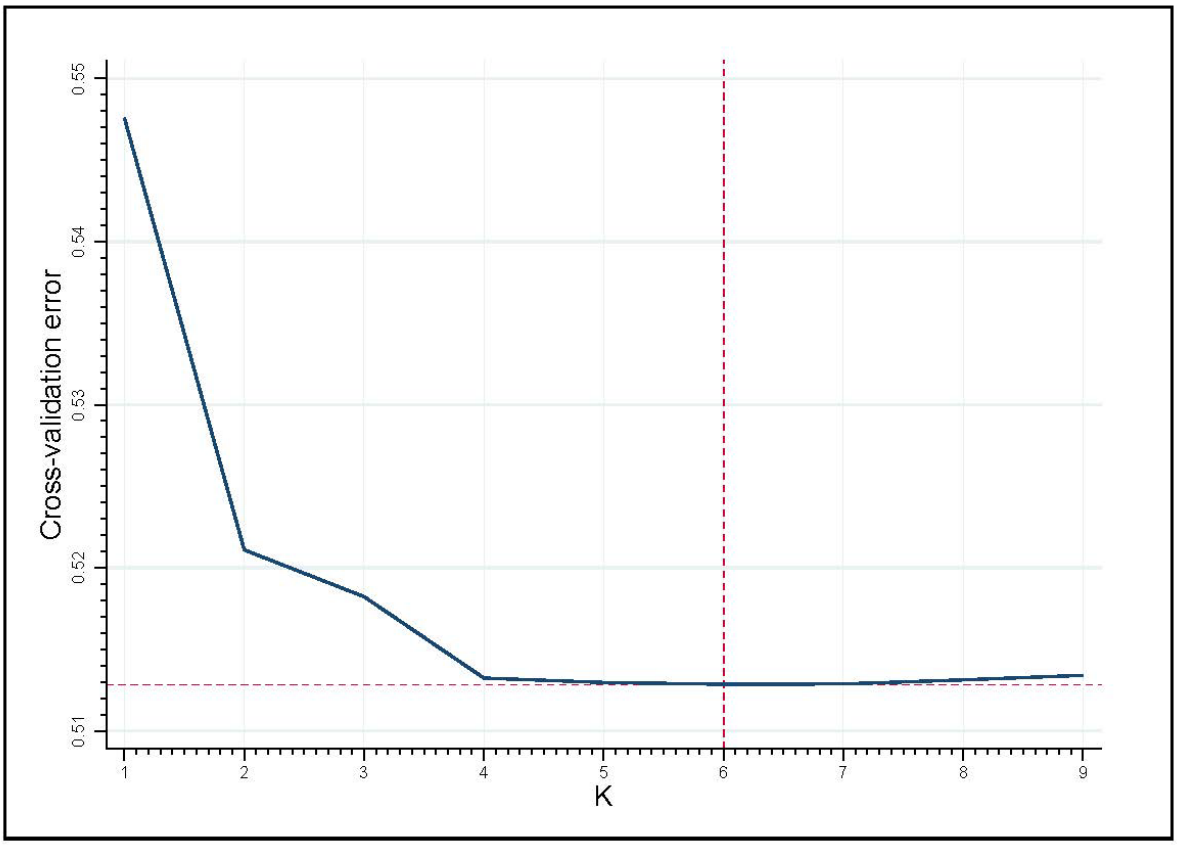
Legend:

X	- Raw X intensity as scanned from the array
Y	- Raw Y intensity as scanned from the array
NORMX	- Normalized X intensity
NORMY	- Normalized Y intensity
R	- Normalized R value
THETA	- Normalized Theta value
LRR	- Log R Ratio
BAF	- B Allele frequency

Supplementary figure 5.2. Cluster plot for rs12379660.



Supplementary figure 5.3. Study schematic and workflow outline of the African OSCC GWAS study.



Supplementary figure 5.4. Cross-validation errors for Admixture analysis, for K=1 to K=9. Optimal K=6, with the lowest cross-validation error of 0.51286.

Supplementary Table 5.1. OSCC GWAS study participants by case-control status.

Study Site	Number of Participants	Total
Population controls:		3217
Johannesburg Cancer Study - Soweto	487	
AWI-Gen - Soweto	1702	
AWI-Gen - Dikgale	1028	
OSCC cases:		1686
Johannesburg Cancer Study - Soweto	942	
WC OSCC - Cape Town	477	
KZN OSCC - Pietermaritzburg	267	
Total:		**4903**

Supplementary Table 5.2. Imputation score (Impute2 score) of all imputed SNPs in the African OSCC GWAS dataset.

Impute2 Score group	Number of SNPs	Percentage
1.00	80 340	0.63%
0.95-0.99	10 062 310	78.34%
0.90-0.94	1 777 621	13.84%
0.85-0.89	562 091	4.38%
0.80-0.84	208 065	1.62%
0.75-0.79	86 133	0.67%
0.70-0.74	38 780	0.30%
0.65-0.69	18 396	0.14%
0.60-0.64	10 258	0.08%
Total	12 843 994	

Supplementary Table 5.3. Allele frequencies of the top SNPs in Chromosome 2 from the OSCC GWAS and the 1KG.

Chr	Position (hg19)	rsID	Effect allele	Non-effect allele	Source	Effect allele frequency - Affected/Cases	Effect allele frequency - Controls/Population	Impute2 Score	P_{LMM} value
2	192207115	rs142741123	C	T	OSCC GWAS	0.0175	0.00715	0.935	$5.490x10^{-8}$
					1KG - Global		0.00440		
					1KG - African		0.01660		
					1KG - East Asian		0		
					1KG - Europe		0		
					1KG - South Asian		0		
					1KG - American		0		
2	192212930	rs141499985	T	G	OSCC GWAS	0.0175	0.00731	0.930	$6.330x10^{-8}$
					1KG - Global		0.00440		
					1KG - African		0.01660		
					1KG - East Asian		0		
					1KG - Europe		0		
					1KG - South Asian		0		
					1KG - American		0		

Supplementary Table 5.4. Linkage disequilibrium independent SNPs from 31 OSCC associated loci with P values < $5x10^{-6}$.

Chr	Position (hg19)	rsID	Effect allele	Non-effect allele	EAF - cases	EAF - controls	Genotyped SNP	Impute2 Score	Odds Ratio	Lower 95% CI	Upper 95% CI	P value (LMM)	P value (Adaptive permutation)
9	96182487	rs12379660	A	G	0.251	0.207	Yes	-	1.278	1.220	1.335	$4.58X10^{-8}$	$3.71X10^{-8}$
2	192207115	rs142741123	C	T	0.018	0.007		0.93	2.662	2.431	2.894	$5.49X10^{-8}$	$3.86X10^{-8}$
8	17801045	rs144188626	A	G	0.037	0.023		1.00	1.823	1.674	1.972	$2.21X10^{-7}$	$1.81X10^{-7}$
1	49055933	rs11205513	G	C	0.281	0.231		1.00	1.257	1.200	1.314	$2.57X10^{-7}$	$5.06X10^{-7}$
16	85580937	rs112854191	A	T	0.110	0.080		0.96	1.401	1.317	1.485	$2.66X10^{-7}$	$5.81X10^{-7}$
7	106320053	rs114448224	T	C	0.012	0.031	Yes	-	0.536	0.380	0.692	$2.91X10^{-7}$	$2.49X10^{-7}$
4	71368476	rs148725713	G	C	0.026	0.010		0.98	2.180	1.984	2.376	$3.38X10^{-7}$	$5.19X10^{-7}$
7	73155477	rs7806956	T	A	0.168	0.207		0.97	0.787	0.726	0.847	$3.46X10^{-7}$	$1.42X10^{-6}$
18	20439688	rs62093212	T	C	0.307	0.353		1.00	0.814	0.762	0.866	$4.42X10^{-7}$	$2.42X10^{-7}$
8	22982360	rs531841075	A	G	0.031	0.013		0.77	1.958	1.783	2.134	$8.53X10^{-7}$	$2.01X10^{-6}$
2	191917370	rs113702517	C	T	0.051	0.066		0.90	0.676	0.573	0.779	$9.37X10^{-7}$	$2.08X10^{-6}$
9	98106714	rs4744456	C	T	0.376	0.418		0.97	0.828	0.779	0.878	$1.14X10^{-6}$	$1.67X10^{-6}$
21	17628926	rs73365656	G	C	0.040	0.021		0.98	1.740	1.593	1.887	$1.30X10^{-6}$	$1.88X10^{-6}$
16	51057903	rs115523423	G	C	0.047	0.034		0.92	1.605	1.479	1.731	$1.44X10^{-6}$	$3.94X10^{-7}$
18	9374236	rs78048962	A	G	0.088	0.118		1.00	0.748	0.670	0.826	$1.75X10^{-6}$	$7.03X10^{-6}$
8	59378164	rs114039795	T	C	0.046	0.031		0.96	1.616	1.486	1.745	$1.97X10^{-6}$	$7.03X10^{-7}$
8	25539830	rs1365016	T	C	0.437	0.395	Yes	-	1.198	1.149	1.247	$2.20X10^{-6}$	$2.73X10^{-6}$
4	176439422	rs4532186	A	G	0.267	0.218	Yes	-	1.233	1.176	1.290	$2.76X10^{-6}$	$3.77X10^{-6}$
5	79551472	rs79319332	C	G	0.223	0.175		0.94	1.255	1.192	1.317	$2.98X10^{-6}$	$1.17X10^{-6}$
10	53313458	rs143380404	G	T	0.036	0.020		0.95	1.750	1.596	1.904	$3.02X10^{-6}$	$1.77X10^{-5}$
16	60045927	rs9783765	G	A	0.234	0.208		0.99	1.240	1.181	1.300	$3.10X10^{-6}$	$9.21X10^{-6}$
1	239870621	rs10802794	C	T	0.291	0.319	Yes	-	0.827	0.775	0.880	$3.11X10^{-6}$	$1.22X10^{-5}$
20	50496977	rs16996300	T	C	0.122	0.135	Yes	-	0.772	0.701	0.844	$3.13X10^{-6}$	$8.30X10^{-6}$
18	29144962	-	G	A	0.030	0.012		0.97	1.917	1.736	2.099	$3.99X10^{-6}$	$2.49X10^{-6}$
6	83695566	rs115648060	A	G	0.016	0.008		0.96	2.311	2.078	2.545	$4.05X10^{-6}$	$1.40X10^{-6}$
2	101053416	rs147274464	G	A	0.009	0.005		0.95	2.913	2.615	3.212	$4.26X10^{-6}$	$1.39X10^{-5}$
2	68310101	rs1868403	C	T	0.203	0.180	Yes	-	1.249	1.187	1.311	$4.34X10^{-6}$	$4.91X10^{-6}$
2	224373759	rs714943	T	C	0.324	0.383		0.98	0.836	0.785	0.886	$4.46X10^{-6}$	$1.15X10^{-5}$
15	45752892	rs73412484	G	A	0.244	0.208	Yes	-	1.239	1.179	1.299	$4.57X10^{-6}$	$1.56X10^{-6}$
19	28568669	rs10417850	A	G	0.020	0.033		0.98	0.592	0.445	0.739	$4.59X10^{-6}$	$1.76X10^{-5}$
2	230919172	rs35963873	G	T	0.027	0.017		0.96	1.833	1.663	2.003	$4.80X10^{-6}$	$5.67X10^{-6}$

Supplementary Table 5.5. Replication of known OSCC risk SNPs and loci.

SNPs present in the African OSCC GWAS and assessed for replication:				Previously published OSCC GWAS:						African OSCC GWAS:					
Chr	Position (hg19)	rsID	Gene (In/nearby)	Allele Source (Non-effect/Effect)	Effect allele freq - Cases	Effect allele freq - Controls	P value (literature)	Odds Ratio (95%CI)	Reference	Allele: African GWAS (Non-effect/Effect)	EAF: Case	EAF: Control	P value (African OSCC)	Odds Ratio (95%CI)	OR Directional Effects Comparison@
2	16680789	rs1430055	*FAM49A*	G/A	0.260	0.230	2.71×10^{-5}	1.18 (1.09-1.28)	Wu et al, 2012	G/A	0.319	0.348	0.7251	0.99 (0.93-1.04)	-
2	202162811	rs13016963	*ALS2CR12/CASP8*	A/G	0.676	0.717	4.29×10^{-13}	0.79 (0.75-0.84)	Wu et al, 2014	G/A	0.372	0.393	0.8074	1.01 (0.96-1.06)	+
2	202202791	rs10201587	*ALS2CR12*	G/A	-	-	8.71×10^{-10}	1.29 (1.19-1.40)	Abnet et al, 2012	A/G	0.380	0.373	0.0716	1.07 (1.02-1.12)	-
2	216318237	rs1437787	*MREG*	C/T	0.110	0.100	4.40×10^{-3}	1.17 (1.05-1.31)	Wu et al, 2012	T/C	0.203	0.221	0.1469	0.94 (0.88-0.99)	+
3	122730910	rs9868873	*SEMA5B*	G/A	-	-	1.00×10^{-7}	1.19 (1.11-1.27)	Wu et al, 2011	A/G	0.155	0.159	0.7233	0.98 (0.91-1.05)	+
3	175647165	rs6772209	*NAALADL2, AC104640.1*	G/A	-	-	3.00×10^{-7}	1.49 (1.28-1.75)	Wu et al 2011	G/A	0.113	0.118	0.5687	1.03 (0.96-1.11)	+
3	186793242	rs2239612	*ST6GAL1*	G/A	0.206	0.182	2.82×10^{-3}	1.12 (1.04-1.20)	Wu et al, 2014	G/A	0.374	0.361	0.2658	1.04 (0.99-1.09)	+
4	100124716	rs1893883	*ADH6*	G/C	0.180	0.140	3.61×10^{-8}	1.30 (1.18-1.43)	Wu et al, 2012	G/C	0.177	0.166	0.6426	1.02 (0.96-1.09)	+
4	100202414	rs1229977	*ADH1A*	C/T	0.140	0.110	1.52×10^{-7}	1.31 (1.19-1.45)	Wu et al, 2012	C/T	0.088	0.069	0.3078	1.08 (0.98-1.17)	+
4	100228466	rs1042026	*ADH1B*	T/C	0.280	0.230	1.33×10^{-10}	1.29 (1.19-1.39)	Abnet et al, 2010	T/C	0.073	0.086	0.6595	0.97 (0.88-1.06)	-
4	100228945	rs17033	*ADH1B*	T/C	0.150	0.120	1.83×10^{-9}	1.36 (1.23-1.50)	Wu et al, 2012	T/C	0.046	0.042	0.6276	1.05 (0.93-1.17)	+
4	100258155	#rs1614972	*ADH1C*	T/C	0.300	0.250	1.62×10^{-9}	1.26 (1.17-1.36)	Wu et al, 2012	C/T	0.390	0.436	0.0113	0.91 (0.86-0.96)	+
4	100260789	rs698	*ADH1C*	A/G	-	-	3.00×10^{-7}	1.12 (1.07-1.17)	McKay et al, 2011	T/C	0.131	0.115	0.1870	1.08 (1.00-1.15)	+
4	100262041	rs1789903	*ADH1C*	C/G	0.120	0.090	1.04×10^{-6}	1.31 (1.18-1.47)	Wu et al, 2012	C/G	0.131	0.115	0.1870	1.08 (1.00-1.15)	+
4	100322786	rs17028973	*ADH7*	C/T	0.370	0.330	2.48×10^{-8}	1.22 (1.14-1.31)	Wu et al, 2012	T/C	0.303	0.295	0.2513	1.05 (1.00-1.10)	-
5	58407771	rs10052657	*PDE4D*	C/A	0.101	0.120	1.15×10^{-6}	0.80 (0.73-0.87)	Wu et al, 2014	C/A	0.136	0.138	0.7690	1.02 (0.95-1.09)	-
5	138861146	rs7447927	*TMEM173*	C/G	-	-	7.72×10^{-20}	0.85 (0.82-0.88)	Wu et al, 2014	C/G	0.047	0.059	0.6317	0.96 (0.85-1.07)	+
5	148904092	rs10058728	*CSNK1A1*	T/A	-	-	5.00×10^{-9}	2.04 (1.61-2.56)	Wu et al, 2011	T/A	0.163	0.182	0.6382	0.98 (0.91-1.04)	-
6	31112737	#rs130079	*CCHCR1*	C/A	0.110	0.080	7.77×10^{-24}	1.39 (1.31-1.49)	Chang et al, 2018	C/A	0.321	0.287	0.0374	1.09 (1.04-1.15)	+
6	31540784	rs1041981	*LTA*	C/A	0.420	0.470	9.37×10^{-22}	0.82 (0.80-0.86)	Chang et al, 2018	A/C	0.378	0.407	0.2737	0.96 (0.91-1.01)	-
6	40802261	rs10484761	*UNC5CL, AL583854.1*	T/C	-	-	7.00×10^{-12}	1.33 (1.23-1.45)	Wu et al, 2011	C/T	0.456	0.466	0.6948	0.99 (0.94-1.03)	+
8	11281273	rs2409764	*FAM167A*	A/G	0.050	0.040	8.00×10^{-4}	1.32 (1.12-1.55)	Wu et al, 2012	A/G	0.490	0.481	0.4948	1.03 (0.98-1.07)	+
8	133119961	rs7822239	*HHLA1*	C/A	-	-	8.00×10^{-6}	1.10 (1.05-1.14)	Wu et al, 2014	C/A	0.149	0.135	0.3783	1.05 (0.98-1.12)	+
10	96066341	rs2274223	*PLCE1*	A/G	0.264	0.209	3.65×10^{-18}	1.35 (1.26-1.45)	Wu et al, 2014	A/G	0.421	0.404	0.6548	1.02 (0.97-1.07)	+
12	112521448	rs4767364	*ALDH2*	G/A	-	-	2.00×10^{-8}	1.13	McKay et al, 2011	A/G	0.115	0.121	0.4977	1.04	-

15	69059084	rs8030672	*ANP32A, CARS1P1*	T/A	-	-	1.00×10^{-11}	(1.08-1.18) 1.56 (1.37-1.78)	Wu et al, 2011	T/A	0.195	0.218	0.6559	(0.96-1.12) 0.98 (0.92-1.04)	-
16	50103734	rs4785204	*HEATR3*	C/T	0.297	0.265	1.83×10^{-5}	1.15 (1.08-1.23)	Wu et al, 2014	C/T	0.090	0.083	0.4417	1.05 (0.97-1.14)	+
16	50148508	rs7206735	*PAPD5*	T/C	0.330	0.290	6.97×10^{-7}	1.21 (1.12-1.30)	Wu et al, 2012	C/T	0.251	0.271	0.2141	0.95 (0.89-1.00)	+
17	7557834	rs1642764	*ATP1B2, TP53*	C/T	-	-	3.10×10^{-13}	0.88 (0.85-0.91)	Wu et al, 2014	C/T	0.191	0.186	0.4619	0.97 (0.90-1.03)	+
21	36357861	rs2014300	*RUNX1*	A/G	0.871	0.843	7.91×10^{-9}	1.28 (1.18-1.39)	Wu et al, 2014	A/G	0.362	0.371	0.0999	0.94 (0.89-0.99)	+
22	29115066	rs4822983	*CHEK2*	C/T	-	-	2.00×10^{-22}	1.27 (1.21-1.34)	Wu et al, 2012	C/T	0.457	0.435	0.1872	1.05 (1.00-1.10)	+
22	29130300	#rs1033667	*CHEK2*	C/T	-	-	1.91×10^{-8}	1.33 (1.20-1.46)	Wu et al, 2014	C/T	0.442	0.397	0.0055	1.11 (1.06-1.16)	+
22	29192670	rs2239815	*CHEK2/ XPB1*	T/C	0.567	0.613	4.09×10^{-7}	0.86 (0.81-0.91)	Wu et al, 2014	C/T	0.182	0.165	0.7024	1.02 (0.95-1.09)	+

@ - "+" Denotes the same direction of effect; "-" denotes the opposite direction of effect.
# - Moderate level of statistical significance: P < 0.05.

SNPs not present in the African OSCC GWAS:				**Previously published OSCC GWAS:**					
Chr	**Position (hg19)**	**rsID**	**Gene (In/nearby)**	**Allele Source (Non-effect/Effect)**	**Effect allele freq - Cases**	**Effect allele freq - Controls**	**P value (literature)**	**Odds Ratio (95%CI)**	**Reference**
2	72360331	rs138478634	*CYP26B1*	G/A	0.030	0.010	9.45×10^{-17}	1.82 (1.58-2.09)	Chang et al, 2018
4	84374480	rs1494961	*HELQ*	T/C	-	-	1.00×10^{-8}	1.12 (1.08-1.17)	McKay et al, 2011
4	100056998	rs3805322	*ADH4*	A/G	0.440	0.490	2.83×10^{-13}	0.80 (0.75-0.86)	Wu et al, 2012
4	100239319	rs1229984	*ADH1B*	C/T	-	-	1.00×10^{-20}	0.64 (0.59-0.71)	McKay et al, 2011
4	100349669	rs1573496	*ADH7*	C/G	-	-	9.00×10^{-17}	0.75 (0.70-0.80)	McKay et al, 2011
6	32056580	rs204900	*TNXB*	A/C	0.140	0.110	3.16×10^{-23}	1.34 (1.26-1.42)	Chang et al 2018
6	32589266	rs35597309	*HLA class II genes*	G/A	-	-	1.18×10^{-7}	1.19 (1.12-1.27)	Wu et al, 2014
9	232803326	rs61271688	*CDKN2B-AS1*	T/C	-	-	7.00×10^{-7}	1.11 (1.07-1.16)	Wu et al, 2014
12	112168009	rs11066015	*ACAD10*	G/A	-	-	7.00×10^{-21}	1.38 (1.29-1.48)	Wu et al, 2011
12	112645401	rs2074356	*HECTD4*	C/T	-	-	2.00×10^{-31}	1.56 (1.45-1.68)	Wu et al, 2011
12	112817783	rs11066280	*RPL6*	T/A	-	-	6.45×10^{-11}	1.22 (1.09-1.37)	Wu et al, 2011
12	112817783	rs11066280	*HECTD4*	T/A	-	-	2.00×10^{-15}	1.3 (1.22-1.38)	Wu et al, 2011
17	7579548	rs1800371	*TP53*	C/T	-	-	-	-	Jennis et al, 2016
17	80040513	rs17848945	*FASN*	G/A	0.020	0.010	1.49×10^{-11}	1.67 (1.44-1.94)	Chang et al 2018
22	29130012	rs738722	*CHEK2*	C/T	0.308	0.254	5.67×10^{-8}	1.32 (1.19-1.45)	Abnet et al, 2011
22	31010417	rs117353193	*TCN2*	G/A	0.050	0.030	1.74×10^{-23}	1.67 (1.51-1.84)	Chang et al 2018

Supplementary Table 5.6. eQTL mapping for the top GWAS hit on Chromosome 9.

Chr	Position (hg19)	Database@	Tissue type	Gene	Gene symbol	Tested allele	Risk increasing allele	P-value of eQTLs	Signed test statistics	FDR of eQTLs	Aligned direction#
9	96182487$	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	A	A	$5.74x10^{-7}$	0.1391	$3.05x10^{-4}$	+
9	96181075	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	T	T	$5.74x10^{-7}$	0.1391	$3.05x10^{-4}$	+
9	96161300	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	T	T	$3.58x10^{-5}$	0.1093	$3.05x10^{-4}$	+
9	96181075	GTEx/v8	Esophagus_Muscularis	ENSG00000188938	*FAM120AOS*	T	T	$4.19x10^{-5}$	0.1073	$2.03x10^{-4}$	+
9	96182487	GTEx/v8	Esophagus_Muscularis	ENSG00000188938	*FAM120AOS*	A	A	$4.19x10^{-5}$	0.1073	$2.03x10^{-4}$	+
9	96178967	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	A	A	$1.21x10^{-4}$	0.1015	$3.05x10^{-4}$	+
9	96172790	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	T	T	$1.21x10^{-4}$	0.1015	$3.05x10^{-4}$	+
9	96177221	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	A	A	$1.21x10^{-4}$	0.1015	$3.05x10^{-4}$	+
9	96173044	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	T	T	$1.21x10^{-4}$	0.1015	$3.05x10^{-4}$	+
9	96179785	GTEx/v8	Esophagus_Mucosa	ENSG00000188938	*FAM120AOS*	A	A	$1.47x10^{-4}$	0.0998	$3.05x10^{-4}$	+

@ - Data source: eQTL data was downloaded from http://www.gtexportal.org/home/datasets. Under the section of GTEx V8, from single tissue eQTL data both GTEx_Analysis_v8_eQTL.tar for significant SNP-gene associations, and all tested pairs of SNP-gene were obtained from GCP (including non-significant pairs). GTEx eQTL v8 contains 54 different tissue types across 30 general tissue types.
# - The direction of effect to gene expression after aligning risk increasing allele of GWAS and tested allele of eQTLs. "+" when the risk-increasing allele increases the expression of the gene, "-" when the risk-increasing allele decreases the expression of the gene.
$ - Top hit in GWAS: rs12379660.

Supplementary Table 5.7. eQTL mapping for the Meta-analysis top hits on Chromosomes 10 and 22.

Chr	Position (hg19)	Database	Tissue type	Gene	Gene symbol	P-value of eQTLs
10	96065694	GTEx/v8	Esophagus - Muscularis	ENSG00000173145.11	*NOC3L*	$3.30x10^{-9}$
22	29130300	GTEx/v8	Esophagus - Muscularis	ENSG00000100154.14	*TTC28*	$7.30x10^{-7}$
22	29130300	GTEx/v8	Esophagus - Mucosa	ENSG00000100209.9	*HSCB*	$9.10x10^{-7}$

@ - Data source: eQTL data was downloaded from http://www.gtexportal.org/home/datasets. Under the section of GTEx V8, from single tissue eQTL data both GTEx_Analysis_v8_eQTL.tar for significant SNP-gene associations, and all tested pairs of SNP-gene were obtained from GCP (including non-significant pairs). GTEx eQTL v8 contains 54 different tissue types across 30 general tissue types.

Supplementary Table 5.8. Data quality control parameters.

Quality control parameters	Thresholds - Inclusion
Genotype data QC:	
SNP selection	Autosomal SNPs only
Individual missingness	<=0.02
SNP missingness	<=0.01
Minor allele frequency	>=0.01
Hardy Weinberg Equilibrium	P>=0.0005
Manifest-genotype sex match	Matched only
Relatedness	pi_hat<=0.18
Heterozygosity	<=0.343 & >=0.15
Post-imputation QC:	
SNP missingness	<0.05
Minor allele frequency	>0.01
Impute2 Score	>=0.6
Hardy Weinberg Equilibrium	p>=0.00001

5.10. References:

Abnet, C.C., Freedman, N.D., Hu, N., et al. (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics*. 42(9):764–767. DOI: 10.1038/ng.649.

Abnet, C.C., Arnold, M. & Wei, W.-Q. (2018) Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology*. 154(2):360–373. DOI: 10.1053/j.gastro.2017.08.023.

Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*. 19(9):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(sup2):1507133. DOI: 10.1080/16549716.2018.1507133.

Arnold, M., Soerjomataram, I., Ferlay, J., et al. (2015) Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*. 64(3):381–387. DOI: 10.1136/gutjnl-2014-308124.

Auton, A., Abecasis, G.R., Altshuler, D.M., et al. (2015) A global reference for human genetic variation. *Nature*. 526(7571):68–74. DOI: 10.1038/nature15393.

Bye, H., Prescott, N.J., Matejcic, M., et al. (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*. 32(12):1855–1861. DOI: 10.1093/carcin/bgr211.

Bye, H., Prescott, N.J., Lewis, C.M., et al. (2012) Distinct genetic association at the *PLCE1* locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis*.

33(11):2155–2161. DOI: 10.1093/carcin/bgs262.

Capowski, E.E., Esnault, S., Bhattacharya, S., et al. (2001) Y box-binding factor promotes eosinophil survival by stabilizing granulocyte-macrophage colony-stimulating factor mRNA. *Journal of immunology (Baltimore, Md. : 1950)*. 167(10):5970–6. DOI: 10.4049/jimmunol.167.10.5970.

Cerami, E., Gao, J., Dogrusoz, U., et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*. 2(5):401–4. DOI: 10.1158/2159-8290.CD-12-0095.

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

Chang, J., Zhong, R., Tian, J., et al. (2018) Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics*. 50(3):338–343. DOI: 10.1038/s41588-018-0045-8.

Chapman, B. V, Wald, A.I., Akhtar, P., et al. (2015) MicroRNA-363 targets myosin 1B to reduce cellular migration in head and neck cancer. *BMC cancer*. 15:861. DOI: 10.1186/s12885-015-1888-3.

Chattopadhyay, R., Das, S., Maiti, A.K., et al. (2008) Regulatory role of human AP-endonuclease (APE1/Ref-1) in YB-1-mediated activation of the multidrug resistance gene MDR1. *Molecular and cellular biology*. 28(23):7066–80. DOI: 10.1128/MCB.00244-08.

Chen, C.Y., Gherzi, R., Andersen, J.S., et al. (2000) Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. *Genes & development*. 14(10):1236–48. Available: http://www.ncbi.nlm.nih.gov/pubmed/10817758.

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):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):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(March):101701. DOI: 10.1016/j.canep.2020.101701.

Chen, X., Li, A., Sun, B.-F., et al. (2019) 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nature cell biology*. 21(8):978–990. DOI: 10.1038/s41556-019-0361-y.

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

Choudhury, A., Sengupta, D., Ramsay, M., et al. (2021) Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*. 30(R1):R56–R63. DOI: 10.1093/hmg/ddaa274.

Cui, R., Kamatani, Y., Takahashi, A., et al. (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*. 137(5):1768–1775. DOI: 10.1053/j.gastro.2009.07.070.

Enzinger, P.C. & Mayer, R.J. (2003) Esophageal Cancer. *New England Journal of Medicine*. 349(23):2241–2252. DOI: 10.1056/NEJMra035010.

Fabregat, A., Sidiropoulos, K., Viteri, G., et al. (2017) Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*. 18(1):142. DOI: 10.1186/s12859-017-1559-2.

Forman, D., Bray, F., Brewster, D.H., et al. (2013) *Cancer Incidence in Five Continents, Vol. X*. Lyon: IARC Scientific Publications, No 164.

Gaudreault, I., Guay, D. & Lebel, M. (2004) YB-1 promotes strand separation in vitro of duplex DNA containing either mispaired bases or cisplatin modifications, exhibits endonucleolytic activities and binds several DNA repair proteins. *Nucleic acids research*. 32(1):316–27. DOI: 10.1093/nar/gkh170.

Griss, J., Viteri, G., Sidiropoulos, K., et al. (2020) ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Molecular & cellular proteomics : MCP*. 19(12):2115–2125. DOI: 10.1074/mcp.TIR120.002155.

GTEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. *Nature genetics*. 45(6):580–5. DOI: 10.1038/ng.2653.

Horwitz, E.M., Maloney, K.A. & Ley, T.J. (1994) A human protein containing a "cold shock" domain binds specifically to H-DNA upstream from the human gamma-globin genes. *The Journal of biological chemistry*. 269(19):14130–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/8188694.

Hyland, P.L., Zhang, H., Yang, Q., et al. (2016) Pathway, in silico and tissue-specific expression quantitative analyses of oesophageal squamous cell carcinoma genome-wide association studies data. *International journal of epidemiology*. 45(1):206–220. DOI: 10.1093/ije/dyv294.

Isaacson, C., Mothobi, P., Hale, M., et al. (2015) Carcinogenic nitrosamines in traditional beer as the cause of oesophageal squamous cell carcinoma in black South Africans. *South African Medical Journal*. 105(8):656. DOI: 10.7196/SAMJnew.7935.

Jassal, B., Matthews, L., Viteri, G., et al. (2020) The reactome pathway knowledgebase. *Nucleic acids research*. 48(D1):D498–D503. DOI: 10.1093/nar/gkz1031.

Kamangar, F., Chow, W.-H., Abnet, C.C., et al. (2009) Environmental causes of esophageal cancer. *Gastroenterology clinics of North America*. 38(1):27–57, vii. DOI: 10.1016/j.gtc.2009.01.004.

Kayamba, V., Bateman, A.C., Asombang, A.W., et al. (2015) HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case–control study. *Cancer Medicine*. 4(4):588–595. DOI: 10.1002/cam4.434.

Kumar, S.K., Feldman, M.W., Rehkopf, D.H., et al. (2015) Limitations of GCTA as a solution to the missing heritability problem. *PNAS*. 113(1):E61-E70. DOI: 10.1073/pnas.1520109113

van Loon, K., Mwachiro, M.M., Abnet, C.C., et al. (2018) The African Esophageal Cancer Consortium: A Call to Action. *Journal of Global Oncology*. 2018(4):1–9. DOI: 10.1200/JGO.17.00163.

López, S., van Dorp, L. & Hellenthal, G. (2015) Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary bioinformatics online*. 11(Suppl 2):57–68. DOI: 10.4137/EBO.S33489.

Matejcic, M., Li, D., Prescott, N.J., et al. (2011) Association of a Deletion of *GSTT2B* with an Altered Risk of Oesophageal Squamous Cell Carcinoma in a South African Population: A Case-Control Study. *PLoS ONE*. 6(12):e29366. DOI: 10.1371/journal.pone.0029366.

Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nature Reviews Genetics*. 9(1):9–14. DOI: 10.1038/nrg2203.

McKay, J.D., Truong, T., Gaborieau, V., et al. (2011) A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. *PLoS Genetics*. 7(3):e1001333. DOI: 10.1371/journal.pgen.1001333.

Menya, D., Maina, S.K., Kibosia, C., et al. (2019) Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case-control study and a pan-African perspective. *International journal of cancer*. 145(1):99–109. DOI: 10.1002/ijc.32086.

Menya, D., Kigen, N., Oduor, M., et al. (2019) Traditional and commercial alcohols and esophageal cancer risk in Kenya. *International journal of cancer*. 144(3):459–469. DOI: 10.1002/ijc.31804.

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

Mulder, N., Abimiku, A., Adebamowo, S.N., et al. (2018) H3Africa: current perspectives. *Pharmacogenomics and personalized medicine*. 11:59–66. DOI: 10.2147/PGPM.S141546.

Murphy, G., McCormack, V., Abedi-Ardekani, B., et al. (2017) International cancer seminars: a focus on esophageal squamous cell carcinoma. *Annals of Oncology*. 28(9):2086–2093. DOI: 10.1093/annonc/mdx279.

Naeger, L.K., McKinney, J., Salvekar, A., et al. (1999) Identification of a *STAT4* binding site in the interleukin-12 receptor required for signaling. *The Journal of biological chemistry*. 274(4):1875–8. DOI: 10.1074/jbc.274.4.1875.

Ohmura, G., Tsujikawa, T., Yaguchi, T., et al. (2015) Aberrant Myosin 1b Expression Promotes Cell

Migration and Lymph Node Metastasis of HNSCC. *Molecular cancer research : MCR*. 13(4):721–31. DOI: 10.1158/1541-7786.MCR-14-0410.

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):1751–6. DOI: 10.1038/sj.bjc.6600338.

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

Ramsay, M., Crowther, N., Tambo, E., et al. (2016) 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*. 1:e20. DOI: 10.1017/gheg.2016.17.

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):djv279. DOI: 10.1093/jnci/djv279.

Schäfer, G., Kabanda, S., van Rooyen, B., et al. (2013) The role of inflammation in HPV infection of the Oesophagus. *BMC Cancer*. 13(1):185. DOI: 10.1186/1471-2407-13-185.

Segal, I., Reinach, S. & de Beer, M., (1988) Factors associated with oesophageal cancer in Soweto, South Africa. *British Journal of Cancer*. 58(5):681–686. DOI: 10.1038/bjc.1988.286.

Sengupta, D., Choudhury, A., Fortes-Lima, C., et al. (2021) Genetic substructure and complex demographic history of South African Bantu speakers. *Nature Communications*. 12(1):2080. DOI: 10.1038/s41467-021-22207-y.

Simba, H., Kuivaniemi, H., Lutje, V., et al. (2019) Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in Genetics*. DOI: 10.3389/fgene.2019.00642.

Sitas, F., Egger, S., Urban, M.I., et al. (2012) InterSCOPE study: Associations between esophageal squamous cell carcinoma and human papillomavirus serological markers. *Journal of the National Cancer Institute*. 104(2):147–58. DOI: 10.1093/jnci/djr499.

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):486. DOI: 10.7196/SAMJ.2017.v107i6.12099.

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

Statistics South Africa. (2006) Mid-year population estimates, South Africa 2005. 27(12).

Statistics South Africa. (2010) Statistical release Mid-year population estimates 2010. *Stats SA*. (July):16.

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):209–249. DOI: 10.3322/caac.21660.

Swiel, Y., Brandenburg, J.T., Hayat, M., et al. (2022) FPGA Acceleration of GWAS Permutation Testing. *bioRXiv*, preprint. DOI: 10.1101/2022.03.11.483235.

Thumbs, A., Borgstein, E., Vigna, L., et al. (2012) Self-expanding metal stents (SEMS) for patients with advanced Esophageal cancer in Malawi: An effective palliative treatment. *Journal of Surgical Oncology*. 105(4):410–414. DOI: 10.1002/jso.23003.

Wang, L.-D., Zhou, F.-Y., Li, X.-M., et al. (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at *PLCE1*. *Nature Genetics*. 42(9):759–763. DOI: 10.1038/ng.648.

Watanabe, K., Taskesen, E., van Bochoven, A., (2017) Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*. 8(1):1826. DOI: 10.1038/s41467-017-01261-5.

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):700–712. DOI: 10.1093/bib/bbw145.

White, R.E., Parker, R.K., Fitzwater, J.W., et al. (2009) Stents as sole therapy for oesophageal cancer: a prospective analysis of outcomes after placement. *The Lancet Oncology*. 10(3):240–246. DOI: 10.1016/S1470-2045(09)70004-X.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available: https://ggplot2.tidyverse.org.

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

Wu, C., Hu, Z., He, Z., et al. (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature genetics*. 43(7):679–684. DOI: 10.1038/ng.849.

Wu, C., Kraft, P., Zhai, K., et al. (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nature genetics*. 44(10):1090–1097. DOI: 10.1038/ng.2411.

Wu, C., Wang, Z.Z., Song, X., et al. (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nature genetics*. 46(9):1001–1006.

DOI: 10.1038/ng.3064.

Zhang, H.-R., Lai, S.-Y., Huang, L.-J., et al. (2018) Myosin 1b promotes cell proliferation, migration, and invasion in cervical cancer. *Gynecologic oncology*. 149(1):188–197. DOI: 10.1016/j.ygyno.2018.01.024.

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

6. CHAPTER 6: Discussion and Conclusions

6.1. Discussion

Our study set out to evaluate the contributions of genetic risk factors to the development of oesophageal squamous cell carcinoma in the South African black population and to identify potentially shared or population-specific risk loci for OSCC. First, we established that the JCS samples that had undergone long-term storage at -30°C were suitable for use in genomic studies. Then we attempted to replicate previous findings and assess whether OSCC risk loci identified in other populations transfer to our South African Black population. Thereafter, through a genome-wide association study, we sought to identify population-specific or shared risk loci for OSCC in populations of diverse ancestry.

The genetic contribution to OSCC disease risk is not well understood. The environmental risk factors linked to OSCC, thus far, are only able to explain a limited proportion of the disease burden (Kamangar et al., 2009; McCormack et al., 2017; Abnet, Arnold & Wei, 2018). The unique distribution of OSCC across distant geographical locations suggests possible shared disease aetiology among populations of diverse ancestry. Previous large-scale genomic studies, conducted in Asian and European populations, have identified robust genetic risk factors for OSCC through the use of GWAS in their respective populations (Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014). To date, most genetic studies investigating the genetic aetiology of OSCC in African populations are small-scale, genetic replication studies or candidate gene studies (reviewed in Simba et al., 2019). Our study is the first GWAS of OSCC in an African black population.

<u>Objective 1: Assessing the quality of DNA samples</u>

Good quality of sufficient quantity is needed to conduct large-scale genetic studies This is highlighted by the importance of the establishment of biobanks and sample repositories to enable such research to take place (Hawkins, 2010; Soo et al., 2017; Coppola et al., 2019). The Johannesburg Cancer Study was established in 1996 (Chen et al., 2020) at a time when technologies for large-scale genomic and genetic studies did not exist. The JCS however had the foresight to biobank valuable biomaterials, from treatment naïve cancer patients to allow for research in the future when new technologies and approaches might

become available. This led to our first objective: "To assess whether the quality and quantity of genomic DNA isolated from archival blood samples are suitable for large-scale genomic assays."

The JCS banked biosamples for over 20 years. This study extracted gDNA from 2,758 archival peripheral blood samples stored for 2 to 19 years at -30°C (Chen et al., 2018) and assessed the quality and quantity of genomic DNA isolated from the blood samples. Our study found that the long-term storage of blood samples at -30°C did not negatively impact the quality and quantity of genomic DNA isolated from the samples. Although previous findings suggested that increased storage duration or increasing age of participants resulted in a reduced DNA yield (Erkeller-Yuksel et al., 1992; AlRokayan, 2000), this was not observed in the JCS. Factors that impacted DNA yield were the volume of blood available for DNA isolation and the HIV status of the patient, which reduced the quantity of peripheral blood mononuclear cells for DNA isolation.

One limitation of our study was that we were not able to perform gel electrophoresis on the entire sample of 2,758 DNA samples. However, the representative subset of samples assessed on gel electrophoresis (n=200) indicated high molecular weight DNA was present in all samples. This study provided strong support for the viability of archival samples banked by the JCS for use in large-scale genetic studies. This has since been confirmed by the high genotyping call rates in our GWAS (97.83% genotyping rate at raw data level, 99.92% genotyping rate at post genotype QC). It has also unlocked the potential for the use of peripheral blood samples from >20,000 cancer patients collected by the JCS over the last 20 years and has led to further GWAS being conducted for breast and cervical cancer.

Objective 2: Replication of known OSCC genetic risk factors

The JCS contributed a sizable portion of the OSCC cases and controls to enable large-scale genetic studies. Together with participants recruited by collaborators at UCT, our group amassed study samples of 1,471 OSCC cases and 1,791 ethnically matched population controls (Chen, Bye, et al., 2019). This enabled us to investigate 7 susceptibility loci previously reported to be associated with OSCC in non-African populations (Abnet et al.,

2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014), for association with OSCC in the SAB population.

In this candidate gene study, we genotyped SNPs from the 7 susceptibility loci in the OSCC study samples from JCS and UCT separately. This was followed by logistical regression modelling of the two study samples separately, and the results were then meta-analysed. Only one SNP (*CHEK2* rs1033667) was significantly associated with OSCC in the SAB population in the meta-analysis of the two study samples (P = 0.002, OR = 1.18 (1.06 – 1.30)). The *PLCE1* locus is the strongest risk-associated locus for Asian OSCC known to date (Wu et al., 2014). No signals of association were detected for the *PLCE1* locus in African OSCC (rs2274223, P = 0.708, OR = 0.98 (0.88 – 1.09)). A key finding from our candidate gene study was the association signal detected in the *CHEK2* gene, which is suggestive of shared disease aetiology between the Chinese and African populations for OSCC. *CHEK2* encodes the protein CHK2, a serine-threonine kinase that plays a key role in the DNA damage response pathway (Zannini, Delia & Buscemi, 2014). Rare mutations in *CHEK2* have been implicated in susceptibility to other cancers such as breast and prostate cancer (Southey et al., 2016), thus it is plausible that *CHEK2* may also confer susceptibility to OSCC.

Both rare and relatively rare variants in *CHEK2* have been associated with several other cancers, including breast, prostate, stomach and kidney cancers (Southey et al 2016; Näslund-Koch et al, 2016) so it is possible that variants in *CHEK2* may also confer susceptibility to OSCC. The results from our trans-ethnic meta-analysis discussed below showing the association of variants in *CHEK2* at genome-wide significance strongly supports a role for *CHEK2* in OSCC susceptibility and suggests that a further detailed study of *CHEK2* in OSCC is warranted using sequencing and fine-mapping approaches in larger sample sizes.

The other key finding was the lack of replication or transferability of other strong risk loci from Asian populations to African populations, particularly for *PLCE1*. This suggests that although the disease is similar phenotypically, there may be population-specific genetic risk factors that arose after the migration of humans out of Africa and are therefore not present in African populations. An alternative explanation for the lack of replication is

that the genotyped SNPs do not tag the causal variants or genes effectively in African populations. Linkage disequilibrium in African populations is generally lower than non-African populations, with short haplotype blocks (Teo, Small & Kwiatkowski, 2010). Thus, the lead SNPs in the Asian populations may be in high LD with the causal variants but not in African populations, which would weaken the association signal.

A limitation of our candidate gene study was that there were no effective means to control for population substructures that might be present within the study samples. Recent genomic studies illustrated the complexity of African genomes (Choudhury et al., 2020, 2021; Sengupta et al., 2021). Rigorous means to control and adjust for population substructure present in African genomes is crucial to ensure the validity of the findings. Ancestry informative markers (AIMS) are useful to detect the presence of population substructure but insufficient to correct for substantial substructure in genetically complex populations. Thus, instead of candidate gene studies or targeted genotyping replication studies, a broader, genome-wide study is likely to be a more fruitful approach.

The fixed-effect model meta-analysis assumes all the studies share the same effect. The random-effect model meta-analysis assumes each study estimates a different underlying true effect that usually follows a normal distribution. Given the known ethnolinguistic differences between the black population in Soweto and the black population in the Western Cape, we felt it was important to test under the random-effect model as well since we cannot assume the underlying cause is the same in both studies.

Objectives 3 & 4: Novel OSCC genetic risk factor discovery and transethnic meta-analysis

Genome-wide association studies are a powerful tool for the identification of genetic risk factors associated with the phenotypes of interest. GWAS have yielded a rich harvest of novel findings in recent decades (reviewed in Mathew, 2008; Sud, Kinnersley & Houlston, 2017). However, an overwhelming proportion of GWAS was conducted exclusively in populations of European ancestry and a limited number in populations of Asian ancestry (reviewed in Pereira et al., 2021). This also applies to OSCC; to date, all OSCC GWAS were conducted in populations of Asian and European ancestry (Abnet et al., 2010; Wang et al., 2010; McKay et al., 2011; Wu et al., 2011, 2012, 2014). Our study is the first OSCC GWAS conducted in a resident African black population.

Through the efforts of the Human Heredity & Health in Africa (H3A) consortium (Mulder et al., 2018) and the availability of the H3Africa Custom ~2.3 million SNP microarray, conducting robust GWAS in African populations has become viable. In addition to technological improvements, collaborative initiatives with other large genomic studies were also essential for the success of our study. The OSCC GWAS was made possible by the availability of good quality genomic DNA which we were able to extract from the JCS and the collaboration with the AWI-Gen study which provided valuable population-matched controls genotyped using the H3A Custom SNP array. The trans-ethnic meta-analysis was facilitated by the collaboration with Abnet et al (2010), who provided the necessary summary statistics from a Chinese OSCC GWAS.

The genetic diversity and continental and regional population substructure present within the African and SAB populations are complex (Choudhury et al., 2020, 2021; Sengupta et al., 2021). Such regional diversity and genomic substructure require careful and extensive correction to avoid genomic inflation and false-positive signals driven by substructure rather than by the disease phenotype. In addition to the traditional, well-established methodologies such as principal component analysis and kinship matrix established in linear-mixed models (Yu et al., 2008; Liu et al., 2011; Chen et al., 2016), our study also incorporated admixture analysis to exclude individuals (both cases and controls), who self-reported as black Africans, but had sizable non-African genomic contributions. Lastly, we incorporated iterative random sub-sampling permutation testing of controls. This iterative permutation testing involved random sub-sampling of controls originating from both the Soweto and Dikgale regions and testing for genetic differences using LMMs. This approach allowed us to identify and remove SNPs that produced false-positive association signals in the OSCC GWAS due to subpopulation-specific differences. Most of the population substructure signals inflating the OSCC GWAS results were accounted for by the genetic differences observed between male and female controls recruited from the Soweto region. Although this finding was unexpected, historically Soweto has been a hot spot for migrant male labourers from other parts of South Africa seeking employment (Statistics South Africa, 2006, 2010), and there are differences in the ethnolinguistic composition of males and females in the AWI-Gen controls which are likely to contribute to the genetic differences. Sengupta et al. (2021) reported a strong

correlation between genetic and linguistic variation in the South Eastern Bantu speakers and also between genetic variation and the geographic spread of these populations. The combination of methods used in this study provided a robust approach to control for the complex and diverse population substructure within our samples.

Our African GWAS for OSCC detected robust association signals on chromosome 9 and chromosome 2. The top signal on chromosome 9 is located upstream of the *FAM120A* gene. The lead SNP on chromosome 9, rs12379660, was also the strongest signal in the meta-analysis with the Chinese OSCC GWAS (P_{meta} = 9.36x10^{-10}, OR (95%CI): 1.51 (1.42-1.60)). This SNP is common in both populations and this finding is suggestive of possible shared risk alleles for OSCC in African and Asian populations.

The variant rs12379660 co-localized with cis-eQTLs in *FAM120AOS*. The cis-eQTLs appear to upregulate the expression of *FAM120AOS* and *FAM120A* in oesophageal mucosa tissue (GTEx Consortium, 2013). The increased expression of *FAM120A* may be suggestive of dysregulation of the oxidative stress-induced survival signalling thus supporting cancer cell survival in environments with various oxidative stresses (Tanaka et al, 2009). Further functional studies, such as gene editing of risk variants in oesophageal epithelial cell lines, may provide evidence of the biological effect of the risk SNPs in this region.
Our OSCC GWAS also detected robust association signals nearing genome-wide significance on chromosome 2 within the *MYO1B* gene. The lead SNP on chromosome 2 is monomorphic in populations of non-African ancestry (Supplementary Table 5.3) and is rare in African populations (EAF_{case} = 1.7%), which is suggestive of a population-specific risk allele for African OSCC. The functional relevance of this finding is not yet clear. This locus would require further replication of the association to achieve genome-wide significance before undertaking any functional analysis of this locus.

The trans-ethnic meta-analysis supported our previous findings that SNPs in *CHEK2* are a risk factor for OSCC in both African and Asian populations. However, the well-established *PLCE1* risk locus currently appears to be restricted to the Chinese population, with little or no evidence of association observed in the African OSCC GWAS. This finding supports the hypothesis of the development of population distinct risk factors after the Out-of-Africa migration (López, van Dorp & Hellenthal, 2015). Similarly, the *MYO1B* risk locus may have

developed independently in the African population. Aside from the findings discussed above, there was very little replication of known OSCC risk loci in the African population (Supplementary Table 5.5). Nevertheless, the genetic heritability for OSCC estimated in the Chinese population and the SAB population were similar (CHN: 38% vs AFR: 33%) and confirmed that genetic risk factors make a substantial contribution to the overall risk of OSCC.

Limitations

Our GWAS was limited by the lack of an independent study sample of African ancestry to replicate our findings. Such validation studies are essential to either confirm or reject the studies in the current OSCC. Although trans-ethnic OSCC GWAS data are available, there are no other data from African populations, so we were unable to test the African-specific OSCC genetic risk factors for replication. The lack of validation study samples also impacted the utilization of our findings for the construction and validation of a viable polygenic risk score (PRS) for African OSCC (Choi & O'Reilly, 2019). PRSs have been developed for several cancers, such as breast and prostate cancer, and have robust capabilities in risk stratification (Lewis & Vassos, 2020; Adeyemo et al., 2021). Specifically, an OSCC PRS could potentially be applied as a screening tool for at-risk population groups for risk stratification and guide early screening and intervention strategies (Lewis & Vassos, 2020).

The clinical application of PRSs is promising, especially when combined with other risk factors. Risk prediction algorithms can combine genetic, lifestyle and environmental risk factors to stratify and identify high-risk individuals in high-incidence populations have been developed, for example, in breast and ovarian cancer. The BOADICEA or Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm is designed to compute a comprehensive breast and/or ovarian cancer risk score using monogenic, polygenic genetic risk variants, lifestyle risk factors and mammographic density (Lee et al, 2019). The risk score that incorporates PRS is being trialled for use to directly inform disease management and screening strategies. An early attempt to incorporate genetic and lifestyle factors into a risk prediction model for OSCC in the Chinese population which incorporated SNPs and their interaction with drinking alcohol produced an AUC of 0.7 (Chang et al, 2013). As our knowledge of the genetic contribution to OSCC expands, the

development of a more effective algorithm for OSCC might provide a meaningful way to identify and stratify at-risk populations and potentially contribute to the development of a screening test for the early diagnosis of OSCC in a low resource setting.

Another limitation of our GWAS study is that, at the current sample size, we were not able to assess the contributions of smaller-effect risk loci to OSCC in African populations. Much larger sample size will be required to detect a larger proportion of the genetic contribution to the risk of OSCC in African populations. Using the Quanto power calculator (Gauderman & Morrison, 2006), for an SNP with MAF of 0.25, OR 1.25, we will require 4,041 cases and 4,041 controls to reach 80% power at an alpha of $5x10^{-8}$. Our GWAS had 21% power to detect such an effect at alpha = $5x10^{-8}$, or 91% at alpha = 0.001.

Future studies

Various large initiatives are currently ongoing, adding to the pool of knowledge for OSCC. The African Esophageal Cancer Consortium (AfrECC) is a partnership of African, European and American scientists and clinicians which was established to reduce the OSCC disease burden and raise awareness, as well as increase African research and clinical capacities in partner countries (Abnet et al., 2022). AfrECC has been conducting case-control studies across Kenya, Zambia, Tanzania and Malawi to identify risk factors for OSCC, and has enrolled over 2,000 OSCC cases and 2,000 controls in preparation for conducting an East African OSCC GWAS. Findings from the AfrECC GWAS will be a valuable resource to allow for future replication, validation and meta-analysis studies of African OSCC.

Prospective enrolment of OSCC cases and controls are also ongoing through our collaborative initiatives with the University of the Witwatersrand, Johannesburg (P.I.s C Mathew, P Ruff, C Penny), the Charlotte Maxeke Johannesburg Academic Hospital Johannesburg (L Prodehl), the Chris Hani Baragwanath Academic Hospital, Soweto (D Surridge), the Grey's Hospital, Pietermaritzburg (L Ferndale), the University of Cape Town (MI Parker) and Columbia University (J Abrams). Valuable biosamples such as fresh/FFPE tissue samples, saliva and peripheral blood samples are being banked for future genetic, genomic and microbiome studies.

Our group has also embarked on whole-exome sequencing of matched blood/tumour pairs from African OSCC patients to identify driver mutations and mutational signatures. To date, 18 pairs of tumour/blood samples from African OSCC patients have been sequenced, with a further 50 pairs of tumour/blood samples being sequenced at the Wellcome Trust Sanger Sequencing centre (P.I. P Jones).

6.2. Conclusion

In conclusion, through the work presented in this thesis, we first demonstrated the viability and feasibility of isolating high molecular genomic DNA from archival blood samples stored for up to 19 years. The DNA samples isolated from archival sources provided a valuable resource for large-scale genetic studies.

Our study also showed that some of the known genetic risk factors for OSCC identified in non-African populations do not replicate or transfer to the African population. Of the 7 risk loci, only one SNP in the *CHEK2* gene showed evidence of association with African OSCC.

We also conducted the first GWAS for OSCC in an African population. Our study has provided evidence of a substantial genetic contribution to OSCC risk in an African population. Our GWAS and trans-ethnic meta-analysis identified novel and shared risk loci for OSCC in populations of diverse ancestry. Additional extensive studies with larger sample sizes are required to detect more of the genetic contribution to African OSCC and to develop a polygenic risk score for OSCC in African populations.

CONSOLIDATED REFERENCES

Abnet, C.C., Arnold, M. & Wei, W.-Q. (2018) Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology*. 154(2):360–373. DOI: 10.1053/j.gastro.2017.08.023.

Abnet, C.C., Freedman, N.D., Hu, N., et al. (2010) A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature Genetics*. 42(9):764–767. DOI: 10.1038/ng.649.

Alexander, D.H., Novembre, J. & Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*. 19(9):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(sup2):1507133. DOI: 10.1080/16549716.2018.1507133.

AlRokayan, S.A.H. (2000) Effect of Storage Temperature on the Quality and Quantity of DNA Extracted from Blood. Pakistan Journal of Biological Sciences. 3(3):392–394. DOI: 10.3923/pjbs.2000.392.394.

Arnold, M., Soerjomataram, I., Ferlay, J., et al. (2015) Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*. 64(3):381–387. DOI: 10.1136/gutjnl-2014-308124.

Auton, A., Abecasis, G.R., Altshuler, D.M., et al. (2015) A global reference for human genetic variation. *Nature*. 526(7571):68–74. DOI: 10.1038/nature15393.

Boua, P.R., Brandenburg, J.T., Choudhury, A., et al. (2020) 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*. 10(February). DOI: 10.3389/fgene.2019.01354.

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):855. DOI: 10.1038/s41467-022-28276-x.

Buniello, A., Macarthur, J.A.L., Cerezo, M., et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. 47(D1):D1005–D1012. DOI: 10.1093/nar/gky1120.

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):661–678. DOI: 10.1038/nature05911.

Bye, H., Prescott, N.J., Lewis, C.M., et al. (2012) Distinct genetic association at the *PLCE1* locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis*. 33(11):2155–2161. DOI: 10.1093/carcin/bgs262.

Bye, H., Prescott, N.J., Matejcic, M., et al. (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*. 32(12):1855–1861. DOI: 10.1093/carcin/bgr211.

Capowski, E.E., Esnault, S., Bhattacharya, S., et al. (2001) Y box-binding factor promotes eosinophil survival by stabilizing granulocyte-macrophage colony-stimulating factor mRNA. *Journal of immunology (Baltimore, Md. : 1950)*. 167(10):5970–6. DOI: 10.4049/jimmunol.167.10.5970.

Cerami, E., Gao, J., Dogrusoz, U., et al. (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*. 2(5):401–4. DOI: 10.1158/2159-8290.CD-12-0095.

Chang, J., Huang, Y., Wei, L., et al. (2013) Risk prediction of esophageal squamous-cell carcinoma with common genetic variants and lifestyle factors in Chinese population. *Carcinogenesis*. 34(8):1782-1786. DOI: 10.1093/carcin/bgt106.

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

Chang, J., Tan, W., Ling, Z., et al. (2017) Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nature communications*. 8:15290. DOI: 10.1038/ncomms15290.

Chang, J., Zhong, R., Tian, J., et al. (2018) Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics*. 50(3):338–343. DOI: 10.1038/s41588-018-0045-8.

Chapman, B. V, Wald, A.I., Akhtar, P., et al. (2015) MicroRNA-363 targets myosin 1B to reduce cellular migration in head and neck cancer. *BMC cancer*. 15:861. DOI: 10.1186/s12885-015-1888-3.

Chattopadhyay, R., Das, S., Maiti, A.K., et al. (2008) Regulatory role of human AP-endonuclease (APE1/Ref-1) in YB-1-mediated activation of the multidrug resistance gene MDR1. *Molecular and cellular biology*. 28(23):7066–80. DOI: 10.1128/MCB.00244-08.

Chelule, P.K., Pegoraro, R.J., Gqaleni, N., et al. (2006) The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. *Disease markers*. 22(5–6):351–4. DOI: 10.1155/2006/980392.

Chen, C.Y., Gherzi, R., Andersen, J.S., et al. (2000) Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. *Genes & development*. 14(10):1236–48. Available: http://www.ncbi.nlm.nih.gov/pubmed/10817758.

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):513–520. DOI: 10.1093/carcin/bgz026.

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):106–113. DOI: 10.1089/bio.2017.0050.

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(March):101701. DOI: 10.1016/j.canep.2020.101701.

Chen, X., Iliopoulos, D., Zhang, Q., et al. (2014) *XBP1* promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature*. 508(7494):103–107. DOI: 10.1038/nature13119.

Chen, X., Li, A., Sun, B.-F., et al. (2019) 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nature cell biology*. 21(8):978–990. DOI: 10.1038/s41556-019-0361-y.

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

Choudhury, A., Ramsay, M., Hazelhurst, S., et al. (2017) Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*. 8(1):1–12. DOI: 10.1038/s41467-017-00663-9.

Choudhury, A., Sengupta, D., Ramsay, M., et al. (2021) Bantu-speaker migration and admixture in southern Africa. Human Molecular Genetics. 30(R1):R56–R63. DOI: 10.1093/hmg/ddaa274.

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 Comms*. 13(2578). DOI: 10.1038/s41467-022-30098-w.

Cui, R., Kamatani, Y., Takahashi, A., et al. (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*. 137(5):1768–1775. DOI: 10.1053/j.gastro.2009.07.070.

Dandara, C., Ballo, R. & Parker, M.I. (2005) CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer letters*. 225(2):275–82. DOI: 10.1016/j.canlet.2004.11.004.

Dandara, C., Li, D.P., Walther, G. et al. (2006). Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis*. 27(4):791–7. DOI: 10.1093/carcin/bgi257.

Desjardins, P. & Conklin, D. (2010) NanoDrop microvolume quantitation of nucleic acids. *Journal of visualized experiments :* JoVE. (45). DOI: 10.3791/2565.

Di Pardo, B.J., Bronson, N.W., Diggs, B.S., et al. (2016) The Global Burden of Esophageal Cancer: A Disability-Adjusted Life-Year Approach. *World journal of surgery*. 40(2):395–401. DOI: 10.1007/s00268-015-3356-2.

Di Pietro, F., Ortenzi, F., Tilio, M., et al. (2011) Genomic DNA extraction from whole blood stored from 15- to 30-years at -20 °C by rapid phenol-chloroform protocol: a useful tool for genetic epidemiology studies. Molecular and cellular probes. 25(1):44–8. DOI: 10.1016/j.mcp.2010.10.003.

Dietzsch, E. & Parker, M.I. (2002) Infrequent somatic deletion of the 5' region of the COL1A2 gene in oesophageal squamous cell cancer patients. *Clinical chemistry and laboratory medicine*. 40(9):941–5. DOI: 10.1515/CCLM.2002.165.

Dietzsch, E., Laubscher, R. & Parker, M.I. (2003) Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *International journal of cancer*. 107(1):38–45. DOI: 10.1002/ijc.11314.

Eltahir, H.A., Adam, A.A.M., Yahia, Z.A., et al. (2012) p53 Codon 72 arginine/proline polymorphism and cancer in Sudan. *Molecular biology reports*. 39(12):10833–6. DOI: 10.1007/s11033-012-1978-0.

Enzinger, P.C. & Mayer, R.J. (2003) Esophageal Cancer. *New England Journal of Medicine*. 349(23):2241–2252. DOI: 10.1056/NEJMra035010.

Erkeller-Yuksel, F.M., Deneys, V., Yuksel, B., et al. (1992) Age-related changes in human blood lymphocyte subpopulations. The Journal of pediatrics. 120(2 Pt 1):216–22. DOI: 10.1016/s0022-3476(05)80430-5.

Fabregat, A., Sidiropoulos, K., Viteri, G., et al. (2017) Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*. 18(1):142. DOI: 10.1186/s12859-017-1559-2.

Ferlay, J., Soerjomataram, I., Dikshit, R., et al. (2015) Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*. 136(5):E359–E386. DOI: 10.1002/ijc.29210.

Ferndale, L., Ayeni, O., Chen, W.C., et al. (2021) Factors Influencing Survival Time in Patients Treated for Oesophageal Cancer with Palliative Intent in South Africa. *ResearchSquare-Preprint*. 1–18. DOI: 10.21203/rs.3.rs-1074694/v1.

Ferndale, L., Sartorius, B., Aldous, C., et al. (2019) Oesophageal cancer in Area 2 of KwaZulu-Natal: predictors of late presentation. *South African Journal of Surgery*. 57(2 SE-General Surgery). Available: http://sajs.redbricklibrary.com/index.php/sajs/article/view/2948.

Forman, D., Bray, F., Brewster, D.H., et al. (2013) Cancer Incidence in Five Continents, Vol. X. Lyon: IARC Scientific Publications, No 164.

Gabriel, S., Ziaugra, L. & Tabbaa, D. (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics*. Chapter 2:Unit 2.12. DOI: 10.1002/0471142905.hg0212s60.

Gamieldien, W., Victor, T.C., Mugwanya, D., et al. (1998) p53 and p16/CDKN2 gene mutations in esophageal tumors from a high-incidence area in South Africa. *International journal of cancer*. 78(5):544–9. DOI: 10.1002/(sici)1097-0215(19981123)78:5<544::aid-ijc3>3.0.co;2-t.

Gauderman, W.J. & Morrison, J.M. (2006) QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe.

Gaudreault, I., Guay, D. & Lebel, M. (2004) YB-1 promotes strand separation in vitro of duplex DNA containing either mispaired bases or cisplatin modifications, exhibits endonucleolytic activities and binds several DNA repair proteins. *Nucleic acids research*. 32(1):316–27. DOI: 10.1093/nar/gkh170.

Goh, L.L., Lim, C.W., Sim, W.C., et al. (2017) Analysis of Genetic Variation in CYP450 Genes for Clinical Implementation. *PloS one*. 12(1):e0169233. DOI: 10.1371/journal.pone.0169233.

Griffiths, L. & Chacon-Cortes, D. (2014) Methods for extracting genomic DNA from whole blood samples: current perspectives. Journal of Biorepository Science for Applied Medicine. (May):1. DOI: 10.2147/BSAM.S46573.

Griss, J., Viteri, G., Sidiropoulos, K., et al. (2020) ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Molecular & cellular proteomics : MCP*. 19(12):2115–2125. DOI: 10.1074/mcp.TIR120.002155.

GTEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. *Nature genetics*. 45(6):580–5. DOI: 10.1038/ng.2653.

GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* (New York, N.Y.). 348(6235):648–60. DOI: 10.1126/science.1262110.

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

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., et al. (2015) The African Genome Variation Project shapes medical genetics in Africa. *Nature*. 517(7534):327–332. DOI: 10.1038/nature13997.

Halec, G., Schmitt, M., Egger, S., et al. (2016) Mucosal alpha-papillomaviruses are not associated with esophageal squamous cell carcinomas: Lack of mechanistic evidence from South Africa, China and Iran and from a world-wide meta-analysis. *International journal of cancer*. 139(1): 85-98. DOI: 10.1002/ijc.29911.

Hara, M., Nakanishi, H., Yoneyama, K., et al. (2016) Effects of storage conditions on forensic examinations of blood samples and bloodstains stored for 20 years. Legal medicine (Tokyo, Japan). 18:81–4. DOI: 10.1016/j.legalmed.2016.01.003.

Hetz, C., Chevet, E. & Harding, H.P. (2013) Targeting the unfolded protein response in disease.

Nature reviews. *Drug discovery*. 12(9):703–19. DOI: 10.1038/nrd3976.

Holland, N.T., Smith, M.T., Eskenazi, B. et al. (2003) Biological sample collection and processing for molecular epidemiological studies. Mutation research. 543(3):217–34. DOI: 10.1016/s1383-5742(02)00090-x.

Horwitz, E.M., Maloney, K.A. & Ley, T.J. (1994) A human protein containing a "cold shock" domain binds specifically to H-DNA upstream from the human gamma-globin genes. *The Journal of biological chemistry*. 269(19):14130–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/8188694.

Hyland, P.L., Zhang, H., Yang, Q., et al. (2016) Pathway, in silico and tissue-specific expression quantitative analyses of oesophageal squamous cell carcinoma genome-wide association studies data. *International journal of epidemiology*. 45(1):206–220. DOI: 10.1093/ije/dyv294.

Isaacson, C., Mothobi, P., Hale, M., et al. (2015) Carcinogenic nitrosamines in traditional beer as the cause of oesophageal squamous cell carcinoma in black South Africans. *South African Medical Journal*. 105(8):656. DOI: 10.7196/SAMJnew.7935.

Jassal, B., Matthews, L., Viteri, G., et al. (2020) The reactome pathway knowledgebase. *Nucleic acids research*. 48(D1):D498–D503. DOI: 10.1093/nar/gkz1031.

Jennis, M., Kung, C.-P., Basu, S., et al. (2016) An African-specific polymorphism in the *TP53* gene impairs p53 tumor suppressor function in a mouse model. *Genes & development*. 30(8):918–30. DOI: 10.1101/gad.275891.115.

Kamangar, F., Chow, W.-H., Abnet, C.C., et al. (2009) Environmental causes of esophageal cancer. *Gastroenterology clinics of North America*. 38(1):27–57, vii. DOI: 10.1016/j.gtc.2009.01.004.

Kayamba, V., Bateman, A.C., Asombang, A.W., et al. (2015) HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case–control study. *Cancer Medicine*. 4(4):588–595. DOI: 10.1002/cam4.434.

Klein, R.J., Zeiss, C., Chew, E.Y., et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 308(5720):385–389. DOI: 10.1126/science.1109557.

Kumar, S.K., Feldman, M.W., Rehkopf, D.H., et al. (2015) Limitations of GCTA as a solution to the missing heritability problem. *PNAS*. 113(1):E61-E70. DOI: 10.1073/pnas.1520109113

Lagisetty, K.H., Beer, D.G., & Chang, A.C. (2017) Squamous cell carcinomas and adenocarcinomas of the esophagus: One treatment does not rule them all. The *Journal of Thoracic and Cardiovascular Surgery*. 154: 1446-7. DOI: 10.1016/j.jtcvs.2017.04.074.

Lee, A., Mavaddat, N., Wilcox, A.N., et al. (2019) BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med*. 21(8):1708-1718. DOI: 10.1038/s41436-018-0406-9.

Lee, S., Park, S., Lee, H., et al. (2018) ChIP-seq analysis reveals alteration of H3K4 trimethylation occupancy in cancer-related genes by cold atmospheric plasma. *Free radical biology & medicine*. 126:133–141. DOI: 10.1016/j.freeradbiomed.2018.08.004.

Li, D., Dandara, C. & Parker, M.I. (2005) Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clinical chemistry and laboratory medicine*. 43(4):370–5. DOI: 10.1515/CCLM.2005.067.

Li, D., Dandara, C. & Parker, M.I. (2010) The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC genetics*. 11:47. DOI: 10.1186/1471-2156-11-47.

Li, D.P., Dandara, C., Walther, G., et al. (2008) Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer. *Clinical chemistry and laboratory medicine*. 46(3):323–8. DOI: 10.1515/CCLM.2008.073.

Li, G., Song, Q., Jiang, Y., et al. (2020) Cumulative Evidence for Associations between Genetic Variants and Risk of Esophageal Cancer. *Cancer Epidemiology Biomarkers & Prevention*. 29(4):838–849. DOI: 10.1158/1055-9965.EPI-19-1281.

Li, W.-Q., Hu, N., Hyland, P.L., et al. (2013) Genetic variants in DNA repair pathway genes and risk of esophageal squamous cell carcinoma and gastric adenocarcinoma in a Chinese population. *Carcinogenesis*. 34(7):1536–1542. DOI: 10.1093/carcin/bgt094.

Liu, W., Snell, J.M., Jeck, W.R., et al. (2017) Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI insight*. 2(22). DOI: 10.1172/jci.insight.98457.

Loots, E., Ferndale, L. & Clarke, D. (2019) Oesophageal cancer in South Africa: the need for a public health response. *South African Journal of Surgery*. 57(2 SE-Editorial). Available: http://sajs.redbricklibrary.com/index.php/sajs/article/view/3094.

López, S., van Dorp, L. & Hellenthal, G. (2015) Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary bioinformatics online*. 11(Suppl 2):57–68. DOI: 10.4137/EBO.S33489.

Machiela, M.J. & Chanock, S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* (Oxford, England). 31(21):3555–7. DOI: 10.1093/bioinformatics/btv402.

Matejcic, M. & Iqbal Parker, M. (2015) Gene-environment interactions in esophageal cancer. *Critical reviews in clinical laboratory sciences*. 52(5):211–31. DOI: 10.3109/10408363.2015.1020358.

Matejcic, M., Gunter, M.J. & Ferrari, P. (2017) Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. *Carcinogenesis*. 38(9):859–872. DOI: 10.1093/carcin/bgx067.

Matejcic, M., Li, D., Prescott, N.J., et al. (2011) Association of a Deletion of *GSTT2B* with an Altered Risk of Oesophageal Squamous Cell Carcinoma in a South African Population: A Case-Control Study. *PLoS ONE*. 6(12):e29366. DOI: 10.1371/journal.pone.0029366.

Matejcic, M., Vogelsang, M., Wang, Y. et al. (2015) NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC cancer*. 15:658. DOI: 10.1186/s12885-015-1105-4.

Mathew, C.G. (2008) New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nature Reviews Genetics*. 9(1):9–14. DOI: 10.1038/nrg2203.

McKay, J.D., Truong, T., Gaborieau, V., et al. (2011) A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. *PLoS Genetics*. 7(3):e1001333. DOI: 10.1371/journal.pgen.1001333.

Menya, D., Kigen, N., Oduor, M., et al. (2019) Traditional and commercial alcohols and esophageal cancer risk in Kenya. *International journal of cancer*. 144(3):459–469. DOI: 10.1002/ijc.31804.

Menya, D., Maina, S.K., Kibosia, C., et al. (2019) Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case-control study and a pan-African perspective. *International journal of cancer*. 145(1):99–109. DOI: 10.1002/ijc.32086.

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

Mulder, N., Abimiku, A., Adebamowo, S.N., et al. (2018) H3Africa: current perspectives. *Pharmacogenomics and personalized medicine*. 11:59–66. DOI: 10.2147/PGPM.S141546.

Murphy, G., McCormack, V., Abedi-Ardekani, B., et al. (2017) International cancer seminars: a focus on esophageal squamous cell carcinoma. *Annals of Oncology*. 28(9):2086–2093. DOI: 10.1093/annonc/mdx279.

Murphy, M.E., Liu, S., Yao, S., et al. (2017) A functionally significant SNP in TP53 and breast cancer risk in African-American women. *NPJ breast cancer*. 3:5. DOI: 10.1038/s41523-017-0007-9.

Mychaleckyj, J.C., Farber, E.A., Chmielewski, J., et al. (2011) Buffy coat specimens remain viable as a DNA source for highly multiplexed genome-wide genetic tests after long term storage. Journal of translational medicine. 9:91. DOI: 10.1186/1479-5876-9-91.

Naeger, L.K., McKinney, J., Salvekar, A., et al. (1999) Identification of a *STAT4* binding site in the interleukin-12 receptor required for signaling. *The Journal of biological chemistry*. 274(4):1875–8. DOI: 10.1074/jbc.274.4.1875.

Naidoo, R., Ramburan, A., Reddi, A. et al. (2005) Aberrations in the mismatch repair genes and the clinical impact on oesophageal squamous carcinomas from a high incidence area in South Africa. *Journal of clinical pathology*. 58(3):281–4. DOI: 10.1136/jcp.2003.014290.

Nakayama, Y., Yamaguchi, H., Einaga, N., et al. (2016) Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. PloS one. 11(3):e0150528. DOI:

10.1371/journal.pone.0150528.

Näslund-Koch, C., Nordestgaard, B.G., & Bojesen, S.E. (2016) Increased Risk for Other Cancers in Addition to Breast Cancer for CHEK2*1100delC Heterozygotes Estimated From the Copenhagen General Population Study. *J Clin Oncol*. 34(11):1208-16. DOI: 10.1200/JCO.2015.63.3594.

NCR. (2018) South African summary statistics of cancer diagnosed histologically in 2014. http://www.nicd.ac.za/wp-content/uploads/2017/03/2014-NCR-tables-1.pdf.

NCR. (2021) South African summary statistics of cancer diagnosed histologically in 2019. https://www.nicd.ac.za/wp-content/uploads/2021/12/NCR_Path_2019_Full_Report_8dec2021.pdf.

Nederhand, R.J., Droog, S., Kluft, C., et al. (2003) Logistics and quality control for DNA sampling in large multicenter studies. Journal of thrombosis and haemostasis : JTH. 1(5):987–91. DOI: 10.1046/j.1538-7836.2003.00216.x.

Ohmura, G., Tsujikawa, T., Yaguchi, T., et al. (2015) Aberrant Myosin 1b Expression Promotes Cell Migration and Lymph Node Metastasis of HNSCC. *Molecular cancer research : MCR*. 13(4):721–31. DOI: 10.1158/1541-7786.MCR-14-0410.

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):1751–6. DOI: 10.1038/sj.bjc.6600338.

Patel, K., Mining, S., Wakhisi, J., et al. (2011) TP53 mutations, human papilloma virus DNA and inflammation markers in esophageal squamous cell carcinoma from the Rift Valley, a high-incidence area in Kenya. *BMC research notes*. 4:469. DOI: 10.1186/1756-0500-4-469.

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

Qiagen®. (2010) FlexiGene® DNA Handbook. Available: http://www.qiagen.com/.

Ramsay, M., Crowther, N., Tambo, E., et al. (2016) 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*. 1:e20. DOI: 10.1017/gheg.2016.17.

Reichenbach, Z.W., Murray, M.G., Saxena, R., et al. (2019) Clinical and translational advances in esophageal squamous cell carcinoma. *In Advances in Cancer Research*. V. 144. Academic Press Inc. 95–135. DOI: 10.1016/bs.acr.2019.05.004.

Richardson, A.J., Narendran, N., Guymer, R.H., et al. (2006) Blood storage at 4 degrees C-factors involved in DNA yield and quality. *The Journal of laboratory and clinical medicine*. 147(6):290–4. DOI: 10.1016/j.lab.2006.01.005.

Robin, J.D., Ludlow, A.T., LaRanger, R., et al. (2016) Comparison of DNA Quantification Methods for Next Generation Sequencing. *Scientific Reports*. 6(1):24067. DOI: 10.1038/srep24067.

Rotimi, C.N., Abayomi, A., Abimiku A., et al. (2014) Research capacity. Enabling the genomic revolution in Africa. *Science*. 344(6190):1346-1348. DOI: 10.1126/science.1251546.

Rotimi, C.N., Bentley, A.R., Doumatey, A.P., et al. (2017) The genomic landscape of African populations in health and disease. *Human Molecular Genetics*. 26(R2):R225–R236. DOI: 10.1093/hmg/ddx253.

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):djv279. DOI: 10.1093/jnci/djv279.

Schäfer, G., Kabanda, S., van Rooyen, B., et al. (2013) The role of inflammation in HPV infection of the Oesophagus. *BMC Cancer*. 13(1):185. DOI: 10.1186/1471-2407-13-185.

Segal, I., Reinach, S. & de Beer, M., (1988) Factors associated with oesophageal cancer in Soweto, South Africa. *British Journal of Cancer*. 58(5):681–686. DOI: 10.1038/bjc.1988.286.

Sengupta, D., Choudhury, A., Fortes-Lima, C., et al. (2021) Genetic substructure and complex demographic history of South African Bantu speakers. Nature Communications. 12(1):2080. DOI: 10.1038/s41467-021-22207-y.

Siewert, J.R. & Ott, K. (2007) Are squamous and adenocarcinomas of the esophagus the same disease? *Seminars in Radiation Oncology*. 17(1): 38-44. DOI: 10.1016/j.semradonc.2006.09.007.

Simba, H., Kuivaniemi, H., Lutje, V., et al. (2019) Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in Genetics*. DOI: 10.3389/fgene.2019.00642.

Sitas, F., Egger, S., Urban, M.I., et al. (2012) InterSCOPE study: Associations between esophageal squamous cell carcinoma and human papillomavirus serological markers. *Journal of the National Cancer Institute*. 104(2):147–58. DOI: 10.1093/jnci/djr499.

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):486. DOI: 10.7196/SAMJ.2017.v107i6.12099.

Southey, M.C., Goldgar, D.E., Winqvist, R., et al. (2016) *PALB2*, *CHEK2* and *ATM* rare variants and cancer risk: data from COGS. *Journal of medical genetics*. 53(12):800–811. DOI: 10.1136/jmedgenet-2016-103839.

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

Statistics South Africa. (2006) Mid-year population estimates, South Africa 2005. 27(12).

Statistics South Africa. (2010) Statistical release Mid-year population estimates 2010. *Stats SA*. (July):16.

Steinberg, K.K., Sanderlin, K.C., Ou, C.Y., et al. (1997) DNA banking in epidemiologic studies. Epidemiologic reviews. 19(1):156–62. DOI: 10.1093/oxfordjournals.epirev.a017938.

Strickland, N.J., Matsha, T., Erasmus, R.T., et al. (2012) Molecular analysis of ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International journal of cancer*. 131(3):623–32. DOI: 10.1002/ijc.26418.

Sud, A., Kinnersley, B. & Houlston, R.S. (2017) Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer*. 17(11):692–704. DOI: 10.1038/nrc.2017.82.

Sun, J., Shen, L., Shah, M.A., et al. (2021) Pembrolizumab plus chemotherapy versus chemotherapy alone for first-line treatment of advanced oesophageal cancer (KEYNOTE-590): a randomised, placebo-controlled, phase 3 study. *The Lancet*. 398(10302):759-771. DOI: 10.1016/S0140-6736(21)01234-4

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):209–249. DOI: 10.3322/caac.21660.

Swiel, Y., Brandenburg, J.T., Hayat, M., et al. (2022) FPGA Acceleration of GWAS Permutation Testing. *bioRXiv*, preprint. DOI: 10.1101/2022.03.11.483235.

Syrmis, M.W., Moser, R.J., Whiley, D.M., et al. (2011) Comparison of a multiplexed MassARRAY system with real-time allele-specific PCR technology for genotyping of methicillin-resistant Staphylococcus aureus. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*. 17(12):1804–10. DOI: 10.1111/j.1469-0691.2011.03521.x.

Teo, Y.-Y., Small, K.S. & Kwiatkowski, D.P. (2010) Methodological challenges of genome-wide association analysis in Africa. *Nature Reviews Genetics*. 11(2):149–160. DOI: 10.1038/nrg2731.

The Cancer Genome Atlas Research Network. (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature*. 541: 169-175. DOI: 10.1038/nature20805.

Thermo Scientific™. (2011) NanoDrop: Assessment of Nucleic Acid Purity. Protoc Prod Manuals. 1–2.

Thumbs, A., Borgstein, E., Vigna, L., et al. (2012) Self-expanding metal stents (SEMS) for patients with advanced Esophageal cancer in Malawi: An effective palliative treatment. *Journal of Surgical Oncology*. 105(4):410–414. DOI: 10.1002/jso.23003.

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):e1001182. DOI: 10.1371/journal.pmed.1001182.

van Loon, K., Mwachiro, M.M., Abnet, C.C., et al. (2018) The African Esophageal Cancer Consortium: A Call to Action. *Journal of Global Oncology*. 2018(4):1–9. DOI: 10.1200/JGO.17.00163.

Victor, T., Du Toit, R., Jordaan, A.M., et al. (1990) No evidence for point mutations in codons 12, 13, and 61 of the ras gene in a high-incidence area for esophageal and gastric cancers. *Cancer research*. 50(16):4911–4. Available: http://www.ncbi.nlm.nih.gov/pubmed/2199031.

Vogelsang, M., Wang, Y., Veber, N., et al. (2012) The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PLoS one*. 7(5):e36962. DOI: 10.1371/journal.pone.0036962.

Vos, M., Adams, C.H., Victor, T.C., et al. (2003) Polymorphisms and mutations found in the regions flanking exons 5 to 8 of the TP53 gene in a population at high risk for esophageal cancer in South Africa. *Cancer genetics and cytogenetics*. 140(1):23–30. DOI: 10.1016/s0165-4608(02)00638-6.

Wang, L.-D., Zhou, F.-Y., Li, X.-M., et al. (2010). Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at PLCE1. *Nature Genetics*. 42(9):759–763. DOI: 10.1038/ng.648.

Watanabe, K., Taskesen, E., van Bochoven, A., (2017) Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*. 8(1):1826. DOI: 10.1038/s41467-017-01261-5.

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):700–712. DOI: 10.1093/bib/bbw145.

Westra, H.-J., Peters, M.J., Esko, T., et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*. 45(10):1238–1243. DOI: 10.1038/ng.2756.

White, R.E., Parker, R.K., Fitzwater, J.W., et al. (2009) Stents as sole therapy for oesophageal cancer: a prospective analysis of outcomes after placement. *The Lancet Oncology*. 10(3):240–246. DOI: 10.1016/S1470-2045(09)70004-X.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available: https://ggplot2.tidyverse.org.

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

Wu, C., Hu, Z., He, Z., et al. (2011) Genome-wide association study identifies three new

susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature genetics*. 43(7):679–684. DOI: 10.1038/ng.849.

Wu, C., Kraft, P., Zhai, K., et al. (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nature genetics*. 44(10):1090–1097. DOI: 10.1038/ng.2411.

Wu, C., Wang, Z.Z., Song, X., et al. (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nature genetics*. 46(9):1001–1006. DOI: 10.1038/ng.3064.

Zaahl, M.G., Warnich, L., Victor, T.C., et al. (2005) Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer genetics and cytogenetics*. 159(1):48–52. DOI: 10.1016/j.cancergencyto.2004.09.017.

Zannini, L., Delia, D. & Buscemi, G. (2014) CHK2 kinase in the DNA damage response and beyond. *Journal of molecular cell biology*. 6(6):442–57. DOI: 10.1093/jmcb/mju045.

Zhang, H.-R., Lai, S.-Y., Huang, L.-J., et al. (2018) Myosin 1b promotes cell proliferation, migration, and invasion in cervical cancer. *Gynecologic oncology*. 149(1):188–197. DOI: 10.1016/j.ygyno.2018.01.024.

Zhang, Y.D., Hurson, A.N., Zhang, H, et al. (2020) Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications*. 11, 2253. DOI: 10.1038/s41467-020-16483-3.

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

APPENDIX

A – Co-author agreement form/ Supervisor declaration

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

31 March 2022

RE: WCHEN 0700425A: Permission to include articles in thesis.

Dear Colleagues,

This letter is in support of Mr WC Chen's PhD submission. Mr WC Chen is submitting three original manuscripts as part of his PhD thesis. Signatures from co-authors were obtained where possible. All co-authors previously consented to the work being submitted for publication.

1. **Chen WC**, 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. https://doi.org/10.1089/bio.2017.0050

A Sobalisa and ST Duze were unreachable.

2. **Chen WC**, 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. https://doi.org/10.1093/carcin/bgz026

H Bye, M Matejcic, A Amar, D Govender, YW Khew, V Beynon and NJ Prescott are collaborators based in the UK. We have not yet received a response from them at the time of the PhD thesis submission.

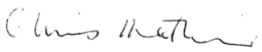
3. **Chen WC**, Brandenburg JT, Choudhury A, Hayat M, Sengupta D, Swiel Y, Babb de Villiers C, Ferndale F, Aldous C, Soo C, Lee S, Curtis C, Newton R, Waterboer T, Sitas F, Bradshaw D, Abnet CC, Ramsay M, Parker MI, Singh E, Lewis CM & Mathew CG (2022) Genome-wide association study of

oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations. Submitted to Nature Communications.

CJ Curtis, D Bradshaw and CC Abnet are active collaborators. No responses were received at the time of the PhD thesis submission.

Please refer to the Permission form attached.

Yours sincerely,



Professor Christopher Mathew, FMedSci

Professor in Human Genetics

Sydney Brenner Institute for Molecular Bioscience

University of the Witwatersrand, South Africa

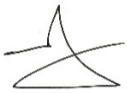
Email: christopher.mathew@kcl.ac.uk

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

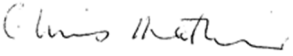
FACULTY OF HEALTH SCIENCES

Declaration: Student's contribution to article(s) and agreement of co-author(s)

I, **Wenlong Carl Chen**, student number **0700425A**, declare that this Thesis/Dissertation/ Research Report is my own work and that I contributed adequately towards the research findings published in the article(s) stated below which are included in my Thesis/Dissertation/Research Report.

Signature of Student  Date: 21 March 2022

Name of Primary Supervisor: Professor Christopher Mathew

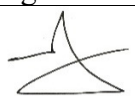
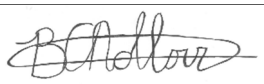
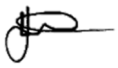
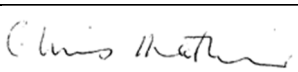
Signature of Primary Supervisor  Date: 18 March 2022

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article 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: **The integrity and yield of genomic DNA isolated from whole blood following long-term storage at -30°C.**

Journal name, year, volume and page numbers: **Biopreservation and Biobanking. 2018. 16(2) 106-113.**

Authors	Name	Signature	Date
1st author/ Primary	Wenlong Carl Chen		21 March 2022
2nd author/ Co-author	Robyn Kerr		18 March 2022
3rd author/ Co-author	Andrew May		18 March 2022
4th author/ Co-author	Babongile Ndlovu		18 March 2022
5th author/ Co-author	Anelisa Sobalisa		
6th author/ Co-author	Sanelisiwe T Duze		
7th author/ Co-author	Lavania Joseph		18 March 2022
8th author/ Senior author	Christopher G Mathew		18 March 2022
9th author/ Senior author	Chantal Babb de Villiers		20 March 2022

Comments by primary supervisor:

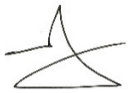
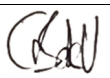
The work presented in this paper was led by the PhD candidate who did the majority of the DNA extractions and quality control analysis. He wrote the paper, with input from his co-supervisors and co-authors. All authors read and approved the manuscript before submission.

C. Mathew

Article 2:

Title: **Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population**

Journal name, year, volume and page numbers: **Carcinogenesis. 2019. 40(4) 513-520.**

Authors	Name	Signature	Date
1st author/ Primary	Wenlong Carl Chen		21 March 2022
2nd author/ Co-author	Hannah Bye		
3rd author/ Co-author	Marco Matejcic		
4th author/ Co-author	Ariella Amar		
5th author/ Co-author	Dhiren Govender		
6th author/ Co-author	Yee Wen Khew		
7th author/ Co-author	Victoria Beynon		
8th author/ Co-author	Robyn Kerr		18 March 2022
9th author/ Co-author	Elvira Singh	(*Dr Elvira Singh passed away on 27 February 2022)	
10th author/ Co-author	Natalie J Prescott		
11th author/ Co-author	Cathryn M Lewis		17 March 2022
12th author/ Co-author	Chantal Babb de Villiers		20 March 2022
13th author/ Co-author	M Iqbal Parker		28 March 2022
14th author/ Senior author	Christopher G Mathew		18 March 2022

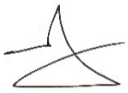
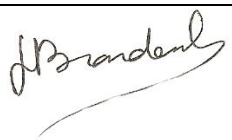
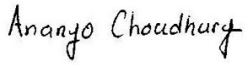
Comments by primary supervisor:

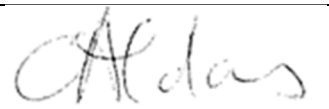
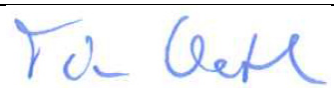
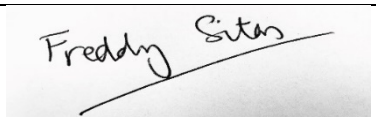
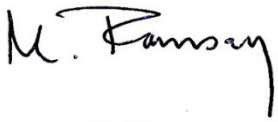
The work presented in this paper was led by the PhD candidate, WC Chen who did all the analysis of the genotype data for the Johannesburg Cancer Study (JCS) cases and controls. The original analysis of the University of Cape Town (UCT) case and control data was done by the joint primary author, Hannah Bye who is second on the author list. The candidate did all the work for the meta-analysis of the JCS and UCT data, which combined the data from the two studies, and the tests for gene x environmental interactions. He also wrote the paper, with input from the co-supervisors and co-authors. All authors read and approved the manuscript before submission. C. Mathew

Article 3:

Title: **Genome-wide association study of oesophageal cancer identifies trans-ethnic risk variants in African and Chinese populations**

Journal name, year, volume and page numbers: **Nature Communications. 2022.**

Authors	Name	Signature	Date
1st author/ Primary	Wenlong Carl Chen		21 March 2022
2nd author/ Co-author	Jean-Tristan Brandenburg		30 March 2022
3rd author/ Co-author	Ananyo Choudhury		29 March 2022
4th author/ Co-author	Mahtaab Hayat		17 March 2022
5th author/ Co-author	Dhriti Sengupta		29 March 2022
6th author/ Co-author	Yaniv Swiel		28 March 2022
7th author/ Co-author	Chantal Babb de Villiers		20 March 2022
8th author/ Co-author	Lucien Ferndale		29 March 2022

9th author/ Co-author	Colleen Aldous		29 March 2022
10th author/ Co-author	Cassandra Soo		16 March 2022
11th author/ Co-author	Lee Sang		23 March 2022
12th author/ Co-author	Charles J. Curtis		
13th author/ Co-author	Rob Newton		17 March 2022
14th author/ Co-author	Tim Waterboer		29 March 2022
15th author/ Co-author	Freddy Sitas		18 March 2022
16th author/ Co-author	Debbie Bradshaw		
17th author/ Co-author	Christian C Abnet		
18th author/ Co-author	Michele Ramsay		20 March 2022
19th author/ Co-author	M. Iqbal Parker		28 March 2022
20th author/ Co-author	Elvira Singh	(*Dr Elvira Singh passed away on 27 February 2022)	
21st author/ Co-author	Cathryn Lewis		17 March 2022
22nd author/ Senior author	Christopher G Mathew		18 March 2022

Comments by primary supervisor:

The candidate did all the analysis and quality control for the genome-wide association study (GWAS) in the black South African population. Summary statistical data from a GWAS in the Chinese population was provided by Dr Abnet (National Cancer Institute, USA), and the meta-analysis of the African and Chinese data was done by the candidate. Additional supervision on the statistical analysis and population genetic aspects was provided by JT Brandenburg, A. Choudhury, D. Sengupta and C. Lewis. Y.Swiel assisted with the adaptive permutation analysis. The candidate wrote the paper, with input from the co-supervisors and co-authors. All authors read and approved the manuscript before submission. C. Mathew

B – Ethics certificates

R14/49 Mr Wenlong Chen et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1706115

NAME: (Principal Investigator)	Mr Wenlong Chen et al
DEPARTMENT:	Human Genetics Sydney Brenner Institute for Molecular Bioscience National Cancer Registry, National Health Laboratory Services
PROJECT TITLE:	The Genetic Aentiology of Oesophageal Cancer in South African Black Populations
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-Study under Primary Study (M160807)
SUPERVISOR:	Prof Christopher Mathew
APPROVED BY:	CB Penny 1/9/2017 Professor CB. Penny Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	31/08/2017

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/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand 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 June and will therefore be due in the month of June each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal ~~Investigator~~ Signature Date 2017/09/01

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

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

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Dr Chantal Babb

CLEARANCE CERTIFICATE	**M120117**
PROJECT	Genetic Research on Blood Samples from the Johannesburg Cancer Case Control Study
INVESTIGATORS	Dr Chantal Babb.
DEPARTMENT	National Cancer Registry and Research
DATE CONSIDERED	27/01/2012
DECISION OF THE COMMITTEE*	Approved unconditionally

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

DATE 27/01/2012 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

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

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...

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M140271

NAME: (Principal Investigator)	Dr Chantal Babb
DEPARTMENT:	Cancer Epidemiology Research Group National Health Laboratory Services
PROJECT TITLE:	Cancer Case-Control Study (Previously M981119/M040445/M090361)
DATE CONSIDERED:	Ad hoc
DECISION:	Renewal Approved
CONDITIONS:	
SUPERVISOR:	

APPROVED BY: _____
Professor PE Cleaton-Jones, Chairperson, HREC (Medical)

DATE OF APPROVAL: 24/03/2014

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 Secretary in Room 10004, 10th floor, Senate House, University.

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**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

28/3/14

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

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

UNIVERSITY OF THE
WITWATERSRAND
JOHANNESBURG

R49 Profs C Mathew & S Hazelhurst; Dr M Cox; Mr Y Swiel

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2111154

NAME: (Principal Investigator)	Profs C Mathew & S Hazelhurst; Dr M Cox; Mr Y Swiel
DEPARTMENT:	Faculty of Health Sciences Sydney Brenner Institute for Molecular Bioscience Medical School University
PROJECT TITLE:	*Investigation of genetic risk factors for African breast, cervical and oesophageal cancer*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Formerly M160807 (expired on 2021/10/20)
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Not applicable
APPROVED BY:	CB Penny Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2021/12/07

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

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 submit details to the Committee. **I agree to submit a yearly progress report**. When a funder requires annual re-certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in December and therefore reports and re-certification will be due in the month of **December** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Chris Mathew — 08 December 2021

Signature of Principal Investigator — Date

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:	CB Penny 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. K Ramsay — Principal Investigator Signature

Date: 22 June 2018

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

R14/49 Prof Michelle Ramsay

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1403107

NAME: (Principal Investigator) Prof Michelle Ramsay

DEPARTMENT: Sidney Brenner Institute for Molecular Bioscience

PROJECT TITLE: Approval for a New Biobank
Sidney Brenner Institute for Molecular Bioscience
(BEC20140302)

DATE CONSIDERED: 25/05/2014

DECISION: Approved unconditionally

CONDITIONS:

SUPERVISOR:

APPROVED BY:

Professor P Cleaton-Jones,
Chairperson, HREC (Medical)

Prof A Dhai,
Chairperson, BEC

DATE OF APPROVAL: 22/07/2015

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 Secretary in Room 10004, 10th floor, Senate House, University.

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**.

M. Ramsay
Principal Investigator Signature

28 July 2015
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

R14/49 Prof Michele Ramsay

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M200469

NAME: (**Principal Investigator**)	Prof Michele Ramsay
DEPARTMENT:	Sydney Brenner Institute for Molecular Bioscience
PROJECT TITLE:	Approval of a Biobank: BEC20200401 Sydney Brenner Institute for Molecular Bioscience 9 Jubilee Road, Parktown, Johannesburg, 2193
DATE CONSIDERED:	Initial Approval - 22/07/2015
DECISION:	Approved unconditionally
CONDITIONS:	Renewal for 5 Years Valid for the period 01 April 2020 to 31 March 2025 Previously BEC20140302 and M1403107
SUPERVISOR:	
APPROVED BY:	Dr N Naran, Co-Chairperson, HREC (Medical) Dr CB Penny, Co-Chairperson, BEC

DATE OF APPROVAL: 10/11/2020

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 **April** and will therefore be due in the month of **April** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

M. Ramsay — 16 November 2020

Principal Investigator Signature — Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

CERTIFICATE OF ETHICS APPROVAL
FOR COLLECTION AND STORAGE

This Certificate is awarded to:

Sydney Brenner Institute for Molecular Bioscience (SBIMB) Biobank
(**Site**) 9 Jubilee Road, Parktown, Johannesburg, 2193

Collection and Storage of Biological Samples

On behalf of The University of The Witwatersrand Johannesburg Research Office

BIOBANK ETHICS COMMITTEE
Sub-Committee of HREC (M) R14\49

Clearance Certificate: BEC20200401 (Previously BEC20140302), M200469 (Previously M1403107)
This certificate is valid for 5 years

10/11/2020

Dr Clement Penny
Co-Chair: Biobank Ethics Committee

10/11/2020

Dr Nitien Naran
Co-Chair: Human Research Ethics Committee

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

Clinical Trials Centre
The University of Hong Kong

Certificat de formation - Training Certificate

Ce document atteste que - this document certifies that

Wenlong Chen

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: 2020/12/24
CID : IcdIuYbvw

Professeur Dominique Sprumont
Coordinateur TRREE Coordinator

FMH Continuing Education Program (5 Credits)
Programme de formation continue (5 Crédits)

Foederatio Pharmaceutica Helvetiae FPH Programmes de formation continue
Continuing Education Programs
Programmes de formation 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 Sciences (SAMS/ASSM/SAMW) (www.samw.ch) - Commission for Research Partnerships with Developing Countries (www.kfpe.ch)

[REV : 20170310]

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

Clinical Trials Centre
The University of Hong Kong

Certificat de formation - Training Certificate

Ce document atteste que - this document certifies that

Wenlong Chen

a complété avec succès - has successfully completed

Research Ethics Evaluation

du programme de formation TRREE en évaluation éthique de la recherche
of the TRREE training programme in research ethics evaluation

Release Date: 2020/12/24
CID : wRgHOeKlpT

Professeur Dominique Sprumont
Coordinateur TRREE Coordinator

FMH Continuing Education Program (5 Credits)
Programme de formation continue (5 Crédits)

Foederatio Pharmaceutica Helvetiae FPH Programmes de formation continue
Continuing Education Programs
Programmes de formation 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 Sciences (SAMS/ASSM/SAMW) (www.samw.ch) - Commission for Research Partnerships with Developing Countries (www.kfpe.ch)

[REV : 20170310]

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

Clinical Trials Centre
The University of Hong Kong

Certificat de formation - Training Certificate

Ce document atteste que - this document certifies that

Wenlong Chen

a complété avec succès - has successfully completed

Informed Consent

du programme de formation TRREE en évaluation éthique de la recherche
of the TRREE training programme in research ethics evaluation

Release Date: 2020/12/24
CID : OLfzcGtcwW

Professeur Dominique Sprumont
Coordinateur TRREE Coordinator

FMH Continuing Education Program (5 Credits)
Programme de Formation continue (5 Crédits)

Foederatio Pharmaceutica Helvetiae FPH Programmes de formation continue
Continuing Education Programs
Programmes de formation 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 Sciences (SAMS/ASSM/SAMW) (www.samw.ch) - Commission for Research Partnerships with Developing Countries (www.kfpe.ch)

[REV : 20170310]

C - Published manuscripts

BIOPRESERVATION AND BIOBANKING
Volume 16, Number 2, 2018
© Mary Ann Liebert, Inc.
DOI: 10.1089/bio.2017.0050

The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term Storage at –30°C

Wenlong Carl Chen,[1–3] Robyn Kerr,[3,4] Andrew May,[3,4] Babongile Ndlovu,[1] Anelisa Sobalisa,[1,4] Sanelisiwe T. Duze,[1,5] Lavania Joseph,[1] Christopher G. Mathew,[2,6] and Chantal Babb de Villiers[3,7]

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

Long-term storage of whole blood can affect the integrity of DNA if it is not done under optimal conditions. The aim of this study was to determine whether long-term storage (2–19 years) of whole blood samples at −30°C had a negative effect on the quality or quantity of genomic DNA that could be recovered at extraction. Genomic DNA was isolated from 2758 whole blood samples collected in 4 mL EDTA vacutainers from 1997 to 2012. DNA was extracted using the Qiagen® FlexiGene® DNA kit. The average storage duration at −30°C was 12 years. The quality and quantity of the isolated DNA were assessed using spectrophotometry (NanoDrop™), a fluorometric assay for double-stranded DNA (Qubit™), and agarose gel electrophoresis. The mean DNA yield per sample was found to be 114 μg from whole blood volumes that ranged from 0.5 to 4 mL. The mean A260/280 ratio and median A260/280 ratios were both 1.8. No correlation was found between the duration of storage and the total yield or the quality of DNA extracted. These data suggest that high-quality DNA can be extracted from whole blood samples that are stored at −30°C for up to 19 years.

Keywords: DNA storage, long-term storage, DNA quality

Introduction

DNA can be isolated from almost all body tissues, including bodily fluids.[1] Due to its relative ease of acquisition, whole blood provides one of the most common sources of high-quantity and high-quality DNA for molecular applications, including clinical and epidemiological studies.[2]

Large genetic epidemiological studies can involve the collection of tens of thousands of valuable whole blood samples that can be, or have been, stored for many years. The blood volume available for DNA extractions is an obvious indicator of DNA yield.[3] However, long-term storage conditions of whole blood can affect the integrity of DNA.[4] Blood samples stored at 4°C for a short period of time will still yield DNA of acceptable quality provided the correct blood collection tubes were used.[1,5] However, the desired temperature for whole blood samples for DNA isolation is at −80°C for long-term storage.[4] Whole blood samples can also be frozen at −20°C for long-term storage,[4,6,7] however, other studies have found lower DNA yields following this approach.[1,7] A study by Di Pietro et al. found that whole blood samples stored for long periods at −20°C yielded high-quality DNA for genotyping studies, although the sample size was small ($n = 82$).[8]

The current study sought to determine whether the duration of long-term storage of whole blood samples, using a large sample size ($n = 2758$), stored at a stable −30°C, has a negative effect on the quality and quantity of the genomic DNA extracted.

Materials and Methods

Samples

Samples analyzed were sourced from the Johannesburg Cancer Case–Control Study (JCS) at the National Cancer Registry (NCR) of South Africa. The JCS is a large, recently (2016) ended case–control study established in 1995.[9] In

[1]National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa.
[2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa.
[3]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
[4]Division of Human Genetics, National Health Laboratory Service, Johannesburg, South Africa.
[5]Centre for Enteric Diseases, National Institute for Communicable Diseases, Johannesburg, South Africa.
[6]Division of Medical Biochemistry, University of Cape Town, Cape Town, South Africa.
[7]Primary Care Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom.

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

brief, the JCS recruited newly diagnosed cancer patients of black African ancestry. These patients were interviewed by trained research nurses and their blood was collected with informed consent for future research as approved by the Research Ethics Committee at the University of the Witwatersrand and described by Urban et al.[9] Blood samples were drawn from patients in 4 mL EDTA vacutainers and were stored in −30°C freezers until further use. More than 20,000 whole blood samples (representing various cancer types) have been collected since March 1995 for the JCS, of which a subset of 2758 blood samples were used for this study. DNA extractions were performed on request for specific studies in collaboration with the JCS. At the time of the study being reported here, DNA had been extracted from 2758 blood samples for four different studies. Ethical approval was obtained from the University of the Witwatersrand Human Research Ethics Committee (Medical), clearance numbers: M140271 and M120117.

Sample storage

All blood samples collected under the JCS were kept at room temperature until collected for transport. The blood samples were then transported in a cooler box with an ice block from the three separate study sites to a centralized storage facility. On the rare occasion that there was an issue with the transportation, samples were frozen at the collection site and collected at the next available time. Average transit time was ≤30 minutes. All samples were frozen in their collection tubes in an upright position at −30°C within a maximum of 24 hours of collection. Miele™ −30°C general freezers were used for sample storage. The freezers' temperatures were monitored with an alarm-based system and backup freezers were available. As part of our risk management strategy, when a freezer failed, the samples were transferred to another freezer before thawing. All the blood samples were frozen in an upright position and thereafter stored on their sides. Although the blood was collected in a 4 mL EDTA tube, not all were filled to capacity; therefore, the volume of the blood sample available for extraction was noted before DNA extraction. Sample information was managed using Microsoft Excel and then incorporated into STATA database.

DNA isolation

Whole blood samples collected between 1997 and 2012 ($n = 2758$) were extracted. DNA extraction procedures occurred in batches between May 2013 and June 2016. Six different technicians were involved over the 4-year period. The technicians were trained on the extraction protocol, blinded to the storage duration of the blood samples, and each technician extracted samples collected over the whole time period. Total genomic DNA was isolated using the FlexiGene® DNA Kit (Qiagen®). This kit was chosen initially for its ability to handle a diverse range of input blood volume and then used subsequently for consistency. Extractions were undertaken following the manufacturer's protocol, with the following minor modifications: the use of a dry heat incubator instead of a hot water bath, and resuspension of the DNA pellet in 1× Tris-EDTA buffer (Invitrogen™) instead of the supplied FG3 buffer. To ensure no wastage and optimum yield, the entire blood volume in a tube was used for DNA extraction. Two fixed protocols, one for samples with <2 mL of blood and the other for >2 mL, were optimized and used as per the manufacturer's procedures. The most noticeable difference between protocols was the use of 50 mL Nunc tubes for samples with blood volumes of >2 mL, rather than the 15 mL Nunc tubes used for samples with blood volumes of <2 mL. Reagent volumes were adjusted as per the sample volume (as per the manufacturer's protocol). The centrifugation conditions were unchanged. No RNAse digestion was incorporated into the DNA isolation protocol.

DNA quantity and quality

DNA quantity and purity were assessed spectrophotometrically using the NanoDrop™ 1000 (Thermo Scientific™) directly after extraction and before storage at −30°C. DNA concentration was measured and DNA purity was calculated through the standard A260/280 and A260/230 ratios. This is a fast and accurate technique for determining DNA concentration of pure samples.[10] DNA/RNA and proteins have their maximum absorbance at 260 and 280 nm, respectively. An A260/280 ratio of ~1.8 is generally accepted as a reflection of a pure DNA sample and a ratio of ~2.0 is generally accepted as pure for RNA. A secondary measurement of the A260/230 ratio is used to determine the purity of DNA against other contaminants. A higher A260/230 is expected for pure DNA (~2.0–2.2). Low A260/230 ratios indicate the presence of contamination or proteins that absorb light at the 230 nm wavelength.[11]

Genomic DNA integrity assessment was done, at a later stage, using gel electrophoresis for 218 samples selected at

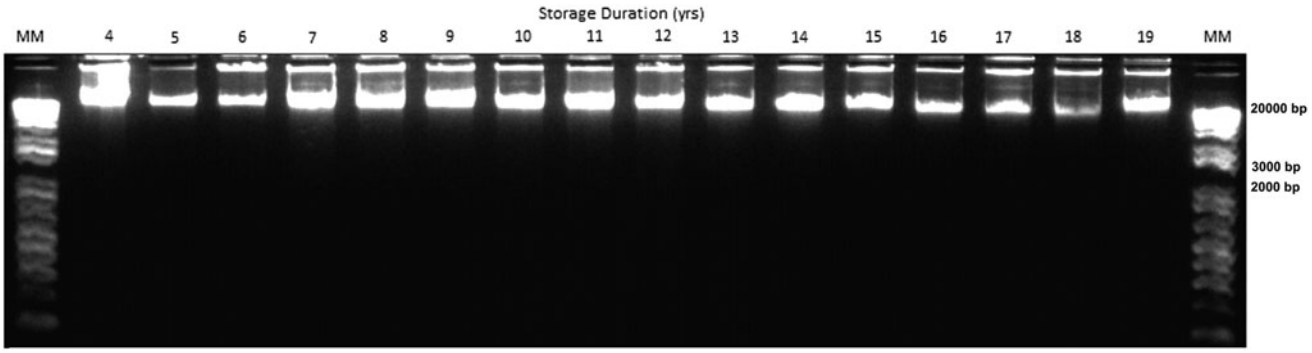


FIG. 1. Gel image showing integrity of genomic DNA from 4 to 19 years of blood sample storage. One representative sample was used from each year of storage (4–19 years). Two microliters of genomic DNA was resolved on 1.0% agarose gel with GelRed™ staining. GeneRuler™ 1 kb plus DNA Ladder MM was used for sizing reference. Electrophoresis was done at 8 V/cm for 45 minutes. MM, molecular marker.

random (4–19 years storage, mean = 12 years) (Fig. 1). Genomic DNA was resolved on 0.8% agarose gel, stained with GelRed™, at 8 V/cm for 45 minutes against the GeneRuler 1 kb Plus DNA Ladder (Thermo Scientific).

DNA quantitation was additionally assessed using a second methodology. Qubit fluorometry was performed using the Qubit™ 3.0 Fluorometer (Thermo Scientific). This method uses the double-stranded DNA (dsDNA) broad range assay to quantify dsDNA, as opposed to single-stranded DNA (ssDNA), present in a sample. Only a subset of samples ($n = 200$) was quantified using this chemistry owing to cost restraints. The 200 samples were randomly selected for year of collection (2–19 years of storage), with concentrations representative of the overall DNA concentration range. This measures dsDNA concentration only, which provides a more useful reading as contaminants, including degraded DNA and RNA, are not measured. The percentage of dsDNA (%dsDNA) was calculated by dividing the Qubit concentration by the NanoDrop concentration and multiplying by 100.

Analysis

Descriptive statistics were compiled by years of storage. Means and medians were calculated for DNA yields, DNA concentrations, blood volume available per sample, DNA yield per milliliter of blood, A260/280 and A260/230. Both means and medians were compiled given the non-normally distributed nature of the data, which may limit the mean accuracy.

Spearman's ranked correlations were performed to determine whether duration of storage (years) of the samples, blood volume available for extraction, and the age of participants correlated with the total DNA yields. Samples ($n = 564$) without blood volume data were excluded.

Kruskal–Wallis H tests, with multiple group comparisons, were performed to determine whether DNA yields, A260/280 and A260/230 ratios, differed between the different years of storage duration. Kruskal–Wallis H tests were also used to determine differences in mean DNA yield per milliliter of blood across different storage durations as well as to determine differences in mean total DNA yields measured by Qubit and differences in mean %dsDNA across different storage durations. A nonparametric test for trend was performed to evaluate the relationship between %dsDNAs across the storage duration of the blood samples. Quantile regression estimating the conditional median was performed to model the effect of storage duration of the blood samples, blood volume available for DNA extraction, and age and gender of the participants had on the %dsDNA.

A two-sample t-test with equal variances was performed to assess the difference in mean DNA yield per 1 mL of blood between the two different DNA extraction protocols as well as mean DNA yield per 1 mL of blood between cancer samples and noncancer samples and between HIV-positive and HIV-negative samples.

Results

Total genomic DNA was isolated from a total of 2758 samples collected between 1997 and 2012 for the JCS (Table 1). NanoDrop data were used to calculate DNA yield, DNA concentration, and to determine A260/280 and A260/230 ratios. Blood samples were stored on average for 11.9 years. The longest storage period of samples before DNA extraction was 19 years ($n = 125$) and the shortest was 2 years ($n = 31$) (Table 1). Information on the volume of blood collected was available for 80% ($n = 2206$) of all samples (Table 1). Freezers were monitored closely so that no thawing of samples (resulting from freezer failures) happened during this study.

The quality (A260/280 and A260/230 ratios), concentration (ng/μL), and the quantity of isolated DNA (μg), and the mean and median concentration per year of storage for the samples are shown in Table 1. The mean A260/280 ratio across all samples was 1.79 and median A260/280 ratio was 1.8, suggesting that good-quality DNA was isolated. The mean A260/230 ratio across all samples was 1.71 and the median A260/230 ratio was 1.78, <2.0, suggesting the presence of proteins, contaminants, or RNA residues that absorb at 230 nm or less.

Genomic DNA from 218 samples was subjected to agarose gel electrophoresis to assess possible degradation. High-molecular-weight staining was observed for all 218 samples. Figure 1 shows a representation of the gel electrophoresis result for a representative subset of samples, which had been stored for different periods of time. A consistent low level of partial degradation (low intensity of smearing from the genomic band down to 3 kb) was noted across most samples analyzed on gel electrophoresis. The consistency of this smearing pattern suggests that this is likely an artifact from the DNA isolation protocol that caused DNA smearing rather than storage-related DNA degradation.

The DNA yields of samples per year of storage clustered tightly together, except for a few outlier samples with very high DNA yields or extremely low DNA yield (Fig. 2). The presence of the outliers illustrates the value of the median as a more informative reporting tool than the mean.

The average yield for all samples was 108.95 μg as measured by NanoDrop, whereas the average yield for the 200 samples measured by Qubit was 86.19 μg per sample. This indicates that ~79% of total DNA measured by spectrophotometry was dsDNA, with the remaining proportion presumed to be ssDNA and/or denatured or degraded DNA. A plot of dsDNA percentage against the storage duration is shown in Figure 3.

A Kruskal–Wallis H test showed that no statistically significant differences exist between the mean total DNA yield by Qubit measurements and the number of years of storage ($p = 0.46$) and no statistically significant differences exist between the mean %dsDNA and the number of years of storage ($p = 0.10$). Nonparametric trend analysis for %dsDNA, quantified by Qubit across the storage durations, indicated a positive trend ($z = 6.24$, $p < 0.001$), with the %dsDNA increasing with increased storage duration of the blood samples. However, storage duration of the blood samples had a negligible effect on %dsDNA after adjusting for blood volume used for extraction, age, and gender ($\beta = 0.01$, $p = 0.045$).

Spearman's ranked correlation was performed to assess the relationship between total DNA yields, storage duration of blood samples, blood volumes available for extractions, and age of participants at recruitment using 2194 samples. After Bonferroni correction, there was a strong positive correlation observed between the total blood volumes

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

Table 1. Descriptive Statistics for the Samples Used for Genomic DNA Isolation Per Years of Storage

Storage duration (years)	*Sample size*	*Mean blood volume*[a] *(mL)*	*A260/280*				*A260/230*				*Mean DNA concentration (ng/μL) per sample*[b]	*Median DNA yield (μg) per sample*[b]	*Mean DNA yield/mL of blood (μg)*[b]	*Mean DNA yield (μg) per sample*[b]
			Median	*Range*	*SD*	*Mean*	*Median*	*Range*	*SD*	*Mean*				
2	31	—	1.81	1.68–1.88	0.03	1.81	1.69	1.24–2.04	0.18	1.69	305.97	110.88	—	122.39
3	25	—	1.80	1.70–1.90	0.05	1.82	1.60	0.29–2.15	0.45	1.56	266.66	98.46	—	106.66
4	54	2.1	1.74	1.52–1.90	0.08	1.73	1.70	0.76–2.13	0.30	1.63	217.16	62.95	28.81	86.86
5	84	2.6	1.79	1.24–1.91	0.08	1.76	1.80	0.36–2.31	0.39	1.69	224.90	59.10	30.70	89.96
6	102	2.6	1.79	1.51–1.95	0.06	1.78	1.88	0.91–2.26	0.34	1.80	195.36	61.60	30.05	78.14
7	212	3.0	1.80	1.51–2.25	0.06	1.81	1.79	0.29–2.58	0.37	1.73	266.64	96.05	37.20	106.66
8	185	3.3	1.81	1.27–1.97	0.06	1.81	1.66	0.26–2.78	0.39	1.68	278.78	102.72	35.40	111.51
9	198	3.1	1.81	1.58–2.18	0.07	1.81	1.66	0.81–2.56	0.42	1.68	285.70	97.68	41.60	114.28
10	218	3.1	1.80	1.62–2.05	0.06	1.81	1.65	0.57–2.41	0.40	1.66	293.74	97.21	40.62	117.50
11	203	3.1	1.80	1.48–2.15	0.07	1.80	1.69	0.30–2.62	0.46	1.68	289.54	103.96	38.67	115.82
12	194	3.1	1.80	1.26–2.31	0.09	1.80	1.64	0.35–2.34	0.41	1.60	291.38	96.57	37.60	116.55
13	172	3.1	1.80	0.71–2.09	0.11	1.78	1.63	0.19–2.60	0.44	1.62	316.77	100.73	40.87	126.71
14	210	2.9	1.80	0.84–2.18	0.10	1.79	1.69	0.23–2.43	0.42	1.61	294.15	101.16	37.04	117.66
15	113	2.9	1.80	0.32–1.89	0.16	1.77	1.95	0.22–2.43	0.43	1.85	281.76	88.59	38.87	112.71
16	339	3.1	1.79	1.02–3.30	0.11	1.79	1.78	0.22–2.44	0.41	1.71	353.03	109.14	45.55	141.21
17	103	3.0	1.80	0.24–1.91	0.16	1.78	2.07	0.51–2.44	0.33	1.98	252.63	78.42	33.68	101.05
18	192	3.1	1.80	1.10–1.96	0.06	1.80	2.08	0.81–2.51	0.25	2.04	273.24	91.02	35.26	109.30
19	124	2.9	1.80	0.56–1.96	0.15	1.77	1.81	0.08–2.49	0.55	1.67	235.57	75.86	32.49	94.23

[a]Mean blood volume per sample.
[b]Mean DNA concentrations and median and mean DNA yields are calculated using NanoDrop™ data.
SD, standard deviation.

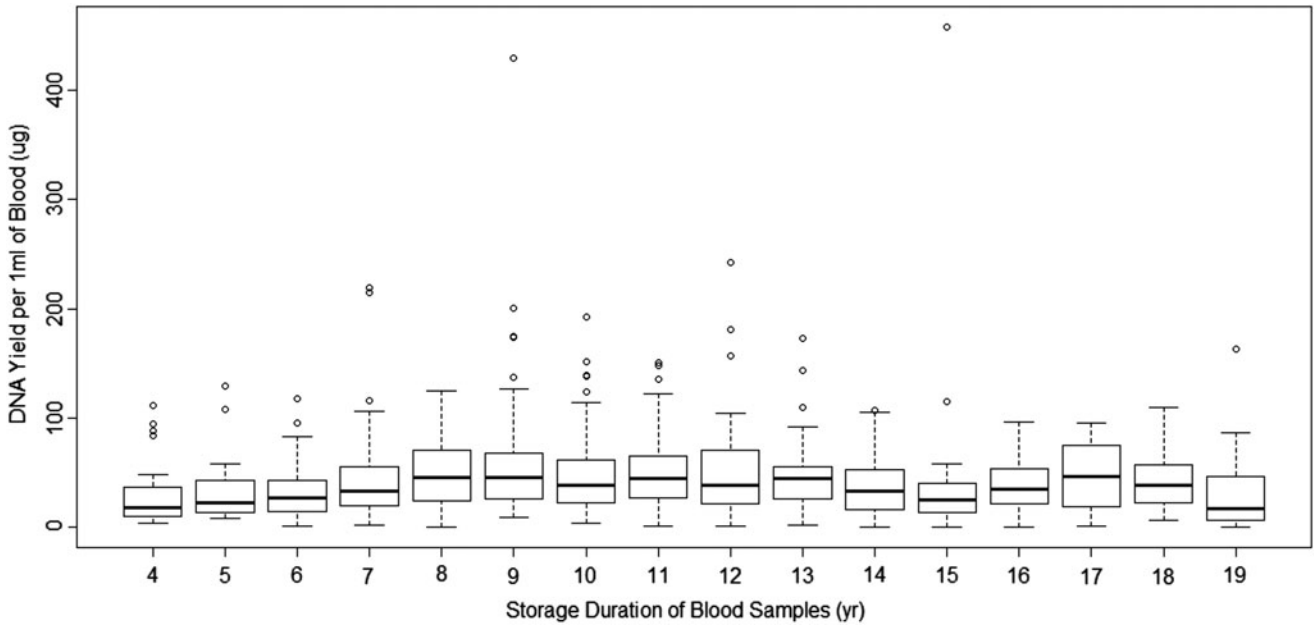


FIG. 2. Range and distribution of DNA yields per 1 mL of blood extracted from frozen whole blood stored over 4–19 years. Blood volume data were not available for 2 and 3 years of storage. No statistically significant difference in yield was observed across the duration of blood storage. DNA yields were measured by NanoDrop.™

available for extractions and the total DNA yields, which was statistically significant, $r_s = 0.2258$, $p < 0.0001$. So, the more total volume of blood available the more total DNA yield there was. There was no evidence of correlation between storage duration of blood samples or age of participants at recruitment with total DNA yields. There was no evidence of correlation between total blood volumes with A260/280 or A260/230 ratios.

A Kruskal–Wallis *H* test, after Bonferroni correction and correcting for technician variability, showed that no statistically significant differences exist between the DNA yields, A260/280 or A260/230 ratios, and the number of years of storage. As expected, total blood volume is a direct indicator of the amount of total DNA available ($r_s = 0.2258$, $p < 0.0001$). Similarly, no statistically significant differences were observed between the mean DNA yield per milliliter of blood and the storage duration of the blood sample ($p = 0.25$).

The FlexiGene DNA Kit (Qiagen) gave an expected DNA yield of 37.5 μg per 1 mL of blood assuming a normal white cell count of 7×10^6 cells/mL blood.[12] The JCS sample yield was between 25.83 and 46.54 μg (mean = 37.77 μg) of DNA per 1 mL of blood.

The effect of the optimized protocols for different sample blood volumes was assessed using a two-sample *t*-test with equal variances. Samples with blood volumes of 2 mL or less ($n = 403$) yielded more DNA per 1 mL of blood

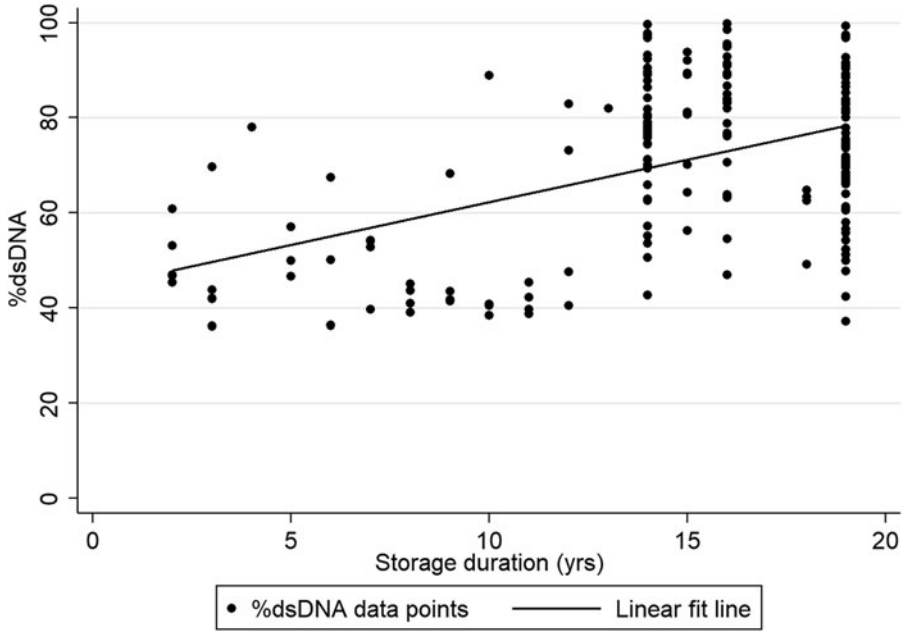


FIG. 3. %dsDNA versus storage duration. The linear fit *line* indicates that a positive trend exists between %dsDNA and the increase of storage duration. %dsDNA, percentage of double-stranded DNA.

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

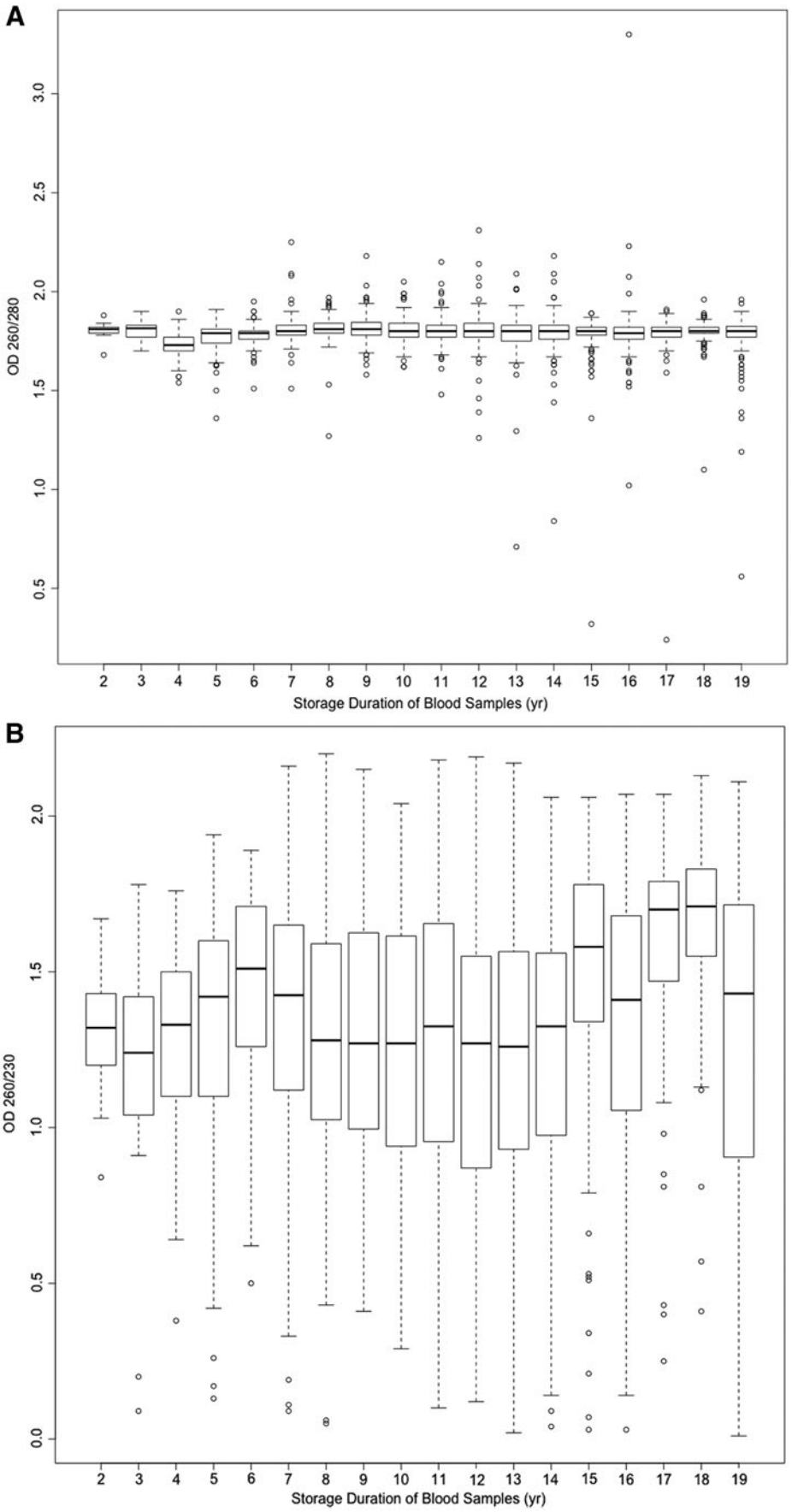


FIG. 4. **(A)** A260/280 ratio distribution of all samples across duration of blood sample storage. A260/280 ratios were measured by NanoDrop. **(B)** A260/230 ratio distribution of all samples across duration of blood sample storage. A260/230 ratios were measured by NanoDrop.

(44.12 μg, 95% CI: 39.73–48.51 μg) when compared to DNA yielded per 1 mL of blood for samples with blood volumes of >2 mL ($n=1792$) (36.79 μg, 95% CI: 35.42–38.16 μg) ($p=0.0001$). The typical DNA yields of the FlexiGene kit for a 2 mL volume would be 33–45 μg/mL. No difference in mean DNA yield per 1 mL of blood was observed between cancer samples ($n=1482$) and noncancer samples ($n=707$) ($p=0.50$). A difference in mean DNA yield per 1 mL of blood was observed between HIV-positive ($n=1851$) and HIV-negative samples ($n=313$), with HIV-negative samples (38.70 μg, 95% CI: 37.23–40.17 μg) yielding ~5 μg DNA more per milliliter than HIV-positive samples (33.36 μg, 95% CI: 29.39–37.32 μg) ($p=0.0076$).

Discussion

This study was undertaken to answer questions about the expected quality and quantity of DNA isolated from whole blood samples that have been stored at −30°C between 2 and 19 years. The concern was that long-term storage of whole blood at −30°C may have a negative effect on the quality and/or quantity of the DNA.[7] Genomic DNA was isolated from 2758 blood samples collected between 1997 and 2012 using the Qiagen FlexiGene DNA kit. To our knowledge this is the largest study of its kind assessing the quality and quantity of DNA isolated from frozen blood samples stored for long periods.

The results from our study are unique from previous findings[1,7] as we used a different DNA extraction method (kit vs. salting out or phenol/chloroform), stored samples at −30°C temperature (compared to room temperature and 4°C), and used a much larger sample size. The large sample size of our study affords greater statistical power to evaluate the relationship between DNA yields and storage duration of the blood samples. A recent publication looking at various storage conditions on forensic examinations of blood samples and blood stains ($n=6$) concluded that to prevent DNA degradation during long-term (20 year) storage, they should be stored at below −20°C,[13] as our samples have been. We have a much larger sample size with varying lengths of storage to assess and have found that the quantity of DNA is not affected, and are fairly confident that the same can be said about the quality.

This study found that long-term storage of whole blood samples at −30°C does not have a negative effect on the quality and quantity of the DNA extracted. The average duration of storage was ~12 years. The mean A260/280 ratios were found to be ~1.8 across all the years of collection, suggesting high purity of the DNA regardless of the duration of blood storage (Fig. 4A). The lower than expected A260/230 ratios indicate the presence of other organic residues (Fig. 4B). The source of the organic residues is likely artifacts resulting from the inadequate removal of organic chemicals used in the extraction chemistry.

The DNA yield was expected to decrease for samples stored for longer periods. Alrokayan reported a 30%–40% decrease in total DNA yield for whole blood samples frozen at −20°C.[7] A decrease in DNA yield was not observed in the JCS samples extracted, as no correlation between storage duration and total DNA yields was observed. Assuming the blood samples are stored within the desired temperature range at stable conditions, DNA can be extracted from these samples after prolonged storage. DNA yields of samples per year of collection clustered tightly together for the 2–19 years, except for a few outliers (Fig. 2).

High-quality dsDNA is preferred for downstream applications such as polymerase chain reaction (PCR), microarray analysis, or next-generation sequencing analysis. Qubit is well known to be a more sensitive method to quantify DNA, compared to the NanoDrop.[14,15] Qubit readings are lower than NanoDrop readings as they do not measure single-stranded, degraded DNA.[14,15] Our results showed that, on average, 79% of total DNA measured by NanoDrop was dsDNA. From the subset of samples analyzed on Qubit, the %dsDNA did not decrease with increased years of storage ($p=0.10$); instead, a positive trend was seen with increasing storage duration but the effect size was minimal ($\beta=0.01$, $p=0.045$). This supports our hypothesis that storage duration of the blood samples does not influence DNA quality. It is worth noting that although DNA was extracted from these archived blood samples, no assessments were done for RNA, micro-RNA, cell-free circulating DNA, or proteins.

The DNA yields are in line with expected yields cited by the manufacturers. With every additional 1 mL of blood volume available for extraction, the DNA yield would be expected to increase by 20.94 μg ($p<0.01$). The DNA yield per milliliter of blood does not increase when more blood was processed. Rather, optimization of the DNA extraction protocol used for different volumes of blood affects the DNA yield per milliliter of blood. Two fixed protocols were used in this study, the protocol for sample volumes of <2 mL of blood yielded ~7 μg ($p=0.0001$) more DNA per milliliter of blood than the protocol used for samples of >2 mL of blood. This is possibly due to the extra time and care given per extraction for the small volumes of these samples because of the limited volume. It is possible that the potential yield could be improved by optimizing the extraction protocol to the exact blood volume available. This would, however, require extra laboratory time and staff.

The patients' age may also influence the number of white cells available.[5,16] In older patients there is often a decline in the quantity of DNA due to a reduction in the number of leukocytes and lymphocytes with age progression.[17] However, this correlation was not observed in this study. As anticipated, samples from individuals who were HIV positive at the time of sampling had a lower DNA yield as a likely consequence of fewer peripheral blood mononuclear cells due to infection. However, no difference in mean yield was observed between cancer samples and noncancer samples.

One limitation of this study was that gel electrophoresis and fluorometry were not done for all 2758 samples to assess the integrity of the extracted DNA at the beginning of DNA storage. Hence the data lack robustness to definitively conclude that long-term storage of blood samples does not negatively affect DNA integrity. However, from the representative subset of samples ($n=200$) that were assessed by gel electrophoresis, the authors are confident that the DNA integrity of DNA extracted from long-term stored blood samples is sound. A comprehensive analysis would include doing a quantitative PCR or a long-range PCR and we have applied for funding to do this. We have used these samples in other genotyping studies and they have produced adequate genotyping call rates (unpublished).

From the results of this large study, the authors are confident that long-term storage (average 11.9 years) of whole blood samples stored in 4 mL EDTA vacutainers at a stable −30°C does not have a negative effect on the quality and quantity of the DNA. The authors are also confident that the DNA integrity assessment done on a representative sample

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

is sufficient to represent the larger cohort. Despite the limitations of this study, we found no evidence that DNA integrity worsened over a prolonged period of blood sample storage. Archived whole blood samples can still be utilized in genetic epidemiology studies provided the blood samples are stored correctly with constant temperature monitoring to prevent freeze/thaw episodes.

Acknowledgments

The authors acknowledge Sisters Gloria Mokwatle, Patricia Rapoho, and Pheladi Kale who carried out the interviews and collected blood specimens from the patients included in this analysis, Mrs. Lettie Bester who prepared the specimens for testing and storage, and Mrs. Margaret Urban and Mr. Tonicah Maphanga for maintaining the JCS database. They thank Dr. Mazvita Sengayi for statistical analysis support and Ms. Cassandra Soo and Ms. Natalie Smyth for fluorescent DNA quantification. They also thank the oncology clinicians and the administration at Charlotte Maxeke Johannesburg Academic Hospital (formerly Johannesburg General Hospital) for assistance and for access to patients in their care, and the patients who gave freely of their time. Research reported in this publication and ongoing extraction of DNA from JCS blood samples are supported by the Grants, Innovation and Product Development (GIPD) Unit of the South African Medical Research Council with funds received from the South African National Department of Health and the U.K. Medical Research Council from the U.K. Government's Newton Fund. The authors thank the DST/NRF internship program for supporting B.N., A.S., S.T.D., and L.J. while interns at CERG. They thank the DAAD/NRF in country scholarship for a fellowship to W.C.C., and the Cancer Association of South Africa (CANSA) for research cost support to C.G.M. The contents are solely the responsibility of the authors and do not necessarily reflect the views of the funding bodies.

Authors' Contributions

W.C.C.: substantial contributions to conception and design, acquisition of data (DNA extraction), analysis and interpretation of data, and final approval of the version to be published. R.K.: interpretation of data, revising the article critically for important intellectual content, and final approval of the version to be published. A.M.: analysis and interpretation of data, revising the article critically for important intellectual content, and final approval of the version to be published. B.N., A.S., S.T.D., and L.J.: acquisition of data (DNA extraction), maintenance of the database and summary report (base of manuscript), and final approval of the version to be published. C.G.M.: interpretation of data, revising the article critically for important intellectual content, and final approval of the version to be published. C.B.d.V. conceived the project, substantial contributions to conception and design, interpretation of data, revising the article critically for important intellectual content, and final approval of the version to be published.

Author Disclosure Statement

No conflicting financial interests exist.

References

1. Nederhand RJ, Droog S, Kluft C, et al. Logistics and quality control for DNA sampling in large multicenter studies. J Thromb Haemost 2003;1:987–991.
2. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988;16:1215.
3. Chacon-Cortes D, Griffiths LR. Methods for extracting genomic DNA from whole blood samples: Current perspectives. J Biorepos Sci Appl Med 2014;2:1–9.
4. Holland NT, Smith MT, Eskenazi B, et al. Biological sample collection and processing for molecular epidemiological studies. Mutat Res 2003;543:217–234.
5. Richardson AJ, Narendran N, Guymer RH, et al. Blood storage at 4 degrees celsius—Factors involved in DNA yield and quality. J Lab Clin Med 2006;147:290–294.
6. Steinberg KK, Sanderlin KC, Ou CY, et al. DNA banking in epidemiologic studies. Epidemiol Rev 1997;19:156–162.
7. Alrokayan SAH. Effect of storage temperature on the quality and quantity of DNA extracted from blood. J Biol Sci 2000;3:392–394.
8. Di Pietro F, Ortenzi F, Tilio M, et al. Genomix DNA extraction from whole blood stored from 15- to 30-years at –20°C by rapid phenol-chloroform protocol: A useful tool for genetic epidemiology studies. Mol Cell Probes 2011;25:44–48.
9. Urban M, Banks E, Egger S, et al. Injectable and oral contraceptive use and cancers of the breast, cervix, ovary, and endometrium in Black South African women: Case–control study. PLoS Med 2012;9:e1001182.
10. Desjardins P, Conklin D. NanoDrop microvolume quantitation of nucleic acids. J Vis Exp 2010. DOI: 10.3791/2565.
11. Thermo Scientific™. Nanodrop: Assessment of nucleic acid purity. Technical Note 52646, 2015;1–3.
12. Qiagen®. FlexiGene® DNA Handbook. 2010. Available at: www.qiagen.com (accessed March 21, 2016).
13. Hara M, Nakanishi H, Yoneyama K, et al. Effects of storage conditions on forensic examinations of blood samples and bloodstains stored for 20 years. Leg Med (Tokyo) 2016;18:81–84.
14. Nakayama Y, Yamaguchi H, Einaga N, et al. Pitfalls of DNA quantification using DNA-binding fluorescent dyes and suggested solutions. PLoS One 2016;11:e0150528.
15. Robin JD, Ludlow AT, LaRanger R, et al. Comparison of DNA quantification methods for next generation sequencing. Sci Rep 2016;6:24067.
16. Mychaleckyj JC, Farber E, Chmielewski J, et al. Buffy coat specimens remain viable as a DNA source for highly multiplexed genome-wide genetic tests after long term storage. J Transl Med 2011;9:91.
17. Erkeller-Yuksel FM, Deneys V, Yuksel B, et al. Age-related changes in human blood lymphocyte subpopulations. J Pediatr 1992;120:216–222.

Address correspondence to:
Wenlong Carl Chen, MSc (Med)
National Cancer Registry
National Health Laboratory Service
Room F50 STIRC Building
1 Modderfontein Drive
Sandringham
Johannesburg 2010
South Africa

E-mail: wenlongc@nicd.ac.za

Downloaded by Kings College London-journal Section from www.liebertpub.com at 04/26/18. For personal use only.

Carcinogenesis, 2019, Vol. 40, No. 4, 513–520

doi:10.1093/carcin/bgz026
Advance Access Publication February 12, 2019
Original Article

ORIGINAL ARTICLE

Association of genetic variants in *CHEK2* with oesophageal squamous cell carcinoma in the South African Black population

Wenlong C.Chen[1,2,3,†,*], Hannah Bye[4,†], Marco Matejcic[5], Ariella Amar[4], Dhiren Govender[6], Yee Wen Khew[4], Victoria Beynon[4], Robyn Kerr[3], Elvira Singh[1,7], Natalie J.Prescott[4], Cathryn M.Lewis[4,8], Chantal Babb de Villiers[3], M.Iqbal Parker[5] and Christopher G.Mathew[2,3,4,*]

[1]National Cancer Registry, National Health Laboratory Service, Johannesburg 2131, South Africa, [2]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg 2000, South Africa, [3]Division of Human Genetics, School of Pathology, National Health Laboratory Service and University of the Witwatersrand, Johannesburg 2000, South Africa, [4]Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London SE1 9RT, United Kingdom, [5]Division of Medical Biochemistry and Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town 7925, South Africa, [6]Division of Anatomical Pathology, University of Cape Town, Cape Town 7925, South Africa, [7]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg 2000, South Africa and [8]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London SE5 8AF, United Kingdom

*To whom correspondence should be addressed. Tel: +44 20 7848 8509; Email: christopher.mathew@kcl.ac.uk
†These authors contributed equally to this work.

Abstract

Oesophageal squamous cell carcinoma (OSCC) has a high incidence in southern Africa and a poor prognosis. Limited information is available on the contribution of genetic variants in susceptibility to OSCC in this region. However, recent genome-wide association studies have identified multiple susceptibility loci in Asian and European populations. In this study, we investigated genetic variants from seven OSCC risk loci identified in non-African populations for association with OSCC in the South African Black population. We performed association studies in a total of 1471 cases and 1791 controls from two study sample groups, which included 591 cases and 852 controls from the Western Cape and 880 cases and 939 controls from the Johannesburg region in the Gauteng province. Thereafter, we performed a *meta*-analysis for 11 variants which had been genotyped in both studies. A single nucleotide polymorphism in the *CHEK2* gene, rs1033667, was significantly associated with OSCC [$P = 0.002$; odds ratio (OR) = 1.176; 95% confidence interval (CI): 1.06–1.30]. However, single nucleotide polymorphisms in the *CASP8/ALS2CR12*, *TMEM173*, *PLCE1*, *ALDH2*, *ATP1B2/TP53* and *RUNX1* loci were not associated with the disease ($P > 0.05$). The lack of association of six of these loci with OSCC in South African populations may reflect different genetic risk factors in non-African and African populations or differences in the genetic architecture of African genomes. The association at *CHEK2*, a gene with key roles in cell cycle regulation and DNA repair, in an African population provides further support for the contribution of common genetic variants at this locus to the risk of oesophageal cancer.

Received: August 24, 2018; **Revised:** December 18, 2018; **Accepted:** February 7, 2019

© The Author(s) 2019. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

Abbreviations

eQTL	expression quantitative trait locus
GxE	gene x environmental
GWAS	genome-wide association studies
JCS	Johannesburg Cancer Study
LD	linkage disequilibrium
MAF	minor allele frequency
OC	oesophageal cancer
OSCC	oesophageal squamous cell carcinoma
SAB	South African Black
SAP	shrimp alkaline phosphatase
SNP	single nucleotide polymorphism
UCT	University of Cape Town

Introduction

Oesophageal cancer (OC) is the eighth most common cancer worldwide (1), with oesophageal squamous cell carcinoma (OSCC) accounting for nearly 90% of all cases of OC worldwide (2). Particularly high incidence rates are observed in regions such as southern Africa, China and Japan. In South Africa, OC was the third most common cancer in males of African ancestry and the fourth most common cancer in females of African ancestry in 2014, with age standardized incidence rates of 4.79 and 3.16 per 100 000, respectively (3). In addition to the recognized risk factors for OSCC of alcohol consumption and tobacco use, polycyclic aromatic hydrocarbons exposure, high food temperatures and other factors such as diet and consumption of *Fusarium*-contaminated maize may also be involved in the development of the disease (reviewed in (4–7)).

Over the past 9 years, several genome-wide association studies (GWAS) have been performed in non-African populations to determine the genetic risk factors involved in the development of OSCC. In 2009, a GWAS for OSCC in the Japanese population found that single nucleotide polymorphisms (SNPs) at *ADH1B* (rs1229984) and *ALDH2* (rs671) were associated with the disease (8), and a European GWAS of upper aerodigestive tract cancers detected genome-wide significant association with OSCC at *ALDH2* (rs4767364) (9). Since 2010, four OSCC GWAS and a *meta*-analysis have been performed in the Chinese population, which identified multiple additional risk loci (10–14). These included SNPs at *PLCE1*, *HEATR3*, *HAP1*, *CHEK2/XBP1*, *ST6GAL1*, *SMG6*, *PTPN2*, *SLC52A3* (*C20orf54*), *PDE4D*, *RUNX1*, *UNC5CL* and *CYP26B1*. However, in a subsequent analysis, the associations at *HAP1*, *SMG6*, *SLC52A3* (*C20orf54*) and *PTPN2* were not confirmed (14).

We previously tested several of these variants for association with OSCC in the South African Black (SAB) population recruited in the Western Cape region (15,16). Coding SNPs in *ADH1B* and *ALDH2* were absent in the SAB population (15), and none of the lead SNPs from the Chinese GWAS studies were associated with OSCC (*PLCE1* rs2274223, *SLC52A3/C20orf54* rs13042395, *PDE4D* rs10052657, *RUNX1* rs2014300, *UNC5CL* rs10484761). However, further analysis of the *PLCE1* locus identified a significant association with the SNP Arg548Leu (rs17417407).

The Johannesburg Cancer Study (JCS) was established in 1995, with the aim of recruiting Black, treatment naïve, patients diagnosed with cancer from the greater Johannesburg area in Gauteng province (17). The JCS recruited all cancer cases, including OSCC, as well as non-cancer, ethnically matched controls from three tertiary academic hospitals in the greater Johannesburg region. The University of Cape Town (UCT) cancer study was established to recruit OSCC patients from the Western Cape region of South Africa through tertiary hospitals (15,18). Here, we investigate the association of seven main-effect susceptibility loci (*CASP8/ALS2CR12*, *TMEM173*, *PLCE1*, *ALDH2*, *ATP1B2/TP53*, *RUNX1* and *CHEK2/XBP1*) identified by GWAS (9–13) with OSCC in SAB populations recruited in these two regions of South Africa.

Materials and methods

Study subjects

The UCT study sample consisted of 591 SAB OSCC cases and 852 ethnically matched non-cancer controls that were available for genotyping. The cases and controls were mainly Xhosa-speakers (97% and 95%, respectively) from the Western or Eastern Cape provinces of South Africa. The JCS samples consisted of 880 SAB OSCC cases and 939 ethnically matched non-cancer controls available for genotyping. The JCS cases and controls were recruited from the greater Johannesburg area in Gauteng province, with self-reported African ancestry. The major linguistic groups in the JCS in both cases and controls were the Nguni (Zulu and Xhosa speakers, 37.8%) and Sotho-Tswana (Southern Sotho, Tswana and Northern Sotho speakers, 39.2%). Control individuals for both study samples had no history of cancer, lived in the same residential areas and had a similar socioeconomic status to the cases. The JCS controls were recruited mainly from the cardiovascular unit at the Charlotte Maxeke Academic Hospital in Johannesburg, a high proportion of whom consumed alcohol.

All patients had a histologically confirmed primary invasive OSCC and were recruited between 1995 and 2016. Smoking status was subdivided into ever-smokers (those who had smoked at some point in their lives) or never-smokers. Drinkers were defined as subjects who consumed alcohol at least once every week or non-drinkers. Whole blood samples were collected, with informed consent, from all subjects and DNA was extracted at the University of Cape Town and the University of the Witwatersrand as previously described (15,19). Ethical approval for the study was obtained from the joint University of Cape Town/Groote Schuur Hospital Research Ethics Committee and the University of Stellenbosch/Tygerberg Hospital Ethics Committee (UCT HREC 040/2005) and the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (M140271, M160807).

SNP selection and genotyping

For the UCT samples, 12 SNPs were selected from the 7 OSCC susceptibility loci described above and genotyped. The list of these SNPs and loci are shown in Table 2, the majority of which were the lead SNPs from published GWAS (9–14). The index SNP for the *TMEM173* locus on chromosome 5q31.2, rs7447927 (14) failed TaqMan assay design; a proxy for this SNP, rs13153461, was genotyped which is in strong linkage disequilibrium (LD) with rs7447927 in both Chinese ($r2 = 0.98$) and African ($r2 = 0.69$ in Yoruban) populations in the 1000 Genome Project Phase 3. The functional SNP in the *TP53* gene (rs1800371, p.P47S), which has only been detected in African populations (20), was also genotyped. For the JCS samples, the lead SNPs from the seven OSCC susceptibility loci and one or more tagging SNPs in LD with the lead SNP were genotyped to provide additional coverage in case of genotype assay failure. Tag SNP selection was based on allele frequencies and LD in the Yoruban (YRI) and Chinese (CHB) populations from the 1000 Genomes Consortium (21), using LDLink 3.0 (22), which resulted in selection of 18 SNPs for genotyping (Table 2). An overview of the study design is shown in Figure 1.

The Agena MassARRAY iPLEX genotyping assay (Agena Bioscience, San Diego) was used to genotype 18 SNPs in the JCS samples. The assay was performed using the iPLEX assay protocol (23). Oligonucleotide primers were designed using the Agena Assay Design Suite (Agena Bioscience). Reactions for the PCR assays were carried out in 5.0 µl volumes in 96-well plates. Each reaction contained 25 ng DNA, 0.5 µl 10X PCR buffer, 2 mM $MgCl_2$, 500 µM deoxynucleoside triphosphate mix, 100 nM primer mix and 1.0 U Taq polymerase, and distilled water (dH_20) made up the remaining volume. Reactions for the shrimp alkaline phosphatase (SAP) post-PCR clean-up were carried out by adding 2.0 µl SAP mix, each containing 0.17 µl 10X SAP buffer, 0.51U SAP enzyme and dH_20 to each reaction well. Reactions for the iPLEX single-base extension were carried out by adding 4.0 µl iPLEX mix, each containing 0.2 µl 10X iPLEX buffer plus, 0.2 µl iPLEX termination mix, 8 µM primer mix, 0.041 µl iPLEX enzyme

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

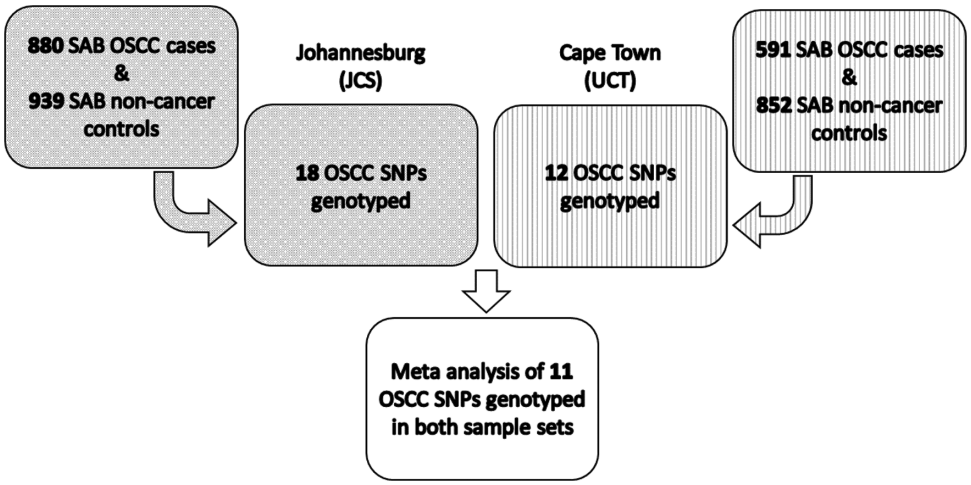


Figure 1. Schematic of the overall study design.

and dH_20 to each reaction well. Primer extension products were spotted onto the SpectroCHIPs and detection of the primer extension products by mass spectrometry was done on the Agena MassARRAY Compact mass spectrometer. All but one of 18 SNPs (rs1642764) produced discrete genotype clusters in the iPLEX assay.

The 12 SNPs for the UCT samples were genotyped using TaqMan SNP assays (Life Technologies, Carlsbad). A TaqMan SNP assay was also used to genotype the SNP (rs1642764) in the JCS samples that had failed in the iPLEX assay. Reactions for the TaqMan SNP assays were carried out in 2.5 µl volumes in 96-well plates. Each reaction contained 20 ng DNA, 1.25 µl ABsolute QPCR ROX mix (Thermo Scientific, Massachusetts) and 0.03 µl 40X TaqMan SNP assay mix (Life Technologies), with the PCR performed on a PTC-0225 DNA Engine (MJ Research, Hampton) according to manufacturer's protocol. Fluorescent levels at the PCR endpoint were determined using a 7900HT Fast Real-Time PCR system (Applied Biosystems, Foster City) and genotypes assigned using SDS 2.2.2 software (Applied Biosystems), with additional manual checks for discrete genotype clusters. High levels of concordance of genotyping results between MassARRAY chemistry and TaqMan chemistry have been described previously, including for clinical implementation (24,25).

Pearson's chi-square (χ^2) test was used to determine deviations from the Hardy–Weinberg equilibrium in controls only, using a cut-off of $P < 0.001$. Call rates for all other SNPs genotyped by either method were >95%.

Statistical analysis

In order to test for association with OSCC in the UCT and JCS study samples, allele frequencies in cases and controls were compared using Pearson's χ^2 test using Plink (version 1.9, https://www.cog-genomics.org/plink/1.9/). Allelic odds ratios (OR) and 95% confidence intervals (CI) were calculated using the common allele as the reference. In view of the differences in the ethno-linguistic composition of the subjects in the UCT and JCS samples and the different genotyping methodology used, a *meta*-analysis of the two data sets for fixed effects and random effects was performed using Plink (version 1.9, https://www.cog-genomics.org/plink/1.9/). A Bonferroni-corrected *P*-value of <0.0045 (0.05/11) was used as a significance threshold to account for the testing of 11 variants in the *meta*-analysis. The proportion of SNPs showing the same direction of effect in the *meta*-analysis as compared to the direction of effect observed in non-African studies was tested using the exact binomial test for difference from 0.5. As secondary analyses, dominant and recessive models were also tested for association with each SNP.

The power of the association tests was determined using Quanto (v1.2.4, http://hydra.usc.edu/gxe/). The average effect size of the loci tested was 1.26 and a range of minor allele frequency (MAF) for the SNPs was used for the power calculation. The power of the UCT study sample to detect an effect size of 1.26 for MAFs of 0.1–0.5 at alpha 0.05 was 48–86%. The power of the JCS samples to detect an effect size of 1.26 for MAFs of 0.05–0.45 at alpha 0.05 was 36–94%. The power of the combined study samples (1471 cases and 1791 controls) in the *meta*-analysis to detect an effect size of 1.26 was calculated for the range of MAFs observed in the combined study sample (0.17–0.46) and was 78–96% at alpha 0.0045.

Gene x environmental (GxE) interactions with tobacco smoking were investigated for SNPs with significant evidence of allelic association ($P < 0.0045$) by testing for association in cases and controls in the combined UCT/JCS study samples stratified by smoking status (ever versus never) in a χ^2 analysis using Stata MP 15.1 (StataCorp, College Station). GxE interactions for alcohol consumption were investigated in a case-only analysis of the combined UCT/JCS study samples. Case-control interactions were not tested owing to the high level of alcohol use in the JCS controls.

Bioinformatic analysis of associated variants to look for potential functional effects was carried out using established tools and resources such as HaploReg (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php), GTEx (https://gtexportal.org), RegulomeDB (http://www.regulomedb.org), ENCODE (https://www.encodeproject.org) and FuncPred (https://snpinfo.niehs.nih.gov/snpinfo/snpfunc.html).

Results

The characteristics of the cases and controls in the JCS and UCT sample sets are shown in Table 1.

The mean age of diagnosis was similar in the JCS and UCT groups (58.2 and 60.2 years, respectively). The male to female ratio was 1.64 in the JCS cases but 0.93 in the UCT sample. Smoking rates were high (>60%) in both groups of cases. The proportion of cases reporting alcohol consumption was higher in the UCT group.

Three of the 12 variants tested, under the allelic additive model, in the SAB UCT study sample were significantly associated with OSCC after Bonferroni multiple testing correction ($P < 0.0042$) (Table 2). These were *XBP1* rs2239815, *CHEK2* rs4822983 and *CHEK2* rs1033667. The other nine SNPs in *CASP8*, *ALS2CR12*, *TMEM173*, *PLCE1*, *ATP1B2/TP53*, *TP53* and *RUNX1* were not significantly associated with the disease, although rs17417407 at *PLCE1* showed nominal evidence of association. In the secondary analysis of the SAB UCT study sample, results under a dominant and a recessive model of inheritance were similar to the additive model (Supplementary Table 1, available at *Carcinogenesis* Online), and all SNPs showing association with OSSC in the dominant or recessive models were also associated in the additive model. None of the 18 variants tested in the SAB JCS study sample showed evidence of

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

Table 1. Characteristics of OSCC cases and controls in the SAB populations

Study Samples		JCS		UCT	
		Case	Control	Case	Control
Total number		880	939	591	852
Age, mean years (SD)		58.2 (10.2)	50.0 (15.5)	60.2(11.3)	48.9 (16.8)
Sex, *n* (%)	Male	545 (61.9)	240 (25.6)	284 (48.1)	342 (40.1)
	Female	332 (37.7)	698 (74.3)	307 (51.9)	507 (59.5)
	Unknown	3 (0.3)	1 (0.1)	0 (0)	3 (0.4)
Smoking status, *n* (%)	Ever-smoker	598 (68.0)	333 (35.5)	364 (61.6)	338 (39.7)
	Never-smoker	276 (31.4)	584 (62.2)	227 (38.4)	504 (59.1)
	Unknown	6 (0.7)	22 (2.3)	0 (0)	10 (1.2)
Alcohol consumption, *n* (%)	Drinker	473 (53.8)	633 (67.4)[a]	370 (62.6)	458 (53.7)
	Non-drinker	404 (45.9)	305 (32.5)[a]	219 (37.1)	390 (45.8)
	Unknown	3 (0.3)	1 (0.1)[a]	2 (0.3)	4 (0.5)

[a]Ascertainment bias of individuals with higher baseline alcohol consumption levels.

Table 2. Case-control association study analysis for SNPs genotyped in seven known risk loci for OSCC in the SAB populations

Chr	Gene	SNP	Alleles	UCT				JCS			
				MAF case	MAF control	*P*-value	OR (95% CI)	MAF case	MAF control	*P*-value	OR (95% CI)
2	*CASP8*	rs10931936	C/T	0.19	0.20	0.391	0.90 (0.71–1.15)	0.22	0.20	0.271	1.09 (0.93–1.29)
2	*ALS2CR12*	rs13016963[a]	G/A	0.35	0.35	0.994	1.00 (0.82–1.22)	0.39	0.38	0.515	1.05 (0.91–1.20)
2	*ALS2CR12*	rs10201587[a]	A/G	–	–	–	–	0.38	0.39	0.783	0.98 (0.86–1.12)
5	*TMEM173*	rs13181561	A/G	–	–	–	–	0.48	0.49	0.275	0.93 (0.82–1.06)
5	*TMEM173*	rs13153461	G/A	0.04	0.05	0.357	0.83 (0.56–1.23)	–	–	–	–
10	*PLCE1*	rs17417407	G/T	0.17	0.21	0.014*	0.76 (0.60–0.95)	0.19	0.19	0.881	0.99 (0.84–1.17)
10	*PLCE1*	rs7084339	G/A	–	–	–	–	0.48	0.46	0.178	1.09 (0.96–1.25)
10	*PLCE1*	rs3765524[a]	T/C	0.47	0.47	0.854	1.02 (0.83–1.26)	0.48	0.46	0.186	1.09 (0.96–1.24)
10	*PLCE1*	rs2274223[a]	A/G	0.42	0.40	0.508	1.06 (0.89–1.26)	0.41	0.43	0.334	0.94 (0.82–1.07)
10	*PLCE1*	rs11187850	A/G	–	–	–	–	0.21	0.19	0.204	1.11 (0.94–1.31)
12	*ALDH2*	rs4767364[a]	A/G	–	–	–	–	0.12	0.11	0.138	1.17 (0.95–1.44)
17	*ATP1B2/TP53*	rs1642764[a]	C/T	0.21	0.20	0.527	1.07 (0.86–1.33)	0.18	0.18	0.982	1.00 (0.85–1.19)
17	*ATP1B2/TP53*	rs1641511	A/G	–	–	–	–	0.39	0.42	0.081	0.89 (0.78–1.02)
17	*TP53*	rs1800371[a]	G/A	0.02	0.03	0.145	0.67 (0.38–1.16)	0.03	0.02	0.117	1.38 (0.92–2.07)
21	*RUNX1*	rs2014300[a]	A/G	0.38	0.40	0.376	0.92 (0.77–1.10)	0.36	0.36	0.795	1.02 (0.89–1.17)
21	*RUNX1*	rs2834718	T/A	–	–	–	–	0.33	0.33	0.667	0.97 (0.84–1.11)
22	*CHEK2*	rs4822983[a]	C/T	0.46	0.39	0.001***	1.32 (1.12–1.56)	0.43	0.42	0.836	1.01 (0.89–1.16)
22	*CHEK2*	rs1033667[a]	C/T	0.44	0.38	0.002**	1.30 (1.10–1.53)	0.42	0.39	0.145	1.10 (0.97–1.26)
22	*XBP1*	rs2239815[a]	C/T	0.21	0.16	0.001***	1.41 (1.15–1.74)	0.16	0.18	0.162	0.88 (0.74–1.05)

MAF, minor allele frequency; C/T, major allele/minor allele; OR, odds ratio, for each minor allele carried; 95% CI, 95% confidence interval.
[a]Previously published lead SNPs (9–14); other SNPs tag the lead SNPs.
*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$.

association with OSCC ($P > 0.05$) under the allelic additive model (Table 2) or the dominant and recessive models (Supplementary Table 1, available at *Carcinogenesis* Online).

A fixed-effect *meta*-analysis was performed for the 11 SNPs genotyped in both study samples under the allelic additive model (Table 3). One SNP, *CHEK2* rs1033667 [$P = 0.002$, OR = 1.18 (1.06–1.30)], was significantly associated with OSCC under this model. This was also the only SNP with evidence of association when analysed under dominant or recessive models (Supplementary Tables 2 and 3, available at *Carcinogenesis* Online). A random-effect *meta*-analysis was also performed to investigate potential heterogeneity between the two studies (Table 3). The only marker with any evidence of association with OSCC under the random-effect model was rs1033667 [$P(R) = 0.032$], and the direction of the effect was the same in both studies. Three variants showed evidence of heterogeneity (*TP53* rs1800371, *CHEK2* rs4822983 and *XBP1* rs2239815), but none of these were associated with OSCC under the random-effect model. Of the 11 variants included in the *meta*-analysis, 9 were genotyped in the Chinese genome-wide scans and 5 had effects in the same direction as in the original Chinese studies ($P = 0.556$).

A comparison of allele frequencies of the 19 SNPs of interest in the UCT and JCS controls with their allele frequencies in the 1000 Genomes project phase 3 populations, Yoruban (YRI), Han Chinese (CHB) and Northern European (CEU), is shown in Figure 2 and Supplementary Table 4, available at *Carcinogenesis* Online. The majority of the SNPs were common in all these populations, but allele frequencies were very different for SNPs such as *TMEM173* rs13153461, which had low frequencies in African populations. In such cases, there would be low power to detect an association of this marker with OSCC in Africans. Also, for SNPs such as *ALDH2* rs4767364 and *XBP1* rs2239815,

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

Table 3. *Meta*-analysis of 11 variants genotyped in the SAB populations

Chr	Gene	SNP	Major allele	Minor allele	*P*	*P*(*R*)	OR (95% CI)	Q	I^2
2	*CASP8*	rs10931936	C	T	0.663	0.886	1.030 (0.902–1.177)	0.186	42.91
2	*ALS2CR12*	rs13016963	G	A	0.588	0.588	1.031 (0.923–1.152)	0.726	0.00
10	*PLCE1*	rs17417407	G	T	0.116	0.310	0.898 (0.786–1.027)	0.061	71.57
10	*PLCE1*	rs3765524	T	C	0.221	0.221	1.072 (0.959–1.197)	0.588	0.00
10	*PLCE1*	rs2274223	A	G	0.708	0.780	0.980 (0.882–1.089)	0.267	18.70
17	*ATP1B2/TP53*	rs1642764	C	T	0.682	0.682	1.028 (0.899–1.176)	0.627	0.00
17	*TP53*	rs1800371	G	A	0.685	0.962	1.070 (0.772–1.483)	0.037*	77.10
21	*RUNX1*	rs2014300	A	G	0.740	0.740	0.982 (0.882–1.094)	0.390	0.00
22	*CHEK2*	rs4822983	C	T	0.028*	0.285	1.123 (1.013–1.245)	0.015*	83.03
22	*CHEK2*	rs1033667	C	T	0.002**	0.032*	1.176 (1.060–1.304)	0.137	54.85
22	*XBP1*	rs2239815	C	T	0.313	0.648	1.071 (0.938–1.758)	0.001***	91.30

OR, odds ratio, for each minor allele carried; 95% CI, 95% confidence interval; *P*(*R*), *P*-values for the random effect *meta*-analysis; Q, Cochrane's Q-statistic *P*-values. I^2, I^2 heterogeneity index; *P*, *P*-value for the fixed effect *meta*-analysis.
*P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001.

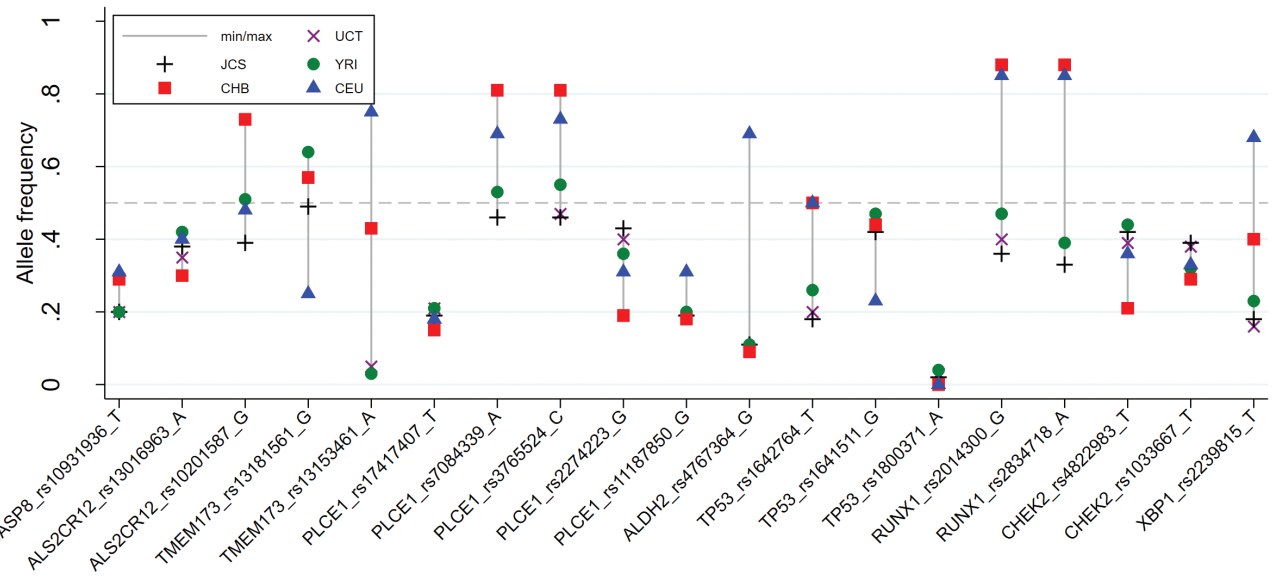


Figure 2. Comparison of the allele frequencies of the 19 SNPs of interest by population.

the rarer allele in one population was the common allele in other populations. Notably, the SNP with the strongest evidence of association with OSCC, *CHEK2* rs1033667, had similar frequencies in all these populations.

The rs1033667 variant was investigated by bioinformatic analysis for evidence of a functional effect. The SNP is located in an intron of *CHEK2*, 90 bp downstream of a donor splice site. Analysis of this variant using a variety of predictive tools (see Materials and methods) did not provide any direct evidence for a functional effect. However, expression quantitative trait locus (eQTL) data show that it is associated with altered expression of several genes in this region in multiple tissue types, including increased expression of *CHEK2* in liver and *HSCB* in oesophageal mucosa and reduced expression of *TTC28* in transformed fibroblasts (26,27). The only SNP currently known to be in strong LD ($r^2 > 0.8$) with rs1033667 in African populations is rs2078555, which is also an intronic SNP in *CHEK2*. Bioinformatic analysis of this variant using the same tools also failed to reveal any direct evidence for a functional effect and showed similar eQTLs to rs1033667 for *CHEK2*, *HSCB* and *TTC28* both in terms of direction of effect and tissue type.

The *CHEK2* variant rs1033667 was tested for GxE interactions with tobacco smoking in a case-control analysis. Tobacco smoking was an independent risk factor for OSCC, OR = 3.04 (2.36–3.90) for ever-smokers ($P < 0.001$) compared to non-smokers. No significant interaction with rs1033667 was observed for tobacco smoking ($P = 0.60$). A case-only analysis was carried out for interaction of rs1033667 with alcohol use (see Materials and methods), but no significant interaction was observed ($P = 0.13$). Although the *ALDH2* SNP rs4767364 was not associated with OSCC in our study, this variant was tested for interaction with alcohol because of strong evidence for interaction between variants in this gene and alcohol use in other populations (8,13), but no interaction was detected ($P = 0.30$).

Discussion

This study investigated seven main-effect susceptibility loci that were previously reported to be associated with OSCC in non-African populations (9–14) for association with this disease in the SAB population. The *meta*-analysis included 1471 cases and 1791 controls from the Western Cape and Gauteng provinces, which is more than double the sample size of any previous African study of oesophageal cancer genetics. Only one SNP (*CHEK2* rs1033667) was significantly associated with

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

OSCC in the SAB population in the *meta*-analysis of the two study samples [*P* = 0.002, OR = 1.18 (1.06–1.30)]. This finding suggests that shared aetiology of the disease may exist between the Chinese populations and the SAB populations involving the *CHEK2* gene. None of the other loci tested showed evidence of transference to the SAB population, even though this study was well powered to detect these associations except for two low-frequency SNPs.

CHK2 (encoded by *CHEK2*) is a serine-threonine kinase with a key role in the DNA damage response pathway. Upon activation by genotoxic stress, it can phosphorylate downstream proteins to lead the activation of DNA repair, cell-cycle arrest, senescence or apoptosis (28). Rare germline mutations in *CHEK2* have been implicated in susceptibility to multiple cancers, including strong evidence of association with breast and prostate cancer (29). The *CHEK2* gene lies within a 100 kb region of chromosome 22 which also contains other genes of interest in relation to cancer, *XBP1*, *HSCB* and *TTC28*. *XPB1* encodes an X-box binding protein 1 which is involved in the unfolded protein response in the endoplasmic reticulum. The stress caused by the accumulation of unfolded proteins activates this response which leads to the restoration of normal protein folding, or in severe cases, apoptosis (30). *XBP1* has been shown to have an important role in the development and progression of triple-negative breast cancer through its control of the HIF1alpha pathway (31). *HSCB* encodes a member of the heat shock cognate B (HscB) family of proteins involved in the synthesis of iron-sulphur clusters and redox reactions of mitochondrial electron transport. Expression of *HSCB* is elevated in breast cancer tissue, and downregulation of this gene has recently been shown to reduce cell proliferation (32). *TTC28* encodes a tetratricopeptide repeat containing protein which may be involved in spindle formation in mitosis. Whole genome sequencing of OSCC tumour tissue has revealed frequent structural alterations in *TTC28*, the significance of which is not yet clear (33). The *CHEK2* SNP rs1033667 is associated with altered expression of *CHEK2* itself and also of *HSCB* and *TTC28*. A detailed fine-map of this region in large sample sizes will be required to address the genes and variants that are driving this association in more detail.

The variants from the other six loci of interest were not significantly associated with OSCC in either of our sample sets or in the *meta*-analysis. These included a low-frequency functional SNP in *TP53*, Pro47Ser, which is unique to African populations (20). However, the SNP has an allele frequency of only 2–3%, so it would require a very large sample size to resolve a possible association with this or any other cancers in African populations. Interestingly, this variant has recently been reported to be associated with pre-menopausal breast cancer in African-American women (34). In our previous study of the SAB population of the Western Cape, we reported evidence of a possible association of a coding SNP in *PLCE1*, rs17417407, with OSCC (16). However, this was not replicated in the JCS sample and the association was not significant in the *meta*-analysis. It is possible that the larger sample size is a more robust reflection of the effect size for this variant, or there may be inter-African population differences that will only be resolved by the analysis of very large sample sizes in these populations (35). As discussed previously, there are several possible reasons for a lack of association in our studies (15,16). First, there may be insufficient power to detect the effect sizes observed in the original GWAS studies. However, we had 78–96% power to detect an effect (OR) of 1.26 for the GWAS-derived SNPs at alpha 0.0046 in the *meta*-analysis, so detection of an association signal should have been possible at most of these loci.

Another possibility is that the actual causal variants in the Chinese population arose after the migration of humans out of Africa and are therefore not present in African populations. A recent exome-wide association study of OSCC reported six new susceptibility loci in the Chinese population (36). However, inspection of the 1000 Genome project phase 3 data for West and East African populations showed that two of the three low-frequency variants (rs117353193 and rs17848945) are monomorphic and one (rs138478634) has a MAF of 0.001; thus, these three loci are unlikely to make a significant contribution to OSCC risk in African populations. The three common associated variants from chromosome 6p21.3 in the exome study will be tested as part of an ongoing GWAS in the SAB population.

An important challenge in the investigation of transference of genetic factors between African and non-African populations is that LD is generally lower across African genomes as compared to Asian and European genomes, with shorter haplotype blocks (37). Thus, the genotyped SNPs may be in high LD with the causal variant in the Chinese population but in low LD in the SAB populations, and hence the genotyped SNP would not 'tag' the causal SNP in an African population well and a disease association might not be observed. Finally, it is possible that differences in population structure between cases and controls could contribute to the weakening of genuine differences in allele frequencies of SNPs between cases and controls (37). This seems unlikely in the UCT sample since cases and controls were predominantly from the same ethno-linguistic group, but the JCS study sample included a broader ethnic mix of SAB participants. In general, one would expect differences in the population structure of cases and controls to increase false positive associations rather than reducing genuine disease associations, but this could only be resolved by high throughput genotyping of very large numbers of SNPs with appropriate statistical correction for any differences observed.

Tobacco smoking is a well-established risk factor for OSCC (reviewed in 4). We therefore investigated the only locus in our study associated with OSCC for interaction with smoking using a case-control analysis. We confirmed tobacco smoking as an independent risk factor for OSCC in this sample but did not find evidence for an interaction between the *CHEK2* variant rs1033667 and smoking. A case-only analysis, rather than a case-control analysis, was done to investigate a gene-environment interaction for the rs1033667 variant with alcohol consumption owing to the very high alcohol consumption rate in the JCS control samples. No differences in allele frequencies of rs1033667 were observed when comparing drinkers with non-drinkers, which is consistent with a previous report that associations with OSCC at the chromosome 22q12 locus did not differ significantly between subgroups with different alcohol drinking status (13). Although the *ALDH2* SNP rs4767364 was not associated with OSSC in our study, this variant was tested for interaction with alcohol use because of strong evidence for interaction between variants in this gene and alcohol in other populations (8,13), but no interaction was detected.

Encouragingly, our current study has identified a SNP in the *CHEK2* locus that is associated with OSCC in the SAB population. This provides support for the hypothesis that at least some risk variants for OSCC will prove to be shared across multiple populations and for the prior evidence of association at *CHEK2* in GWAS and DNA repair pathway analysis of OSCC in Chinese populations (13,14,38). Although most risk variants from other populations do not replicate in the SAB population, an African GWAS is needed to provide deep coverage of the known OSCC

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

risk loci and to identify potential novel risk variants in African populations.

Supplementary material

Supplementary data are available at *Carcinogenesis* online.

Funding

This work was supported by the Cancer Association of South Africa (to C.G.M., C.B.dV. and C.M.L.), the University of the Witwatersrand Research Council (to C.G.M.), the National Research Foundation of South Africa (to C.G.M.: incentive grant to a rated researcher; W.C.C.: German Academic Exchange Service-National Research Foundation Joint In-country Scholarship Programme and the National Research Foundation and Department of Science and Technology Thuthuka fund); Boehringer Ingelheim Fonds travel grant (to W.C.C.); The Grants, Innovation and Product Development Unit (GIPD) of the South African Medical Research Council (SAMRC) and the South African National Department of Health (to E.S.), King's Health Partner's Research and Development Challenge Fund, King's College London (to C.G.M. and H.B.), the South African Medical Research Council Newton fund #045 to the ERICA-SA project (to C.G.M., C.B.dV., C.M.L. and E.S.) and the Wellcome Trust (grant 094491/Z/10/Z to N.J.P.). In addition, this paper represents independent research part-funded by the National Institute for Health Research Biomedical Research Centre at South London and Maudsley National Health Service Foundation Trust and King's College London, and the National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' National Health Service Foundation Trust and King's College London, United Kingdom. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care. M.I.P., M.M. and C.G.M. were supported by a grant from the South African Medical Research Council, with funds received from the South African National Department of Health, the Medical Research Council (United Kingdom) and GlaxoSmithKline via the Newton Fund grant #046. M.M. was recipient of an International Centre for Genetic Engineering and Biotechnology Postdoctoral Fellowship.

Acknowledgements

We thank sisters Gloria Mokwatle, Patricia Rapoho and Pheladi Kale for assistance with sample collection, and Phillip Tombleson (National Institute for Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust and King's College London), Margaret Urban and Tonicah Maphanga (National Cancer Registry, National Health Laboratory Service, South Africa) for assistance in the preparation and management of the sample database; Cassandra Soo and Natalie Smyth (Sydney Brenner Institute for Molecular Bioscience Biobank) and Ariella Rowe (International Centre for Genetic Engineering and Biotechnology) for DNA extractions; Dr Aron Abera and Felicity Azubuike (Inqaba Biotechnical Industries (Pty) Ltd. Southern Africa) for MassARRAY genotyping.

References

1. Ferlay, J. *et al.* (2013) GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer, Lyon.
2. Arnold, M. *et al.* (2015) Global incidence of oesophageal cancer by histological subtype in 2012. Gut, 64, 381–387.
3. National Cancer Registry. *Cancer in South Africa 2014 Full Report*. http://www.nioh.ac.za/assets/files/2014%20NCR%20tables.pdf (1 May 2018, date last accessed).
4. Abnet, C.C. *et al.* (2018) Epidemiology of esophageal squamous cell carcinoma. Gastroenterology, 154, 360–373.
5. Murphy, G. *et al.* (2017) International cancer seminars: a focus on esophageal squamous cell carcinoma. Ann. Oncol., 28, 2086–2093.
6. Matejcic, M. *et al.* (2017) Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. Carcinogenesis, 38, 859–872.
7. Matejcic, M. *et al.* (2015) Gene-environment interactions in esophageal cancer. Crit. Rev. Clin. Lab. Sci., 52, 211–231.
8. Cui, R. *et al.* (2009) Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. Gastroenterology, 137, 1768–1775.
9. McKay, J.D. *et al.* (2011) A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. PLoS Genet., 7, e1001333.
10. Abnet, C.C. *et al.* (2010) A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat. Genet., 42, 764–767.
11. Wang, L.D. *et al.* (2010) Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. Nat. Genet., 42, 759–763.
12. Wu, C. *et al.* (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat. Genet., 43, 679–684.
13. Wu, C. *et al.* (2012) Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat. Genet., 44, 1090–1097.
14. Wu, C. *et al.* (2014) Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nat. Genet., 46, 1001–1006.
15. Bye, H. *et al.* (2011) Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. Carcinogenesis, 32, 1855–1861.
16. Bye, H. *et al.* (2012) Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. Carcinogenesis, 33, 2155–2161.
17. Urban, M. *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 Med., 9, e1001182.
18. Matejcic, M. *et al.* (2011) Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. PLoS One, 6, e29366.
19. Chen, W.C. *et al.* (2018) The integrity and yield of genomic DNA isolated from whole blood following long-term storage at -30°C. Biopreserv. Biobank., 16, 106–113.
20. Jennis, M. *et al.* (2016) An African-specific polymorphism in the TP53 gene impairs p53 tumor suppressor function in a mouse model. Genes Dev., 30, 918–930.
21. The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature, 526, 68–74.
22. Machiela, M.J. *et al.* (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics, 31, 3555–3557.
23. Gabriel, S. *et al.* (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr. Protoc. Hum. Genet., Chapter 2, Unit 2.12.
24. Goh, L.L. *et al.* (2017) Analysis of genetic variation in CYP450 genes for clinical implementation. PLoS One, 12, e0169233.
25. Syrmis, M.W. *et al.* (2011) Comparison of a multiplexed MassARRAY system with real-time allele-specific PCR technology for genotyping of methicillin-resistant Staphylococcus aureus. Clin. Microbiol. Infect., 17, 1804–1810.
26. Westra, H.J. *et al.* (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet., 45, 1238–1243.
27. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. Science, 348, 648–660.
28. Zannini, L. *et al.* (2014) CHK2 kinase in the DNA damage response and beyond. J. Mol. Cell Biol., 6, 442–457.

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

29. Southey, M.C. *et al.*; Australian Ovarian Cancer Study Group (2016) PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. J. Med. Genet., 53, 800–811.
30. Hetz, C. *et al.* (2013) Targeting the unfolded protein response in disease. Nat. Rev. Drug Discov., 12, 703–719.
31. Chen, X. *et al.* (2014) XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. Nature, 508, 103–107.
32. Lee, S. *et al.* (2018) ChIP-seq analysis reveals alteration of H3K4 trimethylation occupancy in cancer-related genes by cold atmospheric plasma. Free Radic. Biol. Med., 126, 133–141.
33. Chang, J. *et al.* (2017). Genomic analysis of oesophageal squamous cell carcinoma identifies alcohol drinking related mutation signature and genomic alterations. Nat. Comm., 8, 15290.
34. Murphy, M.E. *et al.* (2017) A functionally significant SNP in TP53 and breast cancer risk in African-American women. NPJ Breast Cancer, 3, 5.
35. Gurdasani, D. *et al.* (2015) The African genome variation project shapes medical genetics in Africa. Nature, 517, 327–332.
36. Chang, J. *et al.* (2018) Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. Nat. Genet., 50, 338–343.
37. Teo, Y.Y. *et al.* (2010) Methodological challenges of genome-wide association analysis in Africa. Nat. Rev. Genet., 11, 149–160.
38. Wen-Qing, L. *et al.* (2013). Genetic variants in DNA repair pathway genes and risk of esophageal squamous cell carcinoma and gastric adenocarcinoma in a Chinese population. Carcinogenesis, 34, 1536–1542.

Downloaded from https://academic.oup.com/carcin/article-abstract/40/4/513/5316186 by guest on 16 July 2019

D - Plagiarism declaration

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

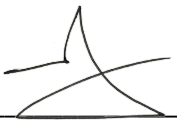
PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Wenlong Chen (Student number: 0700425A) am a student registered for the degree of PhD (Human Genetics) in the academic year 2016.

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:  Date: 31 March 2022

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

31 March 2022

RE: WCHEN 0700425A: Turn-It-In report clarification for PhD Submission

Dear Colleagues,

This letter is in support of Mr WC Chen's PhD submission. The Turn-It-In Originality report for his thesis returned a similarity index of 45%. This is due to the inclusion of his original published manuscripts as part of the thesis submission:

1. **Chen WC**, 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. https://doi.org/10.1089/bio.2017.0050
2. **Chen WC**, 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. https://doi.org/10.1093/carcin/bgz026
3. Simba H, Tromp G, Sewram V, Mathew C, **Chen WC**, Kuivaniemi H (2022) Esophageal Cancer Genomics in Africa: Recommendations for Future Research. Frontiers in Genetics, Human and Medical Genomics, https://doi.org/10.3389/fgene.2022.864575

Please refer to the Originality report attached. The top nine matches, accounting for 35% of the 45% of the similarity index, are all from his 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

Chen_OSCC_PhD_Thesis_Final_20march2022.docx

ORIGINALITY REPORT

45%	42%	38%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	doi.org Internet Source	14%
2	www.repository.cam.ac.uk Internet Source	12%
3	cansa.org.za Internet Source	2%
4	core.ac.uk Internet Source	2%
5	academic.oup.com Internet Source	1%
6	www.frontiersin.org Internet Source	1%
7	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%
8	kclpure.kcl.ac.uk Internet Source	

		1 %
9	Wenlong C Chen, Hannah Bye, Marco Matejcic, Ariella Amar et al. "Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population", Carcinogenesis, 2019 Publication	1 %
10	scholar.sun.ac.za:443 Internet Source	1 %
11	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 %
12	fuma.ctglab.nl Internet Source	<1 %
13	www.medrxiv.org Internet Source	<1 %
14	Gibson, Greg. "A Primer of Human Genetics", Oxford University Press Publication	<1 %
15	www.nature.com Internet Source	<1 %
16	www.nicd.ac.za	

	Internet Source	<1%
17	open.uct.ac.za Internet Source	<1%
18	hdl.handle.net Internet Source	<1%
19	Gaoming Li, Qiuyue Song, Yuxing Jiang, Angsong Cai et al. "Cumulative Evidence for Associations between Genetic Variants and Risk of Esophageal Cancer", Cancer Epidemiology Biomarkers & Prevention, 2020 Publication	<1%
20	Submitted to University of Leicester Student Paper	<1%
21	Bye, H., N. J. Prescott, C. M. Lewis, M. Matejcic, L. Moodley, B. Robertson, C. v. Rensburg, M. I. Parker, and C. G. Mathew. "Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population", Carcinogenesis, 2012. Publication	<1%
22	Submitted to University of Sheffield Student Paper	<1%
23	CHS-NHLBI.org Internet Source	<1%

24	wiredspace.wits.ac.za Internet Source	<1%
25	Submitted to University of Hong Kong Student Paper	<1%
26	www.ncbi.nlm.nih.gov Internet Source	<1%
27	Submitted to Kingston University Student Paper	<1%
28	dceg2.cancer.gov Internet Source	<1%
29	Yaniv Swiel, Jean-Tristan Brandenburg, Mahtaab Hayat, Wenlong Carl Chen, Mitchell A. Cox, Scott Hazelhurst. "FPGA Acceleration of GWAS Permutation Testing", Cold Spring Harbor Laboratory, 2022 Publication	<1%
30	ijins.umsida.ac.id Internet Source	<1%
31	Palwende Boua, Jean-Tristan Brandenburg, Ananyo Choudhury, Hermann Sorgho et al. "Genetic associations with carotid intima-media thickness link to atherosclerosis biology with sex-differences in sub-Saharan Africans", Research Square, 2021 Publication	<1%
	www.illumina.com	

32	Internet Source	<1%
33	www.uniprot.org Internet Source	<1%
34	Ugspace.ug.edu.gh Internet Source	<1%
35	d-nb.info Internet Source	<1%
36	pyrometallurgy.co.za Internet Source	<1%
37	acrabstracts.org Internet Source	<1%
38	etheses.whiterose.ac.uk Internet Source	<1%
39	link.springer.com Internet Source	<1%
40	onlinelibrary.wiley.com Internet Source	<1%
41	browser.epigenomesportal.ca Internet Source	<1%
42	Submitted to The University of Manchester Student Paper	<1%
43	www.genetics.org Internet Source	<1%

44	www.wits.ac.za Internet Source	<1 %
45	Ming, Wu. "The Etiology of Esophageal Cancer in High- and Low- Risk Areas of Jiangsu Province, China.", Wageningen University and Research, 2021 Publication	<1 %
46	lup.lub.lu.se Internet Source	<1 %
47	www.pubfacts.com Internet Source	<1 %
48	journals.plos.org Internet Source	<1 %
49	www.genecards.org Internet Source	<1 %
50	Adebowale A. Adeyemo, Guanjie Chen, Ayo P. Doumatey, Timothy L. Hostelley et al. " A Novel Type 2 Diabetes Locus in sub-Saharan Africans, , is Implicated in Beta Cell Proliferation ", Cold Spring Harbor Laboratory, 2019 Publication	<1 %
51	Segun Fatumo, Tinashe Chikowore, Robert Kalyesubula, Rebecca N Nsubuga et al. "Discovery and fine-mapping of kidney function loci in first genome-wide association	<1 %

study in Africans", Cold Spring Harbor Laboratory, 2020
Publication

52	discovery.ucl.ac.uk Internet Source	<1 %
53	mafiadoc.com Internet Source	<1 %
54	www.mysciencework.com Internet Source	<1 %
55	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	<1 %
56	journals.sagepub.com Internet Source	<1 %
57	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
58	www.fusedworld.com Internet Source	<1 %
59	Lucien Ferndale, Oluwatosin A Ayeni, Wenlong Carl Chen, Colleen Aldous. "Factors Influencing Survival Time in Patients Treated for Oesophageal Cancer with Palliative Intent	<1 %

	in South Africa", Research Square Platform LLC, 2021 Publication	
60	Thomas W Winkler, Felix Grassmann, Caroline Brandl, Christina Kiel et al. "Genome-wide association meta-analysis for early age-related macular degeneration highlights novel loci and insights for advanced disease", Cold Spring Harbor Laboratory, 2019 Publication	<1 %
61	fjfsdata01prod.blob.core.windows.net Internet Source	<1 %
62	researchspace.auckland.ac.nz Internet Source	<1 %
63	www.dakks.de Internet Source	<1 %
64	www.wjgnet.com Internet Source	<1 %
65	archive.org Internet Source	<1 %
66	sph.med.unsw.edu.au Internet Source	<1 %
67	www.diva-portal.org Internet Source	<1 %
68	www.saibw.co.za Internet Source	<1 %

69	Ditte Demontis, Veera Manikandan Rajagopal, Thomas D. Als, Jakob Grove et al. " Genome-wide association study implicates in cannabis use disorder ", Cold Spring Harbor Laboratory, 2018 Publication	<1 %
70	Submitted to King's College Student Paper	<1 %
71	Li, Wen-Qing, Nan Hu, Zhaoming Wang, Kai Yu, Hua Su, Lemin Wang, Chaoyu Wang, Stephen J. Chanock, Laurie Burdett, Ti Ding, You-Lin Qiao, Jin-Hu Fan, Yuan Wang, Yi Xu, Carol Giffen, Xiaoqin Xiong, Gwen Murphy, Margaret A. Tucker, Sanford M. Dawsey, Neal D. Freedman, Christian C. Abnet, Alisa M. Goldstein, and Philip R. Taylor. "Genetic Variants in Epidermal Growth Factor Receptor Pathway Genes and Risk of Esophageal Squamous Cell Carcinoma and Gastric Cancer in a Chinese Population", PLoS ONE, 2013. Publication	<1 %
72	Submitted to University of St Andrews Student Paper	<1 %
73	eprints.soton.ac.uk Internet Source	<1 %
74	peerj.com Internet Source	<1 %

75	Submitted to uu Student Paper	<1%
76	docs.lib.purdue.edu Internet Source	<1%
77	genepi.qimr.edu.au Internet Source	<1%
78	www.biomedcentral.com Internet Source	<1%
79	Submitted to University of Birmingham Student Paper	<1%
80	Xiaoyang Wang, Jianbo Tian, Qianyu Zhao, Nan Yang et al. "Functional characterization of a low-frequency V1937I variant in FASN associated with susceptibility to esophageal squamous cell carcinoma", Archives of Toxicology, 2020 Publication	<1%
81	bmjopen.bmj.com Internet Source	<1%
82	summit.sfu.ca Internet Source	<1%
83	Choi, Jihoon. "Bioinformatics Approaches for Investigating Missing Heritability in Complex Traits", Queen's University (Canada), 2021 Publication	<1%

84	Submitted to Glasgow Caledonian University Student Paper	<1%
85	Hannah Simba, Helena Kuivaniemi, Vittoria Lutje, Gerard Tromp, Vikash Sewram. "Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations", Frontiers in Genetics, 2019 Publication	<1%
86	Junjie Hu, lele zhang, Haoran Xia, Yilv Yan et al. "Single-Cell RNA Sequencing Reveals Tumor Microenvironment Remodeling after Neoadjuvant Immunotherapy in Non-small Cell Lung Cancer", Research Square Platform LLC, 2021 Publication	<1%
87	Submitted to University of Hull Student Paper	<1%
88	www.aasciences.africa Internet Source	<1%
89	Frederick E. Grine, Elsa Gonzalvo, Lloyd Rossouw, Sharon Holt, Wendy Black, José Braga. "Variation in Middle Stone Age mandibular molar enamel-dentine junction topography at Klasies River Main Site assessed by diffeomorphic surface matching", Journal of Human Evolution, 2021	<1%

Publication

90	Michael C Turchin, Gregory Darnell, Lorin Crawford, Sohini Ramachandran. "Pathway Analysis within Multiple Human Ancestries Reveals Novel Signals for Epistasis in Complex Traits", Cold Spring Harbor Laboratory, 2020 Publication	<1 %
91	Zehnder, Tobias. "Computational Approaches for the Prediction of Gene Regulatory Elements and the Analysis of Their Evolutionary Conservation", Freie Universitaet Berlin (Germany), 2021 Publication	<1 %
92	creativecommons.org Internet Source	<1 %
93	discovery.dundee.ac.uk Internet Source	<1 %
94	liebertpub.com Internet Source	<1 %
95	www.annualreviews.org Internet Source	<1 %
96	www.zora.uzh.ch Internet Source	<1 %
97	Bhattacharya, Arjun. "Statistical Methods for Inferring Genetic Regulation Across Heterogeneous Samples and Multimodal	<1 %

Data.", The University of North Carolina at Chapel Hill, 2020
Publication

98	ar.iiarjournals.org Internet Source	<1%
99	repository.lib.fit.edu Internet Source	<1%
100	thescholarship.ecu.edu Internet Source	<1%
101	theses.gla.ac.uk Internet Source	<1%
102	theses.lib.polyu.edu.hk Internet Source	<1%
103	www.nhmrc.gov.au Internet Source	<1%
104	www.tandfonline.com Internet Source	<1%
105	Barel, Gal. "Network Propagation With Node Core for Genotype-Phenotype Associations and Module Identification", Freie Universitaet Berlin (Germany), 2021 Publication	<1%
106	Luyapan, Jennifer. "Pathway-Based Approaches for Understanding Cancer Susceptibility Through Genetic Analyses", Dartmouth College	<1%

	Publication	
107	Nesdoly, Sean. "Characterizing Host-Virus Interactions Through Host Factor Curation and Network Analysis", McGill University (Canada), 2021 Publication	<1 %
108	Raymond, Biaty. "Use of Whole-Genome Sequence Data for Genomic Prediction Across Populations and Species.", Wageningen University and Research, 2021 Publication	<1 %
109	Submitted to University of Leeds Student Paper	<1 %
110	acsess.onlinelibrary.wiley.com Internet Source	<1 %
111	admin.humlab.umu.se Internet Source	<1 %
112	brill.com Internet Source	<1 %
113	cyberleninka.org Internet Source	<1 %
114	etd.lib.metu.edu.tr Internet Source	<1 %
115	jech.bmj.com Internet Source	<1 %

116	researchrepository.murdoch.edu.au Internet Source	<1 %
117	ujcontent.uj.ac.za Internet Source	<1 %
118	www.doria.fi Internet Source	<1 %
119	www.gov.uk Internet Source	<1 %
120	www.gprc.ab.ca Internet Source	<1 %
121	www.jstage.jst.go.jp Internet Source	<1 %
122	www.mdpi.com Internet Source	<1 %
123	www.mixed-models.org Internet Source	<1 %
124	Dzakovich, Michael Paul. "Exploring Metabolic and Genetic Diversity in Tomato Secondary Metabolites", The Ohio State University, 2021 Publication	<1 %
125	Naciye Magusali, Andrew C Graham, Thomas M Piers, Pantila Panichnantakul et al. "Genetic variability associated with OAS1 expression in myeloid cells increases the risk of Alzheimer's	<1 %

disease and severe COVID-19 outcomes", Cold Spring Harbor Laboratory, 2021
Publication

126	Prisca K. Thami, Wonderful T. Choga, Delesa D. Mulisa, Collet Dandara et al. "Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana", Cold Spring Harbor Laboratory, 2020 Publication	<1 %
127	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 %
128	Yizhen Zhong, Minoli Perera, Eric R. Gamazon. "On Using Local Ancestry to Characterize the Genetic Architecture of Human Phenotypes: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations as a Model", Cold Spring Harbor Laboratory, 2018 Publication	<1 %
129	bmccancer.biomedcentral.com Internet Source	<1 %
130	bmcgenet.biomedcentral.com Internet Source	<1 %

131	clok.uclan.ac.uk Internet Source	<1 %
132	nyaspubs.onlinelibrary.wiley.com Internet Source	<1 %
133	pdfs.semanticscholar.org Internet Source	<1 %
134	pure.eur.nl Internet Source	<1 %
135	pure.uva.nl Internet Source	<1 %
136	repository.up.ac.za Internet Source	<1 %
137	repub.eur.nl Internet Source	<1 %
138	research.vu.nl Internet Source	<1 %
139	scholar.sun.ac.za Internet Source	<1 %
140	www.mrc.ac.za Internet Source	<1 %
141	www.omicsonline.org Internet Source	<1 %
142	www2.bio.ku.dk Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		