

**GENOME-WIDE ASSOCIATION STUDY AND POLYGENIC RISK SCORE FOR  
OESOPHAGEAL CANCER IN SOUTH AFRICAN BLACK POPULATION**

Smangaliso Oageng

2451699



UNIVERSITY OF THE  
WITWATERSRAND,  
JOHANNESBURG

Supervisor:

Dr Mahtaab Hayat

Research Report

Submitted in fulfilment of the requirements for the degree

BSc (Honours)


In

Applied Bioinformatics

In the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

## **DECLARATION**

I, Smangaliso Oageng, hereby certify that this report is the result of my own work. It is submitted for the fulfilment of the requirements for the degree of Honours in Applied Bioinformatics at the University of the Witwatersrand, Johannesburg. I confirm that it has not been presented for any degree or examination at this or any other institution.

A handwritten signature in black ink, consisting of a large, stylized 'S' followed by a wavy line.

---

Smangaliso Oageng, signed on the 10<sup>th</sup> day of October 2024

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to thank God for giving me the strength, guidance, and perseverance to complete this work. Without His grace, this would not have been possible.

I extend my heartfelt gratitude to my family and friends for their unwavering support and encouragement throughout this journey. Your belief in me kept me going through challenging times.

A special thank you to my supervisor, Dr Mahtaab Hayat, for her invaluable guidance, mentorship, and patience. Your insights and expertise were instrumental in shaping this project. I am also deeply grateful to Dr Chen for generously providing the data that was essential to this research.

Thank you all for your contributions to the success of this work.

## Table of Contents

DECLARATION .....	i
ACKNOWLEDGEMENTS .....	ii
List of Figures .....	v
List of Tables .....	vi
Chapter 1: ABSTRACT .....	1
Chapter 2: INTRODUCTION .....	2
2.1.    Oesophageal Cancer Overview .....	2
2.2.    Risk factors for OC .....	3
2.3.    Treatment available for OC .....	3
2.4.    Genome-Wide Association Studies .....	4
2.5.    Polygenic Risk Scores .....	6
2.6.    Study Aim and Objectives .....	8
Chapter 3: METHODS .....	8
3.1.    Study design .....	8
3.2.    Study participants .....	9
3.3.    Bio-sampling .....	9
3.4.    Genotyping .....	10
3.5.    Quality control .....	11
3.6.    Population sub-structure control .....	13
3.7.    Testing for sex as a covariate .....	14
3.8.    Separation of base and target population files (70/30) .....	14
3.9.    GWAS LMM modelling .....	15
3.10.    Polygenic Risk Scores (PRS) .....	16
3.10.1.    PRSIce-2 .....	16
3.10.2.    PRScsx .....	17
Chapter 4: RESULTS .....	17
4.1.    Study samples and genotyping .....	17
4.2.    Population sub-structure control .....	18
4.3.    GWAS results .....	19
4.4.    PRS .....	21

Chapter 5: DISCUSSION.....	23
5.1. Study design .....	23
5.2. Population sub-structure control .....	24
5.3. Suggestive significant hits.....	26
5.3.1. Genetic Insights on the Associated SNPs.....	26
5.3.2. Genes that are associated with OSCC from other studies .....	30
5.3.3. Factors that could have influenced the results of this study.....	31
5.3.4. Population Stratification and Genomic Inflation .....	33
5.4. Polygenic Risk Scores .....	33
5.4.1. PRS models comparison.....	33
5.4.2. PRS models comparison in literature .....	34
5.4.3. Factors that could have influenced the PRS models .....	34
Chapter 6: CONCLUSION .....	36
References .....	37

## List of Figures

<b>Figure 1: The estimated age-standardized incidence rates (ASR) for oesophageal cancer in 2022, applicable to all age groups. ....</b>	<b>2</b>
<b>Figure 2: The ancestry of GWAS participants throughout history, in comparison to the global population. ....</b>	<b>4</b>
<b>Figure 3: Outline workflow of GWAS on African OSCC.....</b>	<b>9</b>
<b>Figure 4: Eigenvalue curve for controls and patients of OSCC.....</b>	<b>18</b>
<b>Figure 5: PCA plot of cases and cases with CEU and YRI as reference population. ....</b>	<b>19</b>
<b>Figure 6: The Manhattan plot of 2,250 AWI-Gen, JCS controls, and 1,183 OSCC cases ....</b>	<b>20</b>
<b>Figure 7: QQ plot of 2250 AWI-Gen, JCS controls, and 1183 OSCC cases. ....</b>	<b>21</b>
<b>Figure 8: Distributions and Predictive Performance PRS on the target population using PRSice-2.....</b>	<b>22</b>
<b>Figure 9: Distributions and Predictive Performance PRS on the target population using PRSsxx. ....</b>	<b>23</b>

## List of Tables

**Table 1: Leading SNPs associated to OSCC in South African populations (p-value <  $5 \times 10^{-6}$ ).20**

## Chapter 1: ABSTRACT

Oesophageal squamous cell carcinoma (OSCC) poses a significant health burden in South African populations, with a limited understanding of its genetic underpinnings. Polygenic risk scores (PRSs) are tools with predictive power that, when well curated, can be implemented into health care settings to screen patients and make informed decisions. PRSs use genome-wide association studies (GWAS) data to be created. While GWAS have identified genetic contributions to OSCC in non-African populations, few large-scale studies have been conducted in resident African populations. As a result, the genetic contribution to OSCC in resident African populations is largely unknown, and PRSs have not been generated or tested. This study aimed to carry out a GWAS on South African samples with OSCC, and create and test a PRS. Here, I report a GWAS involving 1 690 individuals with OSCC and 3 217 population-matched controls from three South African locations (Pietermaritzburg, Cape Town, and Soweto) which investigated the genetic aetiology of OSCC in African ancestry. Samples were genotyped using the H3Africa Custom African SNP Array and analysed using the H3ABioNet/H3AGWAS pipeline, incorporating PLINK for quality control and GEMMA for association testing via LMM.

The association analysis identified 11 SNPs with suggestive associations ( $p < 5 \times 10^{-6}$ ) near biologically relevant genes, including *GPC6*, *EPHA4*, *FBXL14*, and *NHLH2*, though no SNPs reached genome-wide significance ( $p < 5 \times 10^{-8}$ ). PRS analysis was conducted using PRSice-2 and PRSsx on the target population dataset consisting of 506 OSCC cases and 965 controls. PRSice-2 explained 1.4% of the variance in OSCC risk, slightly outperforming PRSsx, which explained 1%, although both models exhibited limited predictive power (AUC: 0.6 for PRSice-2 vs. 0.56 for PRSsx).

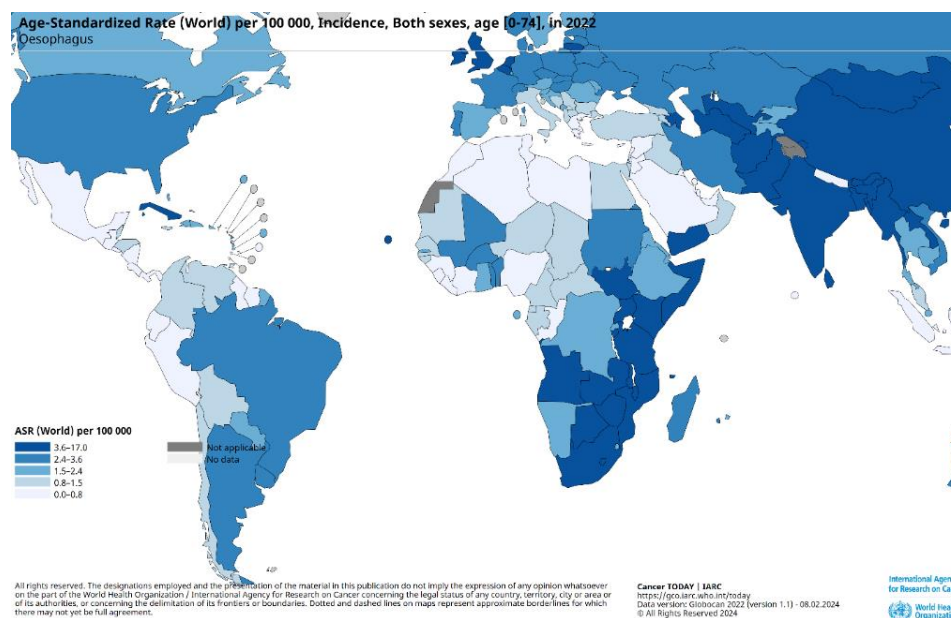
PRSice-2 showed better performance in this study, but PRSsx typically has more accuracy in African populations due to its ability to handle local ancestry. Lack of genotype imputation limited SNP coverage, reducing the models' power. To improve predictive accuracy, future work should incorporate genotype imputation, larger sample sizes, and environmental factors (e.g., smoking). Fine-mapping, meta-analyses, and gene-environment interactions are necessary in future research to better understand OSCC risk and address gaps in genetic research for underrepresented populations.



## Chapter 2: INTRODUCTION

### 2.1. Oesophageal Cancer Overview

Globally, oesophageal cancer (OC) is the eighth most common cancer, and it ranks as the sixth most prevalent cause of cancer-related deaths. In 2020, oesophageal cancer was estimated to have caused 544 100 deaths (Morgan *et al.*, 2022). Typically, the 5-year survival rate for oesophageal cancer is approximately 20%, but it can notably increase for individuals diagnosed in the early stages and decrease for those diagnosed in later stages when the cancer has metastasised (Yang *et al.*, 2020). It was estimated that in 2022 there were 92 322 cases of OC in South Africa (Figure 1) (Sung *et al.*, 2021). Oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (ADC) are known as the two main histological types. Typically, squamous cell carcinoma is seen in the middle and upper segments of the oesophagus. The squamous cells lining the oesophagus are the source of it. Usually, adenocarcinomas develop in the lower oesophageal region, close to the stomach. It arises from glandular cells found in the oesophageal lining, particularly near the junction of the oesophagus and the stomach (Yang *et al.*, 2015).



**Figure 1: The estimated age-standardized incidence rates (ASR) for oesophageal cancer in 2022, applicable to all age groups.**

The data shows particularly high rates in undeveloped nations, while South Africa is experiencing an upward trend in incidence compared

In Western countries, ADC is predominantly found, and it is correlated with Barret's oesophagus and gastric reflux. Barrett's oesophagus is a condition marked by the abnormal change in the oesophageal

lining, where the usual squamous cells are replaced by columnar cells resembling those in the intestines. Whereas, in developing countries, OSCC is predominantly found and of its global incidence, it comprises 85% (Uhlenhopp *et al.*, 2020). Sub-Saharan Africa, central China, and northern Iran are regions of high risk for OSCC. In northern Iran and central China, it was estimated that the incidence rate of OSCC is greater than 100 cases per 100 000 person-year, and high rates that span from 21 to 47 and 14 to 32 cases per 100 000 person-year were also observed in East and South Africa respectively (Then *et al.*, 2020).

## **2.2. Risk factors for OC**

The cause of OSCC is very complex. In Africa, the risk factors that are known include alcohol use, poor nutrition, tobacco smoking, exposure to smoke from cooking fires, limited financial resources and social standing, and dental fluorosis leading to tooth loss (Yang *et al.*, 2020). Case control studies were done in urban Soweto, South Africa and they proved that OSCC is associated with alcohol and smoking, especially the intake of homemade beer and pipe tobacco smoking (Tarazi *et al.*, 2021). The Johannesburg Cancer Study (JCS) revealed that the likelihood of developing OSCC rises when smoking is coupled with consistent alcohol consumption. This likelihood spikes from an odds ratio of 0.9 among non-smokers to 4.4 among smokers (Hull *et al.*, 2020).

A meta-analysis found that the prevalence of human papillomavirus (HPV)-16/18 infection in individuals with OSCC was 18%. However, a detailed examination of tumour tissues from individuals with HPV antibodies in high-incidence areas did not consistently detect the presence of HPV DNA, or HPV mRNA upregulation. As with other malignancies like cervical or oropharyngeal carcinoma, there is no evidence to establish a strong, continuous association between HPV and OSCC, but HPV may have a modest or additive influence, especially in Asian populations (Halec *et al.*, 2016).

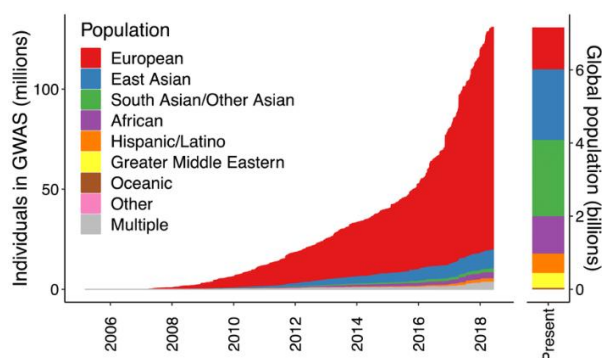
## **2.3. Treatment available for OC**

OSCC often progresses without symptoms, leading to late diagnosis and a poor prognosis. Despite advances in treatment, including immunotherapy, targeted therapies, and chemoradiotherapy, OSCC survival rates remain low, particularly in Africa. Immunotherapy, using immune checkpoint inhibitors like anti-PD-1 and anti-PD-L1 antibodies, has transformed OSCC treatment, especially in advanced cases (He *et al.*, 2021). Targeted therapies focus on pathways like HER2 and EGFR, tailored to tumour characteristics. Chemoradiotherapy, often combining carboplatin and paclitaxel with radiation,

remains the standard for locally advanced OSCC, as shown in improved survival rates in the CROSS study (Thumbs *et al.*, 2012). Palliative care, including stent insertion for dysphagia, remains critical for advanced cases (Ferndale *et al.*, 2023). However, the prognosis remains grim, with median survival times as low as 3.5 months in Mozambique and five-year survival rates ranging from 10% to 20% in East and Southern Africa, underscoring the significant public health challenge posed by OSCC in the region and the need for a screening tool in attempt to identify OSCC before it advances (Come *et al.*, 2018).

## 2.4. Genome-Wide Association Studies

Genome-wide association studies (GWAS) are genetic research methods that are used to find associations between genetic variations with specific diseases or traits across the entire genome. By analysing the genetic makeup of large populations, scientists can pinpoint genetic variations more prevalent in individuals with certain traits or diseases. This approach has been done on many diseases using populations with European ancestry (Figure 2) and it offers valuable insights into the genetic factors underlying various conditions, aiding our understanding of their mechanisms and potentially guiding the development of new treatments or interventions for these populations (Uffelmann *et al.*, 2021). However, populations that consist of African ancestry have been understudied (Figure 2). Risk alleles that are discovered in European populations are not well transferred to non-European populations, particularly African populations. The lack of transferability of risk alleles between these populations is due to different environmental factors, genetic diversity, and allele frequencies (Abdellaoui *et al.*, 2023).



**Figure 2: The ancestry of GWAS participants throughout history, in comparison to the global population.** Around 79% of all GWAS participants are of European ancestry, even though Europeans represent only 16% of the global population (Martin *et al.*, 2019).

Environmental factors vary between African and European populations such as lifestyle, exposure to pathogens, and diet which can interact with SNPs that influence the disease risk and trait which affect the transferability of GWAS results (Abdellaoui *et al.*, 2023). European populations show less genetic diversity compared to African populations. The higher diversity in African populations can lead to distinct genetic structures for diseases and traits, complicating the application of findings across different populations. There might be different allele frequencies for genetic variants that are associated with diseases or traits across populations. Genetic variants that are uncommon in European populations might be more prevalent in African populations, and vice versa. This discrepancy can contribute to variations in the genetic underpinnings of traits among different populations (Kamiza *et al.* 2022). This lack of transferability prevents these African populations from accessing the advantages of precision medicine methods and blocks the possibility of Africa's diverse genetic makeup contributing to innovative breakthroughs in understanding human disease genetics (Abdellaoui *et al.*, 2023). While earlier research has proposed genetic factors contributing to the risk of OSCC, most studies were carried out in Asian populations, with scant data available from African populations (Nariman *et al.*, 2020).

In OSCC, when GWAS was conducted using a Chinese population consisting of 2 044 controls and 2 039 OSCC cases (affected individuals) with independent verification of 8 620 controls and 8 092 controls. Functional variations in *ALDH2* and *ADH1B* were found, and when combined with alcohol and smoking, they increase the risk of developing oesophageal cancer significantly (Cui *et al.*, 2009). When GWAS was performed using Chinese Han descent population by genotyping 1 733 controls and 1 077 OSCC cases, *PLCE1* at 10q23 was discovered which was a previously unknown susceptibility locus (Abnet *et al.*, 2010). When was performed using a Chinese population consisting of 3 302 controls, 2 115 OSCC cases, and 2 240 gastric cancer, it was identified that OSCC and gastric adenocarcinoma share a susceptibility loci in *PLCE1* at 10q23 (Wang *et al.*, 2010). When GWAS was performed using a Chinese population consisting of 2 044 controls and 2 031 OSCC cases utilizing 66 6141 SNPs, multiple loci associations were identified including *HEATR3*, *MTMR3*, *PDE4D*, *STING1*, and *TP53* (Wu *et al.*, 2011). When exome-wide interrogation was performed using European population consisting of 3 880 controls and 3 714 OSCC cases, six novel susceptibility loci in *CCHCR1*, *CYP26B1*, *FASN*, *LTA*, *TCN2*, and *TNXB* were identified (Chang *et al.*, 2018).

In a South African Black (SAB) population, GWAS was performed and OSCC risk loci were identified, and this offered support for a significant role of genetic elements in the susceptibility to OSCC (Chen *et al.*, 2023). On chromosomes 2 and 9 for African OSCC, power signals were detected for *MYO1B* on chromosome 2 and lead SNP on chromosome 9. A novel association between OSCC and the *MYO1B* gene on chromosome 2 was identified, with the lead SNP occurring at a frequency unique to the SAB population. The lead SNP associated with OSCC was pinpointed upstream of the *FAM120A* gene on chromosome 9 and this gene is associated with survival signalling pathways and is frequently overexpressed across different cancer types. More analysis identified another independent SNP close to the *STAT4* gene, this gene is involved in the immune system for cytokine signalling and this suggested that it might be involved in the development of OSCC. A transethnic meta-analysis that was conducted on the Chinese and African populations by the same authors showed that there are unique and common genetic variants among the two populations for OSCC such as *PLCE1* which is unique to the Chinese population and *CHEK2* which is shared between the 2 populations (Chen *et al.*, 2023). All these suggest that there are differences between other populations and the SAB population for genetic contribution to OSCC and that there are novel SNPs associated with OSCC specific to SAB populations. A more comprehensive and significantly larger association study is required to explore the genetic susceptibility in SAB populations for OSCC (Chen *et al.*, 2023).

## **2.5. Polygenic Risk Scores**

Polygenic Risk Score (PRS) is a genetic assessment tool to calculate an individual's likelihood of developing specific diseases or traits by considering SNPs across numerous genes. A GWAS identifies numerous SNPs that individually exert a minor influence on a phenotype, which limits their overall predictive capability. While GWAS can pinpoint mutations linked to specific phenotypes, they often leave a gap in explaining heritability; the SNPs significantly associated with a trait do not fully account for its heritability. In 2010, Yang *et al.* demonstrated that by aggregating many SNPs with small effects, it is possible to estimate the risk of developing a particular phenotype. This insight led to the creation of polygenic risk score methodologies. PRSs are determined by calculating a weighted sum of risk alleles, with the weights usually based on effect sizes obtained from extensive GWAS (Lewis and Vassos, 2020). Different methods are used to calculate these scores including clumping, thresholding, and shrinkage. By clumping and thresholding variants, high linkage disequilibrium

variants are eliminated and only variants that satisfy a p-value threshold in the GWAS are retained (thresholding). Clumping refers to selecting variants that are independent of each other, i.e. selecting variants that are not in LD with each other. Finally, shrinkage involves reducing the effect size of SNPs deemed non-important by the programme being used to remove any bias these SNPs may introduce to the PRS. Subsequently, the PRS is computed by adding up all the risk alleles, weighted by effect size, per individual (Privé *et al.*, 2019). There are multiple, freely available PRS tools that all employ various combinations of the three main basic PRS methods described above.

PRSs are important in many fields such as genetic counselling, personalized medicine, and medical research. In these fields, they can aid in identifying individuals who have a high risk of developing specific diseases to implement early interventions or targeted preventive measures, which can help reduce the burden on health systems in low- to middle-income countries specifically (Lewis and Vassos, 2020). These scores rely on GWAS data as they use GWAS summary statistics such as variants, effect sizes, and/or p values. Since well-powered GWAS are mostly non-African based, these scores work best for non-African populations and work poorly for African populations as they are underrepresented (Kachuri *et al.*, 2024). This is why there is a need to perform PRS on African populations to identify populations with a high risk of developing OSCC to implement preventative measures to reduce challenges faced by South and East African public health.

GWAS and PRS analysis was performed in this study using South African OSCC populations. There are numerous significant advantages of conducting GWAS and PRS analysis for OSCC in the black community in South Africa. Black South Africans have a high prevalence of OSCC, and GWAS can identify genetic risk factors that may contribute to this condition (Mathew *et al.*, 2017). For this population, better early detection, screening, and focused preventive measures can result from this information. Due to genetic variations, PRS derived from European or other ancestry data sometimes perform poorly as predictors in African populations. To predict the risk of OSCC in this population, PRS scores calculated using GWAS data from black South Africans will yield more accurate results. The identification of genetic markers particular to a population that impacts the risk and prognosis of OSCC can inform the creation of more efficient and customized approaches for prevention, screening, and treatment of the disease in South African

## **2.6. Study Aim and Objectives**

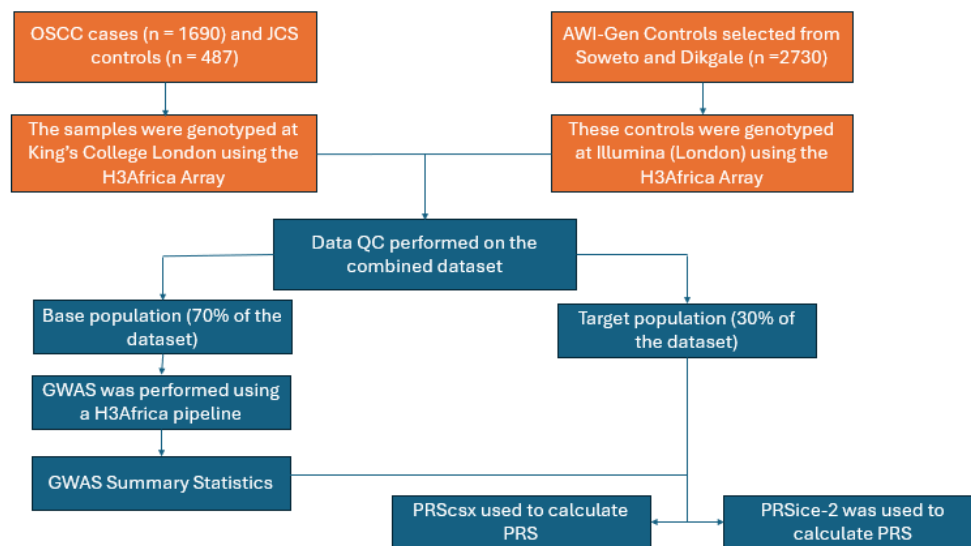
The aim is to enhance our understanding of the genetic contribution in OSCC and assess different PRS methods through GWAS and PRS analysis. The hypothesis is that the genetic variants that are associated with OSCC will be identified and that different PRS methods will have different predictive powers. The objectives were:

1. To carry out quality control using the H3Agwas pipeline.
2. To use the H3Agwas pipeline to carry out a GWAS and determine which SNPs are associated with oesophageal cancer.
3. To compare the predictive ability of two different programs that develop and calculate PRS.

## **Chapter 3: METHODS**

### **3.1. Study design**

This study employed a quantitative approach, utilising a case-control GWAS design to identify genetic variants associated with OSCC in black South Africans. Two PRS methodologies were evaluated to assess their efficacy in predicting OSCC risk. Due to the limited availability of genetic data from black South Africans, the OSCC dataset was split using a 70/30 ratio to carry out the GWAS, and create the PRS in the 70% dataset, and apply the PRS to the 30% dataset. The workflow of how this study was performed is outlined in Figure 3. The Human Research Ethics Committee (HREC) (Medical) of the University of the Witwatersrand granted ethical clearance; certificates M160807 and M2111154 were received by Chen *et al.* 2023.



**Figure 3: Outline workflow of GWAS on African OSCC.**

The orange blocks are for the work that was done by Chen *et al.* (2023) and the blue ones are for the work that I did. Blood samples were obtained from the samples and gDNA was extracted for genotyping using the H3Africa array. QC was done on the combined dataset and the cleaned data was separated into base and target population datasets in a 70/30 ratio. GWAS was performed on the base population using the linear mixed model to identify genetic variants associated with OSCC. The GWAS summary statistics were used to create the PRS and were applied to the target population dataset using PRScsx and PRSice-2.

### 3.2. Study participants

The samples used in this study were obtained from three study sites where it was confirmed that the individuals had OSCC previously, and these study sites are across South Africa. Participants in the study came from Grey's Hospital in Pietermaritzburg, KZN; the University of Cape Town (UCT), Western Cape; and Soweto, Johannesburg, Gauteng. In total, there were 1 690 OSCC cases (477 WC OSCC, 268 KZN OSCC, and 945 JCS-Soweto) and 3 217 population controls (1 702 AWI-Gen – Soweto, 1 028 AWI-Gen - Dikgale and 487 JCS-Soweto) (Chen *et al.*, 2023).

### 3.3. Bio-sampling

Genomic DNA (gDNA) was extracted from peripheral blood samples obtained from all participants in the study. In summary, either the salting-out approach or a kit-based DNA extraction using the Qiagen DNA FlexiGene kit was used to obtain gDNA (Chen *et al.*, 2023). This is a cost-effective, quick, and safe, to extract DNA that makes the salting-out process of deproteinisation easier. Protease K and Sodium Dodecyl Sulfate (SDS) were used to lyse and digest buffy coats from blood samples. Buffy coats are a layer of platelets and white blood cells that form after whole blood is centrifuged. Following digestion, proteins were precipitated by adding a saturated NaCl solution, and



centrifugation was performed. After transferring the DNA-containing supernatant, ethanol was used to precipitate the DNA. The extracted DNA was regularly checked to make sure its purity was high (Mwer *et al.*, 1988). The extracted gDNA was reconstituted in TE buffer (10 mM Tris-HCL, and 0.1 mM EDTA [pH 8.0]) and kept at  $-80^{\circ}\text{C}$  at the University of the Witwatersrand's HREC-approved Sydney Brenner Institute for Molecular Bioscience biobank until needed (Chen *et al.*, 2023).

### 3.4. Genotyping

All the gDNA samples were genotyped using the H3Africa Custom African  $\sim 2.3$  million SNP Array (Illumina). This type of microarray chip is used because of all the populations, those in Africa have the most genetic variety, and the H3Africa array targets more than 2.3 million SNPs important to African populations, and was created expressly to capture this variety (Mulder *et al.*, 2018). Genotyping for every person affected by OSCC and the JCS population control group was done at the Genomics Core Facility, Social Genetics & Development Psychiatry Centre, King's College London. The Illumina FastTrack Sequencing Service in the United States of America was used to genotype the AWI-Gen population control group (Chen *et al.*, 2023). In summary, genotyping was performed by first preparing the samples. During the sample preparation process, DNA was first fragmented, labelled, and then hybridized onto the SNP array. In the hybridization phase, these DNA fragments attach to specific probes on the array that are tailored to target distinct SNPs throughout the genome. After this step, the array is scanned to identify the presence of these SNPs as the array give signals if a fragment with a SNP is bound to its specific complementary probe. The signals resulting from the array are stored as intensities on raw data intensity files (Mulder *et al.*, 2018).

These raw data intensity files needed normalisation and preprocessing for genotype information which is useful because they only had fluorescence intensity as numerical values that represent the amount of DNA that was bound. Using the Illumina Array Analysis Platform (IAAP) Genotyping orchestrated command-line workflow and the Illumina GenCall algorithm, genotype clustering and calling were carried out for all affected OSCC individuals as well as control people (Chen *et al.*, 2023). Genotype clustering and calling are essential processes that follow the hybridization and scanning of Illumina SNP arrays. After obtaining raw data from each SNP site, it is necessary to analyse this data to determine the genotypes. This analysis involves grouping similar signal intensities into clusters that represent the three possible genotypes: homozygous for the major allele (AA), heterozygous (AB), and homozygous for the minor allele (BB). The clustering process categorizes the signals into these

specific groups, while genotype calling assigns each sample to one of these clusters based on its signal intensity (Sipedy et al., 2020).

The IAAP Genotyping Orchestrated Workflow is a comprehensive system that processes raw intensity data, a bead pool manifest, and cluster files as input files. It generates genotype calls and quality metrics for each sample and outputs the results in various formats, including MAP and PED files. The bead pool manifest file contains a detailed list of all the SNPs included in the H3Africa SNP array chip (Chen et al., 2023). It acts as a reference to ensure that the genotyping analysis targets the correct SNPs that the array is designed to detect. Cluster files provide crucial information about each genotype's fluorescent signal intensity thresholds (such as AA, AB, or BB). These files are vital for organizing the raw intensity data into accurate genotype calls, as they define the limits for clustering signal intensities (Morante and Ballestar, 2021). Together, the manifest and cluster files ensure the proper interpretation of the raw data during genotyping, allowing for accurate genotype calls and effective quality control (Chen et al., 2023). The output MAP and PED files are then obtained for quality control in PLINK.

### **3.5. Quality control**

The PLINK software version 1.9 was used to convert MAP and PED files to PLINK binary files (BED, BIM, and FAM formats) by Chen *et al.* (2023). PED files follow a specific structure that includes required columns for genotype data, phenotype, sex, and individual and family IDs. The first six columns contain essential metadata, while the remaining columns store genotype information for each SNP, formatted in a way that PLINK can interpret. The genotype data is represented as pairs of alleles for each SNP, allowing PLINK to properly encode and analyse the genetic information. Along with the PED file, PLINK needs a corresponding MAP file, which provides SNP details such as genetic distance, SNP ID, chromosome number, and base pair position. This MAP file is key for associating the genotype data in the PED file with specific SNPs (Sipedy et al., 2020). This was already done and we received the binary files of the data, namely .fam, .bed and .bim files.

The H3ABioNet/H3AGWAS Pipeline Version 3 was utilized for formatting data and conducting quality control (QC) on the datasets. The H3ABioNet/H3AGWAS pipeline was specifically created to carry out GWAS among African populations because these populations have unique genetic diversity and population structures (Brandenburg *et al.*, 2022). The H3AGWAS pipeline is optimized for QC in African populations by using customized parameters, including minor allele frequency

(MAF) thresholds, flexible Hardy-Weinberg Equilibrium (HWE) cutoffs, Principal Component Analysis (PCA) with African reference panels, relatedness filters, and imputation quality controls. These settings enable it to retain rare variants, manage population structure, correct for platform-specific effects, and minimize technical artifacts, making it highly suitable for the complexity of African genetic data (Brandenburg *et al.*, 2022). This pipeline includes the use of PLINK in the workflow as it is widely used for quality control, and this was used for quality control with specific parameters. Quality control involve accounting for SNP selection, individual missingness, SNP missingness, MAF, HWE, manifest-genotype sex match, relatedness, and heterozygosity.

Only autosomal SNPs were included because association analysis focuses only on autosomal SNPs (non-sex chromosomes). This guarantees that the analysis remains unbiased individual, and it is specifically recommended that individuals should have no more than 2% of their genotype data missing. High levels of missingness can signal poor sample quality, potentially distorting the results. Minimizing missingness makes the dataset robust, ensuring that analyses are conducted on by sex-linked traits and that the genetic variations examined are relevant to the entire population (Truong *et al.*, 2022). Individual missingness was accounted for by only selecting individuals with individual missingness  $\leq 0.02$ . Individual missingness indicates the percentage of missing genotype data for each complete and trustworthy data (Murphy *et al.*, 2021). SNP missingness was accounted for by only selecting SNPs with SNPs missingness  $\leq 0.01$ . According to this criterion, no more than 1% of the genotyping data for any given SNP should be missing in any one person. SNPs with a high percentage of missing data may be unreliable and offer valuable insights into the study. Maintaining low missingness contributes to the genotype data's integrity. MAF was accounted for by only selecting SNPs with  $MAF \geq 0.01$  (common SNPs). The frequency at which the less frequent allele appears at a certain genetic locus within a population is known as MAF. SNPs with an  $MAF \geq 0.01$  offer enhanced statistical power for detecting associations between genetic variants and traits or diseases. This is crucial for preserving the integrity of the analysis, ensuring that identified associations are more likely to represent true signals rather than artifacts stemming from low-frequency variants.

HWE was accounted for by only including SNPs with a  $p \geq 0.00001$ . HWE test evaluates whether the observed genotype frequencies for a SNP agree to the HWE laws. P-value thresholds of 0.00001 indicate the exclusion of SNPs with significant deviations from HWE. Variations from HWE may be a sign of population structure effects, selection, or genotyping errors. This guarantees the validity of

the discovered genetic connections and aids in the identification of potentially troublesome SNPs (Murphy *et al.*, 2021). The Manifest-Genotype Sex Match was done by ensuring that the selected individuals' sex in the genotype data matches the sex listed in the manifest file, ensuring accurate analysis and preventing errors in studies involving sex-linked traits or adjustments by Chen *et al.* (2023). Only individuals with a  $\pi\text{-hat} \leq 0.18$  were selected to account for relatedness.  $\pi\text{-hat}$  is a metric used to assess the relatedness between individuals, with a threshold of 0.18 suggesting that the individuals should be distantly related, such as being first cousins or more distantly related. Because of shared DNA, association results may be skewed by close relatives.  $\pi\text{-hat}$  is calculated based on the proportion of alleles that are shared identical-by-state (IBS). Two alleles are considered IBS if they have the same nucleotide sequence, irrespective of their source. IBS occurs when the same allele is present at a specific locus in both individuals, regardless of whether it was inherited from a common ancestor (Murphy *et al.*, 2021).

Only individuals with heterozygosity rates between this range  $\leq 0.343$  &  $\geq 0.1$  were selected to account for heterozygosity by excluding the ones with low (heterozygosity  $\leq 0.343$ ) or high (heterozygosity  $\geq 0.1$ ) heterozygosity. The heterozygosity rate is an indicator of genetic diversity within an individual's genome, reflecting the proportion of heterozygous loci—locations where an individual possesses two different alleles (one inherited from each parent). A homogenous population or inbreeding may be suggested by extremely low heterozygosity, whereas contamination or incorrectly identified samples may be suggested by extremely high heterozygosity. Maintaining heterozygosity within a particular range aids in keeping the data's quality intact (Murphy *et al.*, 2021). After QC, the files were still in PLINK binary format.

### **3.6. Population sub-structure control**

This study used principal component analysis (PCA) to account for population sub-structure, after the genotype dataset was cleaned. Population sub-structure refers to variations in allele frequencies among different groups of individuals within a population, often arising from geographic isolation. If not addressed, this sub-structure can result in false-positive associations, as genetic differences may be incorrectly identified as links to the phenotype of interest. PCA was done with YRI ( $n = 60$ , Yoruba, Nigeria, filtered 2.6 million SNPs) and CEU ( $n = 60$ , European, filtered 2.3 million SNPs) individuals obtained from Phase 2 HapMap as a reference population. PCA is a statistical method used for reducing the dimensionality of large datasets while preserving important patterns and trends. It

converts a set of correlated variables into a smaller set of uncorrelated variables called principal components (PCs), which represent the majority of the variance in the data.

Standardizing the data is usually the first step in PCA to make sure every variable contributes equally to the analysis. To comprehend how variables vary together, the covariance matrix is calculated. The pairwise covariances between every variable are summarized in this matrix. The covariance matrix's eigenvalues and eigenvectors are computed. The eigenvalues display the amount of variation that is captured by each component, whereas the eigenvectors represent the directions of maximum variance (the major components). Principal components are ordered based on their eigenvalues, and a subset is chosen that captures most of the variance in the dataset. For instance, PC1 and PC2 usually have the highest variance (Murphy *et al.*, 2021). PLINKv1.90 was used to calculate PCs. PLINK produces two file formats namely .eigenval which contains the PCS' eigenvalues and .eigenvec file which has eigenvectors derived from the PCA. The .eigenvec file contains the principal component scores (coordinates) for every individual in the dataset. Each row represents an individual, with the first two columns denoting the family ID (FID) and individual ID (IID), followed by the scores for each principal component (PC1, PC2, etc.). The eigenvalue curve was plotted using the ggplot2 in R to identify PCs that account for most of the variation using .eigenval file. PCs 1 to 6 were included as covariates in the association analysis step because this accounts for genetic differences that are caused by population sub-structure thereby decreasing the risk of getting false-positive associations. The .eigenval file was then used to plot a PC plot of PCs with the highest variations (PC1 and PC2) against each other using ggplot2 on R 4.4.1 to identify samples that might cause population stratification and remove them on the PLINK binary files (Ahmad *et al.*, 2023). 2 controls and 1 case were removed.

### **3.7. Testing for sex as a covariate**

A t-test was used on R to check if there was a statistical difference between the sex of cases and controls. This test was done because the genetic effect differences between females and males can affect the accuracy of the results that were obtained at the end if they were not accounted for.

### **3.8. Separation of base and target population files (70/30)**

The cleaned PLINK binary files had both cases and controls so they were first separated to have files for cases and controls separately. The FIDs and IIDs of the cases and controls were obtained from the FAM file to create text files of cases and controls. The text files were then shuffled to ensure no biases

as this approach reduces the risk of confounding factors. The 70/30 ratio for cases and controls was then calculated to obtain the 70% and 30% text files of cases and controls. The 70% and 30% text files of cases and controls were combined to make the base population and target population text files. PLINK was used to obtain the PLINK binary files of the base and target population by using the text files as the input. GWAS was then performed on the base population with 2 250 controls, and 1 183 cases ( $n = 3433$ ) and PRS analysis on the target population with 964 controls, and 507 cases ( $n = 1471$ ).

### **3.9. GWAS LMM modelling**

GWAS was performed using the H3Bionet/H3Agwas pipeline by utilizing GEMMA LMM to analyse the relationship between OSCC and the genetic variants statistically. Sex and PCs 1 to 6 were used as covariates to ensure that the GWAS results were not affected by confounding factors and the base population PLINK binary files were taken as input files by the software. GEMMA computes a Genetic Relationship Matrix (GRM) that quantifies the pairwise genetic relatedness among individuals. This GRM is crucial for capturing shared genetic variation attributable to ancestry and relatedness, thereby aiding in the control of additional population stratification. GEMMA employs a linear mixed model (LMM) to decompose phenotypic variance into genetic and residual components. The model differentiates between fixed effects (such as covariates), random genetic effects (derived from the GRM), and residual effects. This structure allows GEMMA to correct for individual relatedness while estimating the effects of genetic markers. In conducting association tests, GEMMA evaluates the impact of each genetic variant on the phenotype, taking into account both fixed and random effects from the LMM. The results are compiled into an output file (.assoc) that includes p-values and effect sizes (beta values) for each marker, highlighting which variants show significant associations with the phenotype (Zhou, 2016). Because the LMM generates beta values, they were converted to ORs using a formula described by Cook *et al.* (2016).

A Manhattan plot and Quantile-Quantile (QQ) plot were then plotted using a qqman package on R 4.4.1 using the .assoc file. The Manhattan plot was used to visualize the SNPs associated with OSCC. The QQ plot was used to see if there was any population stratification before further analysis to ensure that the data was accurate. The beta values were transformed into odds ratios using the formula outlined by Cook *et al.* (2016). Due to the extensive number of SNPs being tested for association, the likelihood of false positive associations is mitigated by employing correction methods such as the

Benjamini-Hochberg and Bonferroni approaches for multiple testing. In GWAS, a commonly accepted threshold for significance is  $p < 5 \times 10^{-8}$ , meaning that any SNP with a p-value below this threshold will be deemed significantly associated with OSCC.

### **3.10. Polygenic Risk Scores (PRS)**

Two PRS software were used to generate a PRS in the base dataset (70% dataset) to apply to the target dataset (30% dataset). For each programme, a GWAS had to be carried out on the base dataset first. The summary statistics from the base dataset were used to create the PRS which was applied to the target dataset to assess how well the risk score works. The PLINK binary files of the target dataset are required to apply the PRS.

#### **3.10.1. PRSice-2**

PRSice-2 took the .assoc file of the base population as an input file as it has GWAS summary statistics. The PLINK binary files of the target population were also taken as input files to calculate the PRS for each individual on this dataset. Sex and PCs 1 to 6 were used as covariates to ensure that the PRS results were not affected by confounding factors. PRSice-2 performs clumping to handle linkage disequilibrium (LD) between SNPs. Clumping ensures that only independent SNPs are selected, by removing SNPs that are in high LD with each other. This process involves selecting a lead SNP and excluding any other SNPs within a specified distance that has an LD correlation (usually measured by  $r^2$ ). PRSice-2 calculates  $r^2$  values using maximum likelihood estimates of haplotype frequencies. A haplotype is a set of alleles or genetic markers located on the same chromosome that are inherited together. PRSice-2 tests multiple p-value thresholds to select the best subset of SNPs for constructing the polygenic risk score (Liu *et al.*, 2023). This involves testing whether including SNPs with a higher GWAS significance threshold improves the predictive power of the PRS (Vilhjálmsdóttir *et al.*, 2015). Different p-value thresholds are applied to the GWAS summary statistics to determine which SNPs should be used for score calculation. For each individual in the target dataset, PRSice-2 calculates the PRS by summing the effect sizes of the selected SNPs, weighted by the number of risk alleles the individual carries (Liu *et al.*, 2023).

The PRS results are compiled into an output file (.best) which has the calculated PRS for everyone in the target population including their FIDs and IIDs. A density plot was plotted using ggplot2 on R 4.4.1. A pROC package on R 4.4.1 was used to plot a ROC curve and to estimate the area under the curve (AUC).

### **3.10.2. PRScsx**

PRScsx was also used to generate PRS on the target population dataset. PRScsx took the .assoc file of the base population as an input file as it has GWAS summary statistics. The African (AFR) reference data from the 1000 Genomes Project was used as an LD reference panel. By utilising LD reference panels, PRS-csx can effectively model the correlations among SNPs. This is essential for addressing the dependencies between SNPs and minimizing redundancy when estimating the effect sizes for the PRS.

PRScsx employs a Bayesian shrinkage method to estimate SNP effect sizes across different populations. This approach is vital for minimizing noise and enhancing the accuracy of PRS. By applying a continuous shrinkage prior to the SNP effect sizes, PRScsx encourages the effect sizes of non-significant SNPs to decrease toward zero while preserving the larger effect sizes of significant SNPs (Kurniansyah *et al.*, 2023). This mechanism aids in controlling false-positive associations (Vilhjálmsón *et al.*, 2015). Additionally, the Bayesian model utilizes an LD reference panel to account for correlations between SNPs, ensuring that those in high LD with others have their effect sizes adjusted, thus preventing redundancy or overestimation of their contributions to the PRS (Kurniansyah *et al.*, 2023).

Once the effect sizes were estimated by PRScsx, output text files with SNP effect size for each chromosome were generated. PLINK was then used to calculate PRS for each individual and a .profile was generated. This file included the calculated PRS for each individual and their FIDs and IIDs. A density plot was plotted using ggplot2 on R 4.4.1. A pROC package on R 4.4.1 was used to plot a (ROC) curve and to estimate the area under the curve (AUC). PRScsx was then used to calculate the accuracy metrics of the PRS model.

## **Chapter 4: RESULTS**

### **4.1. Study samples and genotyping**

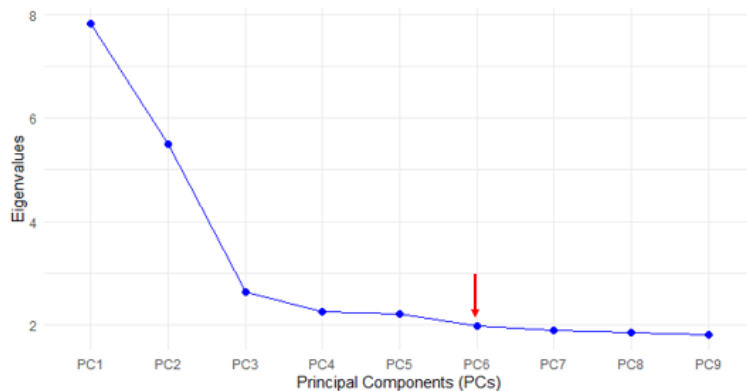
Histologically confirmed oesophageal cancer cases were selected from three study sites, along with ethnically matched population controls from the AWI-Gen and JCS studies, for genotyping and association analysis. All case participants were recruited from Grey's Hospital in Pietermaritzburg, KwaZulu-Natal; the University of Cape Town in the Western Cape; and Soweto in Johannesburg, Gauteng. There were 1 690 OSCC cases (477 WC OSCC, 268 KZN OSCC, and 945 JCS-Soweto)



and 3 217 population controls (1 702 AWI-Gen – Soweto, 1 028 AWI-Gen - Dikgale and 487 JCS-Soweto) that were sampled. All study participants identified themselves as being of African ancestry.

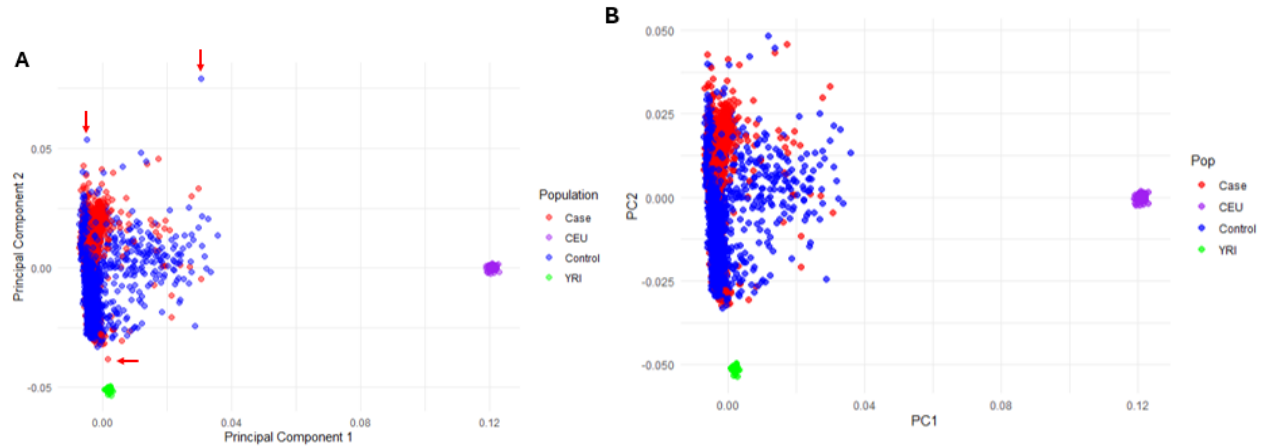
#### 4.2. Population sub-structure control

PCA was carried out using genotyped 1 656 285 SNPs. 10 PCs were calculated/constructed using PLINK. This principal component analysis (PCA) was conducted on the combined genotyped dataset (cases and controls). The selection of principal components (PCs) was guided by the eigenvalue curve (Figure 4), where visual inspection identified a point of inflection (indicated by an arrow in Figure 4) showing that PCs 1 through 6 captured the majority of the variance in the dataset. These PCs were utilized as covariates in the linear mixed model (LMM) for association analysis. YRI (n = 60) and CEU (n = 60) individuals were included as reference together with all the cases and controls in the PCA to identify individuals in the study participants that might cause population stratification and might be outliers (Figure 5). Two controls and one case were removed from the samples (indicated by an arrow in Figure 5A) because they were outliers indicating that they are too different from the dataset and might cause population stratification. The PC analysis plots shows that all the other samples were well matched as the cases and controls are overlapping with each other and they are not clustered with YRI and CEU populations (Figure 5B).



**Figure 4: Eigenvalue curve for controls and patients of OSCC.**

The arrow marks the point of inflection, indicating the number of principal components that account for most variation in the dataset.



**Figure 5: PCA plot of cases and cases with CEU and YRI as reference population.**

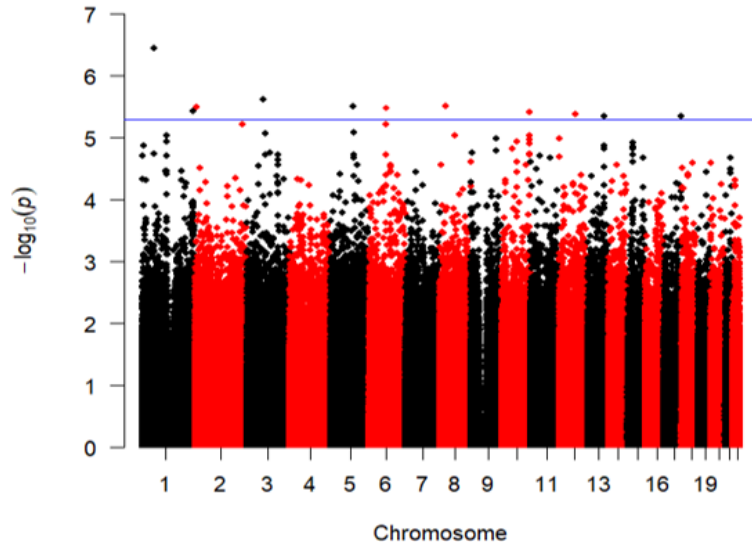
A) PCA plot illustrating the clustering of cases and controls within the study population. The plot shows two control individuals, and one case individual (indicated by red arrows) do not cluster together with the study population, indicating potential population stratification. The YRI and CEU individuals are distinctly separated from the study population. B) PCA plot following the removal of individuals contributing to population stratification. The removal of these outliers enhances the clarity of the study population's structure, facilitating more accurate association analysis in GWAS for oesophageal cancer. CEU = European, YRI = Yoruban.

After addressing the issues related to population substructure in the study, the subsequent step involved testing for sex as a covariate and separating the dataset files into base and target population data. After applying all quality control measures and correcting for population structure, 1 688 cases and 3 215 population controls were included in the analysis. There were 1 699 678 genotyped SNPs and after QC 1 656 285 SNPs were retained. The t-test showed that there is a significant difference in sex between cases and controls ( $p\text{-value} < 2.2 \times 10^{-16}$ ) indicating that sex needed to be accounted for as a covariate. After the file separation, SNPs of the base population were tested for association with African oesophageal cancer.

### 4.3. GWAS results

GWAS was performed on 1 183 OSCC cases and 2 250 population controls. This was done using an LMM technique because it corrects for individual relatedness while estimating the effects of genetic markers. In the LMM, the binary phenotype was analysed using PCs 1 through 6 and sex as covariates.

There were no SNPs that were found to be associated with OSCC as none of them had a p-value  $<5 \times 10^{-8}$  however there are 11 SNPs that show suggestive association with OSCC (p-value  $<5 \times 10^{-6}$ ) shown by both the Manhattan plot in Figure 6 and Table 1.



**Figure 6: The Manhattan plot of 2,250 AWI-Gen, JCS controls, and 1,183 OSCC cases**

The blue line represents a suggestive significance threshold with a p-value of  $5 \times 10^{-6}$ . 11 SNPs have a suggestive association with OSCC (p-value  $<5 \times 10^{-6}$ ) as they are above the blue the line. Displayed in relation to the  $-\log$  p-value.

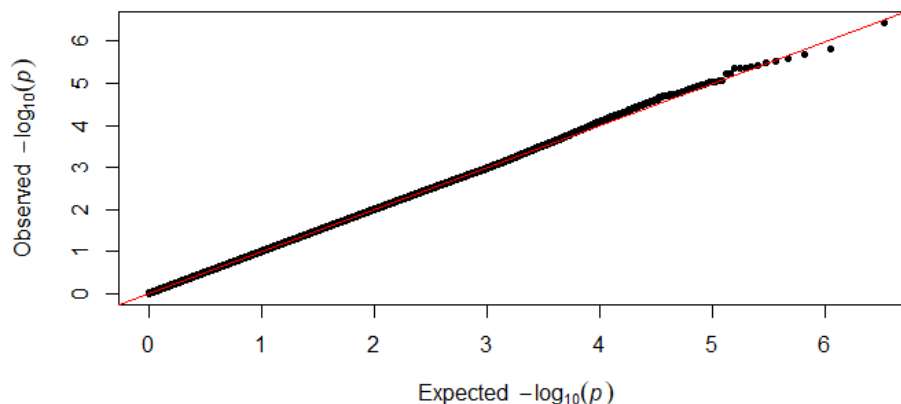
**Table 1: Leading SNPs associated to OSCC in South African populations (p-value  $< 5 \times 10^{-6}$ ).**

Ch r	rs_ID	Genes (within/ close to)	Position (GRCh37)	E A	A A	EAF	OR(95%CI )	P_value
1	rs11627995 2	<i>LOC12490415</i> 6	56826428	A	G	0.01 3	2.64(2.39- 2.88)	$3.66 \times 10^{-7}$
12	rs10879942	<i>LOC10536984</i> 4	76017475	T	G	0.21 0	1.29(1.22- 1.36)	$4.27 \times 10^{-6}$
13	rs9589807	<i>GPC6</i>	94309026	G	A	0.19 7	1.30(1.23- 1.38)	$4.46 \times 10^{-6}$
17	rs7220255	<i>LINC01993</i>	76289418	T	C	0.38 6	0.81(0.75- 0.87)	$4.47 \times 10^{-6}$
6	rs62413399	<i>SH3BGRL2</i>	80297080	T	C	0.04 6	1.64(1.50- 1.78)	$6.04 \times 10^{-6}$
2	rs11635681 5	<i>EPHA4</i>	22229853 6	A	G	0.12 6	1.36(1.27- 1.45)	$6.13 \times 10^{-6}$
5	rs10070308	<i>FBXL17</i>	10728162 1	T	C	0.26 2	1.25(1.19- 1.32)	$8.44 \times 10^{-6}$
3	rs1580082	<i>LINC00971</i>	84843143	C	A	0.24 3	0.79(0.72- 0.86)	$8.71 \times 10^{-6}$

10	rs1171728	<i>C10orf143</i>	13190756 8	A	G	0.03 7	1.70(1.54- 1.85)	9.15× 10 <sup>-6</sup>
8	rs13883104	<i>SLCO5A1</i>	70726388	T	C	0.03 9	1.67(1.52- 1.82)	9.28× 10 <sup>-6</sup>
1	rs2797179	<i>NHLH2</i>	11635765 4	T	G	0.39 4	1.23(1.17- 1.28)	9.35× 10 <sup>-6</sup>

Chr = chromosome number, rs\_ID = SNP identifier, EA = effect allele, AA = alternate allele, EAF = effect allele frequency, OR = odds ratio, 95%CI = 95% confidence interval, OR calculated with reference to EA

In Figure 7, the Quantile-Quantile (QQ) plot displays the expected ( $-\log_{10}(p)$ ) versus the observed ( $-\log_{10}(p)$ ). The expected and observed p-values align closely with the reference line, with deviations only appearing at the higher end of the plot. If deviations occurred earlier, it would indicate the presence of genomic inflation in the dataset. The genomic inflation factor ( $\lambda$ ) was 1.01, indicating no evidence of inflation in the test statistics. Genomic inflation in GWAS refers to a systematic bias that results in an overrepresentation of small p-values, creating the illusion of more significant associations than truly exist. This phenomenon often arises from population stratification, which involves differences in ancestry between the case and control groups.



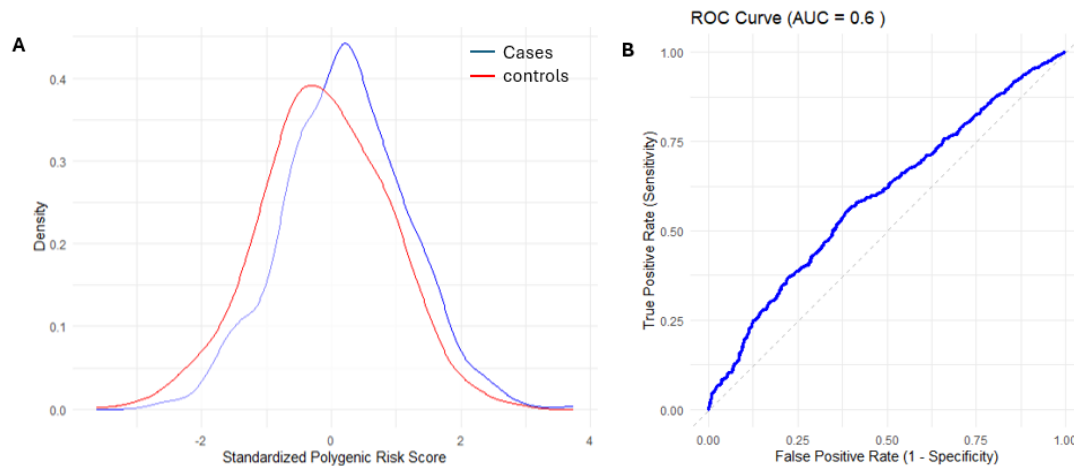
**Figure 7: QQ plot of 2250 AWI-Gen, JCS controls, and 1183 OSCC cases.**

The observed p-value line (black line) aligns closely with the null line (dashed red line) without deviating very early (genomic inflation factor  $\lambda = 1.01$ ). The points that deviate below the expected line at the absolute top show that there are no top hits (SNPs associated with OSCC).

#### 4.4. PRS

PRSice-2 was used to analyse the performance of PRS on the South African OSCC GWAS. PRS analysis was done on the target population with 506 OSCC cases and 965 controls using 296 121 SNPs after clumping. This PRS model, built using the target population, accounted for approximately 1.4% of the variance observed in the South African dataset. As shown in Figure 8A, there was minimal

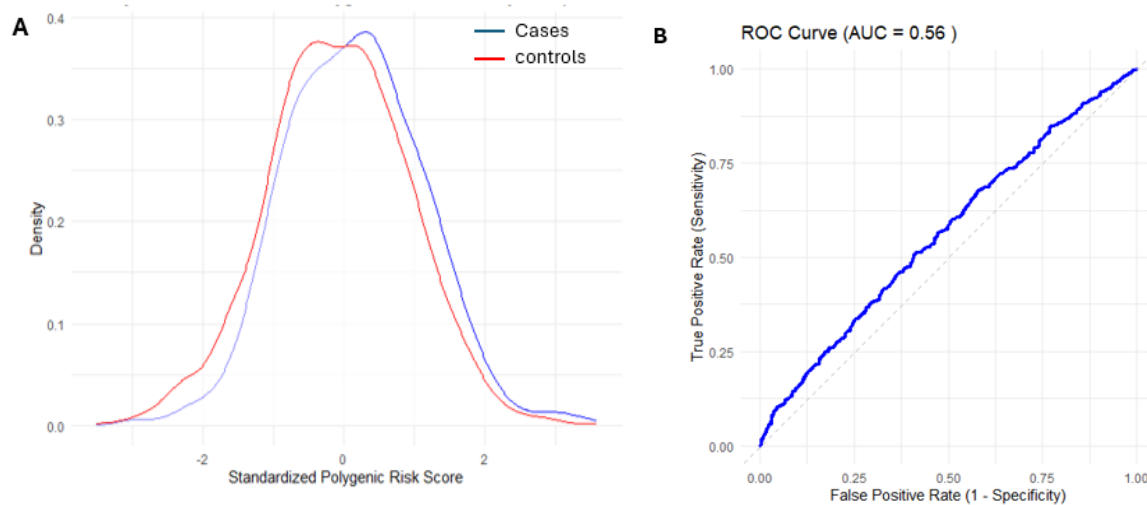
differentiation between cases and controls. For this model, the Area under the Curve is 0.6 (Figure 8B).



**Figure 8: Distributions and Predictive Performance PRS on the target population using PRSice-2.**

A) The density plot illustrates the distribution of standardized PRS for both cases (in blue) and controls (in red). The plot highlights the overlap and shift in PRS distribution between the two groups, with cases having slightly higher scores compared to controls however this is minimal. B) The Receiver Operating Characteristic (ROC) curve evaluates the PRS's ability to differentiate between cases and controls. The area under the curve (AUC) is measured at 0.6, indicating a moderate capacity of the PRS to predict group membership. The diagonal grey dashed line signifies a random classifier that lacks any discriminative ability.

PRScsx was also used to analyse the performance of PRS on the South African OSCC GWAS. PRS analysis was done on the target population with 506 OSCC cases and 965 controls using 557 946 SNPs estimated by PRScsx. This PRS model, built using the base population, accounted for approximately 1% of the variance observed in the South African dataset. As shown in Figure 9A, this also has minimal differentiation between cases and controls. For this model, the AUC=0.56 (Figure 9B).



**Figure 9: Distributions and Predictive Performance PRS on the target population using PRScsx.**

A) Density plot showing the distributions of the standardized Polygenic Risk Score (PRS) for cases (blue line) and controls (red line). The plot highlights the overlap and shift in PRS distribution between the two groups, with cases having slightly higher scores compared to controls. B) ROC curve demonstrating the performance of the PRS in distinguishing cases from controls. The AUC is 0.56, suggesting a low but above-random predictive ability of the PRS to differentiate between cases and controls.

## Chapter 5: DISCUSSION

Genetic risk loci associated with OSCC have been extensively characterised in Asian and European populations, where the majority of GWAS have been conducted on them. According to the literature, one GWAS has been conducted in resident sub-Saharan African populations for OSCC. Therefore, uncommon variations with great penetrance have a major role in the genetic aetiology of OSCC among resident African communities, especially in the black population of South Africa. The results of the OSCC GWAS and PRS analysis in a resident black population of South Africa are reported in this study. There were 2 250 population controls and 1 183 cases of OSCC in the GWAS base dataset. The target dataset for PRS analysis included 506 OSCC cases and 965 population controls.

### 5.1. Study design

Samples were collected from three locations in South Africa: Grey's Hospital in Pietermaritzburg, the University of Cape Town, and Soweto in Johannesburg. The study included 1 690 cases of OSCC and 3 217 controls. gDNA was extracted from peripheral blood using either the salting-out method or the Qiagen DNA FlexiGene kit. The samples were genotyped with the H3Africa Custom African SNP

Array. The H3Africa Custom African SNP Array is utilized for GWAS in African populations because it is specifically crafted to capture the extensive genetic diversity found in these populations. African populations exhibit the highest genetic variation among all human groups, and many SNPs common in these populations are not adequately represented in standard SNP arrays. The H3Africa array includes over 2.3 million SNPs relevant to African populations, making it more effective at identifying unique genetic associations. By providing a better representation of African genetic diversity, this array improves the power of GWAS to uncover genetic factors linked to complex traits and diseases in African populations, minimizing the likelihood of missing important associations that may not be captured by less specialized arrays (Mulder *et al.*, 2018).

Genotyping for both OSCC cases and controls was conducted at King's College London and the Illumina FastTrack Sequencing Service in the USA. The Illumina GenCall algorithm was employed for genotype clustering and calling by Chen *et al.* (2023). QC measures were implemented using PLINK software and the H3ABioNet/H3AGWAS Pipeline, focusing on filtering based on individual and SNP missingness, MAF, HWE, sex matching, relatedness, and heterozygosity. PC analysis was performed using YRI and CEU reference populations to account for population stratification in the study participants. A t-test assessed whether sex significantly affected the case-control distribution after which it was included as a covariate in the GWAS. The cases and controls were separated, shuffled, and divided into a 70/30 ratio to create base and target populations for further analysis.

GWAS was conducted on and the PRS was generated in the base population, while the PRS was tested on the target population. The H3ABioNet/H3AGWAS pipeline with GEMMA LMM was utilized for GWAS, correcting for relatedness with a genetic relationship matrix (GRM) (Mulder *et al.*, 2018; Zhou, 2016). Covariates included sex and principal components 1-6. PRS analysis was carried out using PRSice-2 and PRS-CSx software; PRSice-2 involved clumping and multiple p-value thresholds to construct PRS, while PRS-CSx employed a Bayesian shrinkage method using African LD information from the 1000 Genomes Project (Choi and O'Reilly, 2019; Ruan *et al.*, 2022). PRS-CSx is reported to perform well in diverse datasets (Ruan *et al.*, 2022). This comprehensive analysis aims to evaluate the association between genetic variants and OSCC while addressing unique population structure and stratification issues, and other confounding factors that could influence result accuracy.

## 5.2. Population sub-structure control

The PCA conducted on the genotyped SNPs (Figure 5) highlights the importance of correcting for population stratification and ensuring that the study participants form a genetically homogeneous group. In this analysis, PCA was performed on the combined dataset of 1 688 cases and 3 215 controls, using 1 656 285 genotyped SNPs that were retained after stringent quality control. The inclusion YRI and CEU populations in the PCA as reference groups allowed for the identification and exclusion of individuals that exhibited population substructure, which could confound the association analysis.

The selection of PCs for inclusion as covariates in the subsequent association analysis was based on the eigenvalue curve (Figure 4). Visual inspection of the curve revealed a point of inflection after the sixth component, suggesting that PCs 1 through 6 captured the majority of the variance in the dataset. Using these PCs as covariates in the LMM helped account for potential population stratification and minimized confounding effects due to population structure. This step is crucial, especially in studies involving African populations, which are characterized by high levels of genetic diversity (Sengupta *et al.*, 2021).

The PCA plots presented in Figure 5 demonstrated that, following the removal of two control individuals and one case individual that exhibited distinct genetic differences from the study population, the remaining samples showed minimal population stratification. The distinct clustering of YRI and CEU individuals, compared to the study samples, further highlighted that these outliers were not representative of the broader study population. The removal of these genetically distinct individuals reduced the risk of confounding, thereby increasing the accuracy and reliability of subsequent association analysis (Murphy *et al.*, 2021).

Another important step involved testing whether sex was a significant covariate. The t-test results indicated a highly significant difference in sex between cases and controls ( $p\text{-value} < 2.2 \times 10^{-16}$ ), suggesting that sex could act as a potential confounder in the association analysis. Including sex as a covariate in the model helped to adjust for these differences and ensured that the associations identified were not simply due to gender imbalances between cases and controls (Kwitshana *et al.*, 2020).

After applying quality control measures and correcting for population structure and sex, the dataset consisted of 1,688 cases and 3,215 population controls. Of the initial 1,699,678 genotyped SNPs,



1,656,285 SNPs were retained for further analysis. The subsequent file separation into base (70%) and target (30%) populations enabled the focused investigation of SNPs that might be associated with OSCC in African populations.

This rigorous approach to addressing population stratification and other confounders is essential for accurate and reliable GWAS findings. By accounting for population structure and sex differences, this study ensures that the identified genetic variants are more likely to be truly associated with oesophageal cancer, thereby enhancing the power of the GWAS analysis. Future studies can build upon these findings to further explore the genetic architecture of oesophageal cancer in African populations and uncover specific variants that contribute to disease risk and progression.

### **5.3. Suggestive significant hits**

GWAS conducted on 1 183 OSCC cases and 2 250 controls using LMM has revealed 11 SNPs with suggestive associations to OSCC in South African populations. While none of these SNPs reached the strict genome-wide significance threshold ( $p < 5 \times 10^{-8}$ ), their p-values below  $5 \times 10^{-6}$  suggest potential involvement in OSCC risk. It is likely that there were no genome-wide significant SNPs due to the power of the study. Including an imputed dataset will improve the power as will increasing sample numbers. These findings warrant further investigation to confirm their role in the disease and understand their biological mechanisms.

The Manhattan plot (Figure 6) clearly indicates the presence of suggestive associations, as shown by the 11 SNPs crossing the blue line, which represents a p-value threshold of  $5 \times 10^{-6}$ . These SNPs could serve to understand genetic variants that might contribute to OSCC susceptibility in African populations.

#### ***5.3.1. Genetic Insights on the Associated SNPs***

The table of leading SNPs (Table 1) provides detailed information about these variants, including their chromosomal positions, effect alleles, and odds ratios (OR). Notably, some of these SNPs exhibit odds ratios greater than 1, indicating an increased risk of OSCC for carriers of the effect allele, while others have OR values less than 1, suggesting a protective effect of the effect allele. Below, is a discussion of the eleven identified SNPs, their gene locations, functions, and potential deleterious effects (Cook et al., 2016).

On chromosome 1, rs116279952 shows the strongest association with OSCC ( $p = 3.66 \times 10^{-7}$ ), with an odds ratio of 2.64 (95% CI: 2.39-2.88), indicating that individuals carrying the effect allele (A) are at a significantly increased risk of developing OSCC. The effect allele, A, is rare with a frequency of 0.013. Given its low frequency and high OR, this SNP could be a candidate for further functional studies to understand its role in OSCC. This SNP is found near a pseudogene called the uncharacterized *LOC124904156* gene. A pseudogene is a non-functional DNA segment resembling a gene, often viewed as a "genomic fossil" due to mutations or deletions that prevent protein coding (Cheetham *et al.*, 2020). *LOC124904156*, a lncRNA gene, does not encode a protein but may influence gene expression or interact with molecular processes. LOC-prefixed genes, identified via sequencing, are often studied in population genetics and disease contexts (Statello *et al.*, 2021). lncRNAs play a vital role in cancer development by regulating gene expression, chromatin structure, and splicing processes. They can function as either enhancers or silencers of protein-coding genes, thereby influencing the expression of both oncogenes and tumor suppressor genes. Moreover, lncRNAs can recruit chromatin-modifying proteins, which changes DNA accessibility and gene expression patterns. Additionally, they interact with splicing factors to produce different protein isoforms that may interfere with cellular pathways linked to cancer (Mattick *et al.*, 2023). Investigating this lncRNA's regulatory role could be relevant to OSCC risk.

rs10879942 on Chromosome 12 shows an association with OSCC ( $p = 4.27 \times 10^{-6}$ ), with an OR of 1.29 (95% CI: 1.22-1.36), indicating that individuals carrying the effect allele (T) are at a significantly increased risk of developing OSCC. This variant's effect allele frequency (EAF) is 0.21, indicating that 21% of the population carries the effect allele (T). The modest increase in risk suggests that this SNP may contribute to genetic susceptibility to OSCC. This SNP is located within an uncharacterized *LOC105369844* gene. This gene is a lncRNA gene that may have regulatory roles in gene expression or other molecular functions (He *et al.*, 2023). It could be valuable to investigate whether this lncRNA has any regulatory relevance to the risk of OSCC.

rs9589807 on Chromosome 13 shows an association with OSCC ( $p = 4.46 \times 10^{-6}$ ), with an OR of 1.30 (95% CI: 1.23-1.38), indicating that individuals carrying the effect allele (G) are at a significantly increased risk of developing OSCC. With an EAF of 0.197, this SNP is present in a notable proportion of the population. This SNP is located within the *GPC6* gene. The *GPC6* gene encodes glypican-6, a cell surface proteoglycan involved in cell signalling pathways that regulate growth, differentiation,

adhesion, and migration, supporting tissue formation and repair. Abnormal *GPC6* expression has been associated with various cancers, including breast, colorectal, and ovarian cancers, as its dysregulation can drive tumour growth and metastasis through its impact on these signalling pathways (Liu *et al.*, 2020). Similar mechanisms could potentially be involved in the development of OSCC.

rs7220255 on Chromosome 17 shows an association with OSCC ( $p = 4.47 \times 10^{-6}$ ), with an OR of 0.81 (95% CI: 0.75-0.87), indicating that individuals carrying the effect allele (T) have a significantly protective benefit against developing OSCC. This SNP is located near the *LINC01993* gene. *LINC01993* is a long intergenic non-coding RNA (lincRNA) gene. Like other lincRNAs, *LINC01993* does not code for a protein but may have important regulatory roles in gene expression and cellular processes. Although specific functions for *LINC01993* are still being researched, lincRNAs are generally involved in various biological processes, including regulation of gene transcription, chromatin remodelling, and cellular signalling (Reid *et al.*, 2024). It could be valuable to investigate whether this lincRNA has any regulatory relevance to the protective effect against OSCC.

rs62413399 on Chromosome 6 shows an association with OSCC ( $p = 6.04 \times 10^{-6}$ ), with an OR of 1.64 (95% CI: 1.50-1.78), indicating that individuals carrying the effect allele (T) are at a significantly increased risk of developing OSCC. Its low EAF of 0.046 indicates that the effect allele (T) is relatively rare, yet its substantial OR suggests it may be a significant risk factor for OSCC. This SNP is located within the *SH3BGRL2* gene. *SH3BGRL2* is a gene that encodes a protein vital for essential cellular functions like signal transduction, cell proliferation, and apoptosis. SNPs within *SH3BGRL2* can influence its expression and functionality, potentially leading to pathway dysregulation. Increased expression or activity from these SNPs may create an environment conducive to cancer development by enhancing cell survival, while mutations that disrupt its normal function could contribute to tumour progression by changing how cells respond to growth signals (Li *et al.*, 2020).

rs116356815 on Chromosome 2 shows an association with OSCC ( $p = 6.13 \times 10^{-6}$ ), with an OR of 1.36 (95% CI: 1.27-1.45), indicating that individuals carrying the effect allele (A) are at a significantly increased risk of developing OSCC. The EAF for this variant is 0.126, indicating that it is present in a significant portion of the population. This SNP is located within *EPHA4*, a gene that encodes the Ephrin type-A receptor 4, a member of the Eph receptor family involved in key cellular functions such as cell signaling, development, and tissue patterning. *EPHA4* is crucial for bidirectional signalling pathways that regulate cell adhesion, migration, and proliferation, playing an important role in

neuronal development and angiogenesis. Additionally, it is associated with cancer biology, influencing tumor growth and metastasis. SNPs near or within *EPHA4* can affect its expression and function, potentially leading to dysregulation of its signaling pathways. Variations in *EPHA4* expression from these SNPs may alter cellular behaviors like migration and adhesion. This disruption can enhance tumor invasion and metastasis of cancer (Nikas *et al.*, 2022).

rs10070308 on Chromosome 5 shows an association with OSCC ( $p = 8.44 \times 10^{-6}$ ), with an OR of 1.25 (95% CI: 1.19-1.32), indicating that individuals carrying the effect allele (T) are at a significantly increased risk of developing OSCC. With an EAF of 0.262, this variant's effect allele (T) is relatively common, suggesting it may play a role in genetic susceptibility. This SNP is located within *FBXL17*, a gene that encodes a protein from the F-box protein family, known for its distinctive F-box motif. These F-box proteins are essential in the ubiquitin-proteasome system, a vital mechanism for protein degradation and regulation within cells. Single nucleotide polymorphisms (SNPs) found near or within the *FBXL17* gene may affect its expression or functionality. Changes in *FBXL17* expression caused by these SNPs could disrupt the normal process of ubiquitination, resulting in the accumulation of proteins that might promote tumour development or alter cell cycle regulation. If the function of *FBXL17* is impaired, it may lead to the initiation and progression of cancers, by allowing abnormal protein levels to persist and disrupt critical cellular pathways (Mason *et al.*, 2020).

rs1580082 on Chromosome 3 shows an association with OSCC ( $p = 8.71 \times 10^{-6}$ ), with an OR of 0.79 (95% CI: 0.72-0.86), indicating that individuals carrying the effect allele (C) have a protective effect against the development of OSCC. The EAF of 0.243 indicates that a substantial proportion of the population carries the effect allele (C). This SNP is located near *LINC00971* a lncRNA gene that is implicated in various biological processes, particularly gene regulation and cellular signaling. Although the specific functions of *LINC00971* are still being explored, lncRNAs like this one are recognized for their ability to influence gene expression through interactions with chromatin, transcription factors, and other regulatory proteins (Cook *et al.*, 2020). It could be valuable to investigate whether this lncRNA has any regulatory relevance to the protective effect against OSCC.

rs1171728 on Chromosome 10 shows an association with OSCC ( $p = 9.15 \times 10^{-6}$ ), with an OR of 1.70 (95% CI: 1.54-1.85), indicating that individuals carrying the effect allele (A) are at a significantly increased risk of developing OSCC. The EAF for this variant is 0.037, suggesting it is rare, but its substantial OR indicates a strong effect on OSCC risk. This SNP is situated in *C10orf143* a gene with

a less characterized role, but it is believed to participate in various cellular functions, possibly including cell growth and development, variants might disrupt its function and lead to cancer development (Liu *et al.*, 2019). rs138831048 on Chromosome 8 shows an association with OSCC ( $p = 9.27 \times 10^{-6}$ ), with an OR of 1.67 (95% CI: 1.52-1.82), indicating that individuals carrying the effect allele (T) are at a significantly increased risk of developing OSCC. rs138831048 is located within the *SLCO5A1* which encodes a solute carrier organic anion transporter essential for the uptake of various organic anions and plays a role in drug metabolism and clearance. Variants near this gene may affect its expression or functionality, potentially altering the metabolism of substances. This influence could impact cancer susceptibility or the effectiveness of treatment (Sutherland *et al.*, 2020).

Finally, the SNP rs2797179 on chromosome 1 is associated with OSCC, presenting a p-value of  $9.35 \times 10^{-6}$  and an OR of 1.23 (95% CI: 1.17-1.28) indicating that individuals carrying effect allele (T) have an increased risk of developing OSCC. With an EAF of 0.394, this variant's effect allele (T) is relatively common, suggesting that it may contribute to genetic susceptibility to OSCC. This SNP is situated close to the *NHLH2*. *NHLH2* is implicated in transcriptional regulation and may contribute to the development of the nervous system and other tissues. SNPs near *NHLH2* could influence its expression or the regulation of downstream target genes, potentially resulting in changes in cellular differentiation and a heightened risk of cancer (Carraro *et al.*, 2021).

All SNPs were investigated in the GWAS catalogue and there were no existing associations with OSCC reported with these SNPs. The identified SNPs offer important insights into the genetic factors linked to OSCC risk in South African populations. Many of these variants are found within or near genes that are essential for critical cellular processes, such as cell signalling, regulation of gene expression, and cellular growth, suggesting that variations in these genes may play a pivotal role in OSCC development and progression. Their potential harmful effects emphasize the necessity of understanding how these genetic variations may influence the development of OSCC and highlight the need for additional functional studies to clarify their roles in disease susceptibility.

### **5.3.2. Genes that are associated with OSCC from other studies**

The genetic landscape of OSCC differs considerably among various populations, illustrating the intricate interplay of genetic, environmental, and lifestyle factors that contribute to cancer susceptibility (Morgan *et al.*, 2022). Multiple GWAS conducted in diverse populations have

pinpointed numerous genetic loci associated with these cancers, revealing both shared and population-specific risk factors (Chen et al., 2023).

In comparison to other GWAS, no gene was found to overlap in our study. In the Chinese population, several notable genetic variants linked to OSCC have been discovered. For example, variations in *ALDH2* and *ADH1B* have been shown to interact with alcohol consumption and smoking, significantly heightening the risk of oesophageal cancer (Cui *et al.*, 2009). A GWAS conducted in a Chinese population identified multiple loci associated with OSCC, including *HEATR3*, *MTMR3*, *PDE4D*, *STING1*, and *TP53* (Wu et al., 2011). Additionally, the identification of *PLCE1* as a susceptibility locus for OSCC through GWAS involving Han Chinese individuals highlights its significance in cancer pathogenesis (Abnet *et al.*, 2010). Subsequent research further confirmed the association of *PLCE1* with both OSCC and gastric adenocarcinoma, underscoring its role as a shared susceptibility locus (Wang et al., 2010). Identification of novel susceptibility loci, such as *CCHCR1*, *CYP26B1*, *FASN*, *LTA*, *TCN2*, and *TNXB* in a European population (Chang *et al.*, 2018), enhances our understanding of the genetic factors influencing OSCC.

Conversely, the South African Black (SAB) population displays unique genetic markers associated with OSCC. GWAS conducted in this population identified susceptibility loci, including *MYO1B* on chromosome 2 and a lead SNP near the *FAM120A* gene on chromosome 9, which is linked to survival signalling pathways often overexpressed in various cancers. Furthermore, the discovery of an independent SNP near the *STAT4* gene suggests a possible role in immune signalling and OSCC development (Chen et al., 2023). This study was done using the same dataset that was used in this however none of the SNPs that were found to be in suggestive association with OSCC are located near or within *FAM120A* and *MYO1B*. This is attributed to the low power of this study having only used genotyped SNPs, whereas Chen *et al.* (2023) used an imputed dataset. This is a good example of how underpowered studies, as is the case in most GWASs in sub-Saharan Africa, fail to identify variants associated with phenotypes.

The transethnic meta-analysis revealed both unique and shared genetic variants among Chinese and African populations. While *PLCE1* was specific to the Chinese population, *CHEK2* was common to both groups (Chen et al., 2023). This highlights the necessity for large, population-specific studies to identify novel SNPs that may not be captured in non-African analyses.

### 5.3.3. *Factors that could have influenced the results of this study*

There were no SNPs that were found to be associated with OSCC in this study however the GWAS that was performed using the same data found that 2 genes (*FAM120A* and *MYO1B*) are associated with OSCC. The main reason for this is that the genotyped SNPs went through genotype imputation before LMM was performed to identify SNPs that are associated with OSCC. This is to obtain more genetic data from microarray studies. Genotype imputation involves predicting the genotypes of untyped markers based on the information provided by the typed markers that were genotyped, such as linkage disequilibrium (LD) and haplotypes. Even individuals who are not directly related can share common genomic regions inherited from a shared ancestor if they have similar ancestry. Consequently, imputation depends on the assumption that the reference sample used is a recent ancestor of the sample being imputed. This ensures that the haplotype blocks in the reference sample closely resemble those in the imputed sample, leading to a more accurate imputed dataset. The resulting imputed dataset contains a greater number of variants, which can be incorporated into studies, thereby enhancing the power of GWAS. This is especially crucial when sample sizes are limited (Naj, 2019).

Genotype imputation can be done using the Sanger Imputation Service. Genotype imputation through the Sanger Imputation Service comprises several essential steps to ensure precise inference of untyped markers. Initially, researchers submit their genotyped data, which undergoes a quality control process to eliminate low-quality SNPs and samples. Following this, the service employs a reference panel, usually derived from extensive population datasets, to utilize existing haplotype information and linkage disequilibrium (LD) patterns. The submitted data is phased to reconstruct haplotypes, enabling the Sanger service to infer missing genotypes based on shared ancestry and genomic regions. Advanced statistical algorithms are then applied during the imputation process to estimate the likelihood of unobserved genotypes, resulting in a more extensive dataset that encompasses many additional variants. After imputation, a quality control step ensures that the inferred genotypes meet specific confidence thresholds before providing researchers with the final imputed dataset, which is ready for association testing and analysis of the phenotype of interest (Hui *et al.*, 2020).

A small sample size can reduce the statistical power of a GWAS, making it difficult to detect associations, particularly for SNPs with small effect sizes. Increasing the sample size enhances the chances of identifying significant associations. Additionally, many complex traits have heritable components that may not be captured by common SNPs. This missing heritability can result from rare

variants, structural variants, or gene-environment interactions that are not included in the GWAS (Alquah et al., 2020). Furthermore, the impact of environmental factors or the interactions between genotype and environment may not be fully accounted for, which could influence the observed associations. If environmental exposures are overlooked, it may obscure the genetic effects (Chen et al., 2023).

#### ***5.3.4. Population Stratification and Genomic Inflation***

The Quantile-Quantile (QQ) plot (Figure 7) shows that substantial genomic inflation, which would otherwise distort the association signals, does not affect the GWAS results. The observed p-values' alignment with the expected values particularly in the lower range indicates that the population stratification was well controlled for in the study design, reducing the likelihood of false-positive results (Chen et al., 2023). Population stratification was properly accounted for since the LMM model genomic inflation factor lambda was 1.01 (Figure 7). The SNPs with suggestive associations, or those with a few deviations at the higher end of the QQ plot, suggest that these polymorphisms may indeed have an impact on OSCC risk.

In this study, the exclusion of individuals with potential population substructure (as identified by PCA) and the use of an LMM approach helped account for relatedness and ancestry, which are crucial for avoiding spurious associations in populations with high genetic diversity, such as African populations. The absence of early deviation in the QQ plot provides confidence in the robustness of the GWAS findings.

### **5.4. Polygenic Risk Scores**

The analysis of Polygenic Risk Scores (PRS) for oral squamous cell carcinoma (OSCC) in the South African population utilized two distinct methods: PRSice-2 and PRSsx. Both analyses were conducted on a target population consisting of 506 OSCC cases and 965 controls, providing insights into the predictive performance of the PRS in this specific cohort.

#### ***5.4.1. PRS models comparison***

The PRS model generated by PRSice-2 accounted for approximately 1.4% of the variance in the dataset, while the model derived from PRSsx explained about 1% of the variance. This indicates that PRSice-2 is slightly more effective in capturing the genetic contribution to OSCC in the South African population compared to PRSsx. As illustrated in Figures 8A and 9A, both methods displayed minimal



differentiation between cases and controls. In PRSice-2, cases exhibited a higher average standardized PRS compared to controls, although the difference was not substantial. PRScsx also demonstrated a slight increase in PRS for cases but with considerable overlap in distributions, further underscoring the challenge of distinguishing between the two groups.

The Area Under the Curve (AUC) for the PRSice-2 model was 0.6 (Figure 8B), indicating a moderate ability to differentiate between cases and controls. Conversely, PRScsx achieved an AUC of 0.56 (Figure 9B), suggesting a lower, yet still above-random, predictive ability. The AUC values indicate that while both models struggle to effectively classify individuals, PRSice-2 performs better in this regard. Considering the metrics assessed, PRSice-2 emerges as the superior method for analysing PRS in this specific South African OSCC GWAS. Its slightly higher variance explained (1.4% vs. 1%), better differentiation between cases and controls, and higher AUC (0.6 vs. 0.56) collectively suggest that PRSice-2 offers a more robust model for predicting OSCC risk in this population.

#### **5.4.2. PRS models comparison in literature**

According to the literature, PRScsx generally performs better than PRSice-2 in African populations, particularly due to the unique genetic diversity and structure of these groups (Ruan *et al.*, 2022). Studies assessing polygenic risk scores (PRS) across different ancestries have shown that PRScsx achieves higher predictive accuracy in African populations compared to PRSice-2. For example, one study revealed that PRS derived from European GWAS data performed poorly in individuals of African ancestry when using PRSice-2, whereas PRScsx significantly improved performance ( $R^2$  values of 0.055 versus 0.0032 for African ancestry) (Ruan *et al.*, 2022).

PRScsx is specifically designed to manage the complexities of genetic diversity and population stratification more effectively than PRSice-2. It integrates local ancestry information and can draw on genetic data from multiple ancestries, improving its predictive power in admixed populations like those of African descent. The trans-ancestry portability issue is well-documented, with PRS based on European ancestry data often showing limited accuracy in non-European populations. PRScsx addresses this challenge more successfully by incorporating shared genetic effects across different ancestries, which is critical for enhancing PRS accuracy in African cohorts (Ruan *et al.*, 2022).

While PRSice-2 can work well with European populations, its performance tends to decline when applied to African populations due to differences in linkage disequilibrium and allele frequencies. This

highlights the importance of using methods like PRScsx that are better suited for diverse genetic backgrounds (Ruan *et al.*, 2022).

#### **5.4.3. Factors that could have influenced the PRS models**

Including covariates like sex and principal components in PRSice-2 can significantly improve the model's accuracy by reducing the impact of confounding factors, resulting in more reliable predictions. On the other hand, if PRScsx did not incorporate these covariates, it may have generated biased estimates, reducing its predictive power. However, PRScsx, by design, models population-specific allele frequencies and linkage disequilibrium patterns, allowing it to maintain strong performance even without covariate adjustments (Ruan *et al.*, 2022).

Research has demonstrated that PRSice-2 often performs well when applied to well-curated datasets with appropriate covariate adjustments. One study found that PRSice-2 achieved better predictive accuracy when controlling for covariates compared to cases where no adjustments were made. In African populations, genetic diversity presents challenges for polygenic risk scoring, but PRSice-2's ability to incorporate local ancestry and adjust for population stratification through covariate control can improve performance (Keat *et al.*, 2023).

The fact that PRScsx and PRSice-2 explain only 1% and 1.4% of the variance in the data, respectively, can be attributed to several factors, particularly the absence of genotype imputation. Without genotype imputation, many variants that could influence the polygenic risk score (PRS) are excluded from the analysis. Imputation helps infer the genotypes of untyped variants by leveraging linkage disequilibrium (LD) patterns, greatly increasing the number of SNPs available for analysis and boosting the power of GWAS. Without imputation, many potentially informative variants are missed, which results in lower explained variance. Imputation fills in missing data points using LD patterns, expanding SNP coverage and improving the ability to detect associations (Naj, 2019). Since imputation was not performed in this case, the models have fewer SNPs to analyse, limiting their predictive power and reducing their capacity to explain genetic variance.

OSCC is influenced by both genetic and environmental factors. Without considering environmental influences, such as tobacco and alcohol consumption, the genetic contribution might seem smaller. Additionally, gene-environment interactions, which may not be fully captured by the current PRS models, could account for some of the missing variance (Sugimura *et al.*, 2005). Furthermore, the sample size in this study may not be sufficient to capture the complete genetic architecture of the

phenotype. Large sample sizes are typically required in GWAS to identify variants with small effect sizes, and smaller datasets often lead to lower variance explained by PRS due to reduced statistical power (Soo *et al.*, 2023).

To address the low variance explained by PRSice-2 and PRScsx in the current analysis, several strategies can be implemented to improve predictive accuracy and tackle the challenges associated with missing genetic data and environmental factors. One key solution is to perform genotype imputation using high-quality reference panels, such as the 1000 Genomes Project or the Haplotype Reference Consortium (HRC). Imputation helps infer untyped genotypes by leveraging linkage disequilibrium patterns, thereby increasing the number of SNPs available for analysis. This broader SNP coverage would enhance the power of the study, potentially leading to a greater proportion of variance explained by the PRS as more SNPs would be used to explain OSCC (Hui *et al.*, 2020).

Additionally, increasing the sample size by combining data from similar studies or collaborating with international consortia could significantly boost statistical power. Conducting meta-analyses across multiple cohorts would improve the detection of associations with SNPs that have small effect sizes (Nariman *et al.*, 2020). A larger sample size would enhance the statistical power, allowing for more robust findings and a higher variance explained by the PRS. PRScsx may also be correcting for the slight population substructure that comes with pooling together participants from different parts of South Africa. Although the PCA plot showed no visible stratification, there may be underlying differences that the PRS programme is picking up on and correcting for, resulting in a lower  $r^2$ .

Incorporating environmental covariates such as smoking, and alcohol use, as well as modelling gene-environment interactions, could also improve the analysis. Including these factors in both GWAS and PRS models would help account for additional variation that might be overlooked when focusing solely on genetic components. This approach provides a more comprehensive understanding of the combined genetic and environmental influences on OSCC risk.

## **Chapter 6: CONCLUSION**

This study identified 11 SNPs with suggestive associations with OSCC in a South African population using a GWAS with an LMM. Although none of the SNPs reached genome-wide significance ( $p < 5 \times 10^{-8}$ ), their p-values below  $5 \times 10^{-6}$  suggest potential involvement in OSCC risk. These findings

underscore the importance of exploring genetic susceptibility in African populations, which are often underrepresented in genetic studies.

Several of the identified SNPs are located near genes involved in critical biological processes such as cell signalling, regulation of gene expression, and cellular growth. Notably, SNPs with higher odds ratios, such as rs116279952 in *LOC124904156* and rs1171728 near *C10orf143*, highlight their potential role as high-risk variants for OSCC, while others, such as rs7220255 near *LINC01993*, suggest protective effects. These results provide insights into potential genetic variants contributing to OSCC susceptibility in African populations.

The PRS analysis for OSCC in the South African population revealed modest predictive power using both PRSice-2 and PRScsx models. PRSice-2 accounted for 1.4% of the variance, outperforming PRScsx, which explained 1%. Both models demonstrated limited ability to distinguish between cases and controls, with PRSice-2 achieving a slightly higher AUC of 0.6 compared to 0.56 for PRScsx.

Despite these modest results, the study highlights the potential for polygenic risk scores to provide insights into genetic susceptibility in African populations, though improvements such as genotype imputation and the incorporation of environmental factors are necessary. These steps could enhance the performance of PRS models, making them more effective tools for understanding the genetic risk of OSCC in diverse populations.

To fully understand the role of these variants in OSCC risk, further studies are necessary. Fine-mapping techniques could help pinpoint causal variants by narrowing down the region surrounding the SNPs and identifying those with the strongest association. Genotype imputation is essential for increasing SNP coverage in the PRS models, and combining data from larger, multi-ethnic cohorts could enhance the statistical power of future analyses. Incorporating gene-environment interactions will provide a more comprehensive understanding of how genetic predisposition and environmental factors, such as smoking and alcohol use, jointly influence OSCC risk. Replication studies in diverse African populations and meta-analyses are also critical for validating these findings and understanding the genetic architecture of OSCC in African cohorts.

## References

- Abdellaoui, A., Yengo, L., Verweij, K.J. and Visscher, P.M., 2023. 15 years of GWAS discovery: realizing the promise. *The American Journal of Human Genetics*, 110(2), pp.179-194.
- Abnet, C.C., Freedman, N.D., Hu, N., Wang, Z., Yu, K., Shu, X.O., Yuan, J.M., Zheng, W., Dawsey, S.M., Dong, L.M. and Lee, M.P., 2010. A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics*, 42(9), pp.764-767.
- Ahmad, W.M.A.W., Ghazali, F.M.M. and Yaqoob, M.A., 2023. *Basic Statistical Analysis Using RStudio Software*. Penerbit USM.
- Brandenburg, J.T., Clark, L., Botha, G., Panji, S., Baichoo, S., Fields, C. and Hazelhurst, S., 2022. H3AGWAS: a portable workflow for genome wide association studies. *BMC bioinformatics*, 23(1), p.498.
- Carraro, R.S., Nogueira, G.A., Sidarta-Oliveira, D., Gaspar, R.S., Dragano, N.R., Morari, J., Bobbo, V.C., Araujo, E.P., Mendes, N.F., ZanESCO, A.M. and Tobar, N., 2021. Arcuate nucleus overexpression of NHLH2 reduces body mass and attenuates obesity-associated anxiety/depression-like behavior. *Journal of Neuroscience*, 41(48), pp.10004-10022.
- Chang, J., Zhong, R., Tian, J., Li, J., Zhai, K., Ke, J., Lou, J., Chen, W., Zhu, B., Shen, N. and Zhang, Y., 2018. Exome-wide analyses identify low-frequency variant in CYP26B1 and additional coding variants associated with esophageal squamous cell carcinoma. *Nature Genetics*, 50(3), pp.338-343.
- Cheetham, S.W., Faulkner, G.J. and Dinger, M.E., 2020. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nature Reviews Genetics*, 21(3), pp.191-201.
- Chen, W.C., Brandenburg, J.T., Choudhury, A., Hayat, M., Sengupta, D., Swiel, Y., de Villiers, C.B., Ferndale, L., Aldous, C., Soo, C.C. and Lee, S., 2023. Genome-wide association study of esophageal squamous cell cancer identifies shared and distinct risk variants in African and Chinese populations. *The American Journal of Human Genetics*, 110(10), pp.1690-1703.
- Choi, S.W. and O'Reilly, P.F., 2019. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*, 8(7), p.giz082.
- Come, J., Castro, C., Morais, A., Cossa, M., Modcoicar, P., Tulsidás, S., Cunha, L., Lobo, V., Morais, A.G., Cotton, S. and Lunet, N., 2018. Clinical and pathologic profiles of esophageal cancer in Mozambique: a study of consecutive patients admitted to Maputo Central Hospital. *Journal of Global Oncology*, 4, pp.1-9.

- Cook, J.P., Mahajan, A. and Morris, A.P., 2016. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *European Journal of Human Genetics*, 25(2), pp.240-245.
- Cook, J.P., Mahajan, A. and Morris, A.P., 2020. Fine-scale population structure in the UK Biobank: implications for genome-wide association studies. *Human Molecular Genetics*, 29(16), pp.2803-2811.
- Cui, R.I., Kamatani, Y., Takahashi, A., Usami, M., Hosono, N., Kawaguchi, T., Tsunoda, T., Kamatani, N., Kubo, M., Nakamura, Y. and Matsuda, K., 2009. Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*, 137(5), pp.1768-1775.
- Ferndale, L., Ayeni, O.A., Chen, W.C., Aldous, C. and Thomson, S.R., 2023. Development and internal validation of the survival time risk score in patients treated for oesophageal cancer with palliative intent in South Africa. *South African Journal of Surgery*, 61(1), pp.36-44.
- Halec, G., Schmitt, M., Egger, S., Abnet, C.C., Babb, C., Dawsey, S.M., Flechtenmacher, C., Gheit, T., Hale, M., Holzinger, D. and Malekzadeh, R., 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), pp.85-98.
- He, C., Wu, X., Lin, L., Liu, C., Li, M., Jiang, C., Xu, Z. and Fang, B., 2023. Causal relationship between atrial fibrillation and stroke risk: a Mendelian randomization. *Journal of Stroke and Cerebrovascular Diseases*, 32(12), p.107446.
- He, S., Xu, J., Liu, X. and Zhen, Y., 2021. Advances and challenges in the treatment of esophageal cancer. *Acta Pharmaceutica Sinica B*, 11(11), pp.3379-3392.
- Hui, R., D'Atanasio, E., Cassidy, L.M., Scheib, C.L. and Kivisild, T., 2020. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific Reports*, 10(1), p.18542.
- Hull, R., Mbele, M., Makhafola, T., Hicks, C., Wang, S.M., Reis, R.M., Mehrotra, R., Mkhize-Kwitshana, Z., Hussain, S., Kibiki, G. and Bates, D.O., 2020. A multinational review: oesophageal cancer in low to middle-income countries. *Oncology letters*, 20(4), pp.1-1.
- Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D.J., Martin, I., Kullo, I.J., Kenny, E.E., Pasaniuc, B., Polygenic Risk Methods in Diverse Populations (PRIMED) Consortium Methods Working

Group Auer Paul L. 20 Conomos Matthew P. 21 Conti David V. 22 23 Ding Yi 24 Wang Ying 19 25 26 Zhang Haoyu 27 28 Zhang Yuji 29, Witte, J.S. and Ge, T., 2024. Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, 25(1), pp.8-25.

Kamiza, A.B., Toure, S.M., Vujkovic, M., Machipisa, T., Soremekun, O.S., Kintu, C., Corpas, M., Pirie, F., Young, E., Gill, D. and Sandhu, M.S., 2022. Transferability of genetic risk scores in African populations. *Nature Medicine*, 28(6), pp.1163-1166.

Keat K, Hui D, Xiao B, Bradford Y, Cindi Z, Daar ES, Gulick R, Riddler SA, Sinxadi P, Haas DW, Ritchie MD. Leveraging Multi-Ancestry Polygenic Risk Scores for Body Mass Index to Predict Antiretroviral Therapy-Induced Weight Gain. *Pac Symp Biocomput.* 2023;28:233-244.

Kurniansyah, N., Goodman, M.O., Khan, A.T., Wang, J., Feofanova, E., Bis, J.C., Wiggins, K.L., Huffman, J.E., Kelly, T., Elfassy, T. and Guo, X., 2023. Evaluating the use of blood pressure polygenic risk scores across race/ethnic background groups. *Nature communications*, 14(1), p.3202.

Lewis, C.M. and Vassos, E., 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1), p.44.

Li, D.D., Deng, L., Hu, S.Y., Zhang, F.L. and Li, D.Q., 2020. SH3BGRL2 exerts a dual function in breast cancer growth and metastasis and is regulated by TGF- $\beta$ 1. *American journal of cancer research*, 10(4), p.1238.

Liu, J.Q., Liao, X.W., Wang, X.K., Yang, C.K., Zhou, X., Liu, Z.Q., Han, Q.F., Fu, T.H., Zhu, G.Z., Han, C.Y. and Su, H., 2020. Prognostic value of Glypican family genes in early-stage pancreatic ductal adenocarcinoma after pancreaticoduodenectomy and possible mechanisms. *BMC gastroenterology*, 20, pp.1-23.

Liu, W., Zhuang, Z., Wang, W., Huang, T. and Liu, Z., 2021. An improved genome-wide polygenic score model for predicting the risk of type 2 diabetes. *Frontiers in genetics*, 12, p.632385.

Liu, J., Yao, Y., Hu, Z., Zhou, H. and Zhong, M., 2019. Transcriptional profiling of long-intergenic noncoding RNAs in lung squamous cell carcinoma and its value in diagnosis and prognosis. *Molecular genetics & genomic medicine*, 7(12), p.e994.

Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M. and Daly, M.J., 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4), pp.584-591.

- Mason, B., Flach, S., Teixeira, F.R., Manzano Garcia, R., Rueda, O.M., Abraham, J.E., Caldas, C., Edwards, P.A. and Laman, H., 2020. Fbxl17 is rearranged in breast cancer and loss of its activity leads to increased global O-GlcNAcylation. *Cellular and Molecular Life Sciences*, 77, pp.2605-2620.
- Mathew, C.G., 2017. Abstract IA8: The genetics and genomics of African esophageal cancer. *Cancer Research*, 77(22\_Supplement), pp.IA8-IA8.
- Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A. and Gingeras, T.R., 2023. Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nature reviews Molecular cell biology*, 24(6), pp.430-447.
- Morante-Palacios, O. and Ballestar, E., 2021. shinyÉPICO: a graphical pipeline to analyze Illumina DNA methylation arrays. *Bioinformatics*, 37(2), pp.257-259.
- Morgan, E., Soerjomataram, I., Rumgay, H., Coleman, H.G., Thrift, A.P., Vignat, J., Laversanne, M., Ferlay, J. and Arnold, M., 2022. The global landscape of esophageal squamous cell carcinoma and esophageal adenocarcinoma incidence and mortality in 2020 and projections to 2040: new estimates from GLOBOCAN 2020. *Gastroenterology*, 163(3), pp.649-658.
- Mulder, N., Abimiku, A.L., Adebamowo, S.N., de Vries, J., Matimba, A., Olowoyo, P., Ramsay, M., Skelton, M. and Stein, D.J., 2018. H3Africa: current perspectives. *Pharmacogenomics and personalized medicine*, pp.59-66.
- Murphy, A.E., Schilder, B.M. and Skene, N.G., 2021. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. *Bioinformatics*, 37(23), pp.4593-4596.
- MWer, S., Dykes, D. and Polesky, H., 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids res*, 16(3), p.1215.
- Naj, A.C., 2019. Genotype imputation in genome-wide association studies. *Current Protocols in Human Genetics*, 102(1), p.e84.
- Nariman-Saleh-Fam, Z., Saadatian, Z., Nariman-Saleh-Fam, L., Ouladsahebmadarek, E., Tavakkoly-Bazzaz, J. and Bastami, M., 2020. An association and meta-analysis of esophageal squamous cell carcinoma risk associated with PLCE1 rs2274223, C20orf54 rs13042395 and RUNX1 rs2014300 polymorphisms. *Pathology & Oncology Research*, 26, pp.681-692.



Nikas, I., Giaginis, C., Petrouska, K., Alexandrou, P., Michail, A., Sarantis, P., Tsourouflis, G., Danas, E., Pergaris, A., Politis, P.K. and Nakopoulou, L., 2022. EPHA2, EPHA4, and EPHA7 expression in triple-negative breast cancer. *Diagnostics*, 12(2), p.366.

Privé, F., Vilhjálmsson, B.J., Aschard, H. and Blum, M.G., 2019. Making the most of clumping and thresholding for polygenic scores. *The American journal of human genetics*, 105(6), pp.1213-1221.

Ruan Y, Lin YF, Feng YA, Chen CY, Lam M, Guo Z; Stanley Global Asia Initiatives; He L, Sawa A, Martin AR, Qin S, Huang H, Ge T. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet*. 2022 May;54(5):573-580.

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S. and Gómez-Olivé, F.X., 2021. Genetic substructure and complex demographic history of South African Bantu speakers. *Nature communications*, 12(1), p.2080.

Sugimura, T., Kumimoto, H., Tohnai, I., Fukui, T., Matsuo, K., Tsurusako, S., Mitsudo, K., Ueda, M., Tajima, K., & Ishizaki, K. (2005). Gene–environment interaction involved in oral carcinogenesis: molecular epidemiological study for metabolic and DNA repair gene polymorphisms. *Journal of Oral Pathology & Medicine*, 34(10), 580-585.

Soo, C.C., Brandenburg, J.T., Nebel, A., Tollman, S., Berkman, L., Ramsay, M. and Choudhury, A., 2023. Genome-wide association study of population-standardised cognitive performance phenotypes in a rural South African community. *Communications biology*, 6(1), p.328.

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

Sutherland, R., Meeson, A. and Lowes, S., 2020. Solute transporters and malignancy: establishing the role of uptake transporters in breast cancer and breast cancer metastasis. *Cancer and Metastasis Reviews*, 39, pp.919-932.

Statello, L., Guo, C.J., Chen, L.L. and Huarte, M., 2021. Gene regulation by long non-coding RNAs and its biological functions. *Nature reviews Molecular cell biology*, 22(2), pp.96-118.

Tarazi, M., Chidambaram, S. and Markar, S.R., 2021. Risk factors of esophageal squamous cell carcinoma beyond alcohol and smoking. *Cancers*, 13(5), p.1009.

Then, E.O., Lopez, M., Saleem, S., Gayam, V., Sunkara, T., Culliford, A. and Gaduputi, V., 2020. Esophageal cancer: an updated surveillance epidemiology and end results database analysis. *World journal of oncology*, 11(2), p.55.

Thumbs, A., Borgstein, E., Vigna, L., Kingham, T.P., Kushner, A.L., Hellberg, K., Bates, J. and Wilhelm, T.J., 2012. Self-expanding metal stents (SEMS) for patients with advanced Esophageal cancer in Malawi: An effective palliative treatment. *Journal of surgical oncology*, 105(4), pp.410-414.

Truong, V.Q., Woerner, J.A., Cherlin, T.A., Bradford, Y., Lucas, A.M., Okeh, C.C., Shivakumar, M.K., Hui, D.H., Kumar, R., Pividori, M. and Jones, S.C., 2022. Quality control procedures for genome-wide association studies. *Current Protocols*, 2(11), p.e603.

Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T. and Posthuma, D., 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), p.59.

Uhlenhopp, D.J., Then, E.O., Sunkara, T. and Gaduputi, V., 2020. Epidemiology of esophageal cancer: update in global trends, etiology and risk factors. *Clinical journal of gastroenterology*, 13(6), pp.1010-1021.

Vilhjálmsen, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R. and Hayeck, T., 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4), pp.576-592.

Wang, L.D., Zhou, F.Y., Li, X.M., Sun, L.D., Song, X., Jin, Y., Li, J.M., Kong, G.Q., Qi, H., Cui, J. and Zhang, L.Q., 2010. Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies a susceptibility locus at PLCE1. *Nature genetics*, 42(9), pp.759-763.

Wu, C., Hu, Z., He, Z., Jia, W., Wang, F., Zhou, Y., Liu, Z., Zhan, Q., Liu, Y., Yu, D. and Zhai, K., 2011. Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature genetics*, 43(7), pp.679-684.

Yang, J., Liu, X., Cao, S., Dong, X., Rao, S. and Cai, K., 2020. Understanding esophageal cancer: the challenges and opportunities for the next decade. *Frontiers in oncology*, 10, p.1727.

Yang, X., Zhu, H., Qin, Q., Yang, Y., Yang, Y., Cheng, H. and Sun, X., 2015. Genetic variants and risk of esophageal squamous cell carcinoma: a GWAS-based pathway analysis. *Gene*, 556(2), pp.149-152.

Zhou, X., 2016. Gemma user manual. *University of Chicago, Chicago, IL*.