

https://dspace.org/

Germline mutation landscape of Oesophageal Squamous Cell Cancer in South African patients

Dangor, Zahraa

2025

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/48339

Dangor, Zahraa . (2025). Germline mutation landscape of Oesophageal Squamous Cell Cancer in South African patients [Master`s dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/48339

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

GERMLINE MUTATION LANDSCAPE OF OESOPHAGEAL SQUAMOUS CELL CANCER IN SOUTH AFRICAN PATIENTS

Zahraa Dangor

Student number: 2208826

Supervisor: Dr Wenlong Carl Chen

Co-supervisors: Dr Jean-Tristan Brandenburg and Prof Christopher Mathew



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A research report (in the format of a "submissible" paper) submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework

Johannesburg, 2025

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
DIVISION OF HUMAN GENETICS
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Hospital Street, Johannesburg, 2001 | PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 | [M] +27 78 080 8841 | [F] +27 11 489 9226 | [E] human.genetics@nhls.ac.za

30 November 2024

Re: Research report submitted to the Faculty of Health Sciences, University of the Witwatersrand, by Mrs Zahraa Dangor in partial fulfilment of the requirements for the degree Master of Science in Medicine (Genomic Medicine)

Dear examiner,

Thank you for agreeing to examine the research report submitted by Mrs Zahraa Dangor (student number 2208826) entitled "Germline Mutation Landscape of Oesophageal Squamous Cell Cancer in South African populations". This research report is being submitted by the candidate in the format of a "submissible" paper as per the Style Guide for Thesis, Dissertation and Research Reports (updated March 2016) by the University of the Witwatersrand Faculty of Health Sciences. This research report contributes 33% towards the final mark of the MSc (Med) Genomic Medicine degree for which the candidate is enrolled.

The candidate and her supervisors have chosen to submit the paper to the journal Frontiers in Genetics. This international journal is a peer-reviewed and is published by Frontiers Media SA. The submissible paper attached is therefore written in concordance with the author guidelines of the stipulated journal. These guidelines have been attached for your reference in the appendices section of this document. The only deviation from these guidelines is that the manuscript has been presented as one complete document (including all tables and figures) for ease of marking, rather than being submitted separately.

The research protocol has also been attached as an appendix for the purpose of providing an extended literature review and further contextualise the study. The protocol has already been assessed and approved by the Faculty of Health Sciences Post-Graduate Assessors committee and is therefore not for examination.

Please refer to the Table of Contents for a complete list of the contents provided in this research report.

Zahraa Dangor (candidate)

Dr Wenlong Carl Chen (supervisor)

Prof Christopher Mathew (co-supervisor)

Dr Jean-Tristan Brandenburg (co-supervisor)

Zahraa Dangor (candidate)

This report is intended solely to record the observations and/or opinion of the writer. It does not constitute a medico-legal report

Declaration

I, Zahraa Dangor, declare that this Research Report (in the format of a “submissible” paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Zahraa Dangor

Student number: 2208826

26th day of February 2025 in Johannesburg

Contribution of Candidate

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

I, Zahraa Dangor, student number 2208826, declare that this Research Report, entitled

"Germline mutation landscape of Oesophageal Squamous Cell Cancer in South African patients" is my own work and that I contributed significantly towards the research findings presented in the paper intended for publication below.



28 February 2025

Student Date

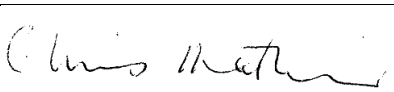


28 February 2025

Primary supervisor Date

Agreement by co-authors:

By signing this declaration, the co-authors listed below agree to the use of the research findings intended for a published article the student as part of their Research Report.

Authors	Name	Signature	Date
1st	Mrs Zahraa Dangor		28 February 2025
2nd	Dr Wenlong Carl Chen		28 February 2025
3rd	Dr Jean-Tristan Brandenburg		28 February 2025
4th	Prof Christopher Mathew		28 February 2025

Abstract

Oesophageal squamous cell cancer is highly prevalent in the Black South African population, yet its germline mutation landscape remains poorly understood. This study aimed to identify pathogenic germline variants in known cancer predisposition genes in a cohort of 16 Black South African patients diagnosed between the ages of 26 and 38. Whole exome sequencing data were generated using the Agilent SureSelect Human All Exome Capture assay (version 8) and sequenced on the Illumina NextSeq 2000 platform. Sequencing data were analysed using a curated set of 109 genes implicated in oesophageal squamous cell cancer, derived from literature, clinical panels, and the Fanconi Anaemia pathway. Variants were filtered and prioritised using a stepwise approach: exonic and splicing variants were retained, while downstream, upstream, intergenic, intronic, and synonymous single nucleotide variants were excluded. Population frequency filtering was applied using gnomAD exomes, removing variants with a minor allele frequency >1%. Variants classified as benign or likely benign based on American College of Medical Genetics and Association for Molecular Pathology guidelines were excluded, while pathogenicity was further assessed using in silico prediction tools. Seventeen candidate variants were identified in 14 genes, including 15 missense variants classified as variants of uncertain significance and two frameshift deletions classified as likely pathogenic. Notably, rs5742973, a “hot VUS,” was identified in *PMS1*, and two frameshift deletions were found in *MSH3* (rs766700228 and rs759756736). The findings highlight the potential involvement of the mismatch repair and Fanconi Anaemia pathways in oesophageal squamous cell cancer . This study highlights the importance of germline studies in underrepresented populations, providing a basis for future research on the genetic aetiology of oesophageal squamous cell cancer and its potential applications in risk prediction, early diagnosis, and targeted therapies.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Dr Carl Chen, Dr Jean-Tristan Brandenburg and Prof Christopher Mathew for their invaluable guidance, support, and encouragement throughout the course of this research. Their expertise and mentorship have been instrumental in shaping this work.

To my MSc classmates and lecturers, thank you for fostering a collaborative and stimulating environment that made this journey both enriching and enjoyable.

A special thanks to my friends and family for their unwavering support and belief in me. Your encouragement has been a constant source of motivation.

Finally, to my husband, thank you for your love, patience, and unyielding support throughout this year. Your strength and encouragement have been my anchor, and I am deeply grateful for everything you have done to help me persevere.

Table of Contents

Declaration iii
Contribution of Candidate iv
Abstract v
Acknowledgements vi
List of figures ix
List of tables ix
List of abbreviations x
1. Introduction 2
2. Materials and methods 4
2.1 Study participants 4
2.2Development of a curated set of cancer predisposition genes associated with OSCC 5
2.3 Whole exome sequencing 5
2.4 Data processing, variant calling and annotation 6
2.5 Variant filtering and prioritisation 8
2.6 Variant classification 10
3. Results 10
3.1. Development of curated gene set for Oesophageal Squamous Cell Cancer 10
3.2 Patient characteristics and Whole Exome Sequencing Summary Statistics 11
3.3 Identification and classification of candidate variants 12
3.3.1 Truncating variants classified as likely pathogenic 12
3.3.2 Missense variants classified as variants of uncertain significance 13
4. Discussion 14
5. Conclusion 18
Statement of ethics 18
Conflict of interest statement 18
Data availability statement 19
References 19
Appendix A: Supplementary Tables and Figures 25
Appendix B: Ethics Clearance Certificate 29
Appendix C: Approved Research Protocol 31

Appendix D: Plagiarism Declaration and Turnitin Report 32
Appendix E: Author guidelines 36

List of figures

Figure 1. Inclusion and exclusion criteria applied during bioinformatic analysis, with the number of variants shown at each step……8

Supplementary Figure 1. Script for data preprocessing in RStudio……24

List of tables

Table 1. Curated set of cancer predisposition genes derived from three sources……10

Table 2. VUSs identified in cancer predisposition genes……12

Table 3. Summary of germline variants and zygosity across patient samples harbouring more than one variant……13

Table 4. Likely pathogenic frameshift deletions……14

Supplementary Table 1. 72 variants remaining prior to pathogenicity prediction and classification……25

List of abbreviations

ACMG/AMP	American College of Medical Genetics and Genomics and Association for Molecular Pathology
FA	Fanconi Anaemia
FATHMM	Functional Analysis through Hidden Markov Models
GERP	Genomic Evolutionary Rate Profiling
gnomAD	Genome Aggregation Database
GWAS	Genome-wide association studies
JCS	Johannesburg Cancer Study
LP	Likely Pathogenic
MAF	Minor Allele Frequency
MMR	Mismatch Repair
OC	Oesophageal Cancer
OAC	Oesophageal Adenocarcinoma
OSCC	Oesophageal Squamous Cell Cancer
P	Pathogenic

PolyPhen2	Polymorphism Phenotyping v2
PROVEAN	Protein Variation Effect Analyzer
REVEL	Rare Exome Variant Ensemble Learner
SNV	Single Nucleotide Variants
UTRs	Untranslated Regions
WES	Whole-Exome Sequencing
VEST	Variant Effect Scoring Tool
VUS	Variant of uncertain significance

Research report in the format of a "submissible" paper

To be submitted to the Frontiers in Genetics journal.

Research Article

Germline Mutation Landscape of Oesophageal Squamous Cell Cancer in South African Patients

Zahraa Dangor[1], Jean-Tristan Brandenburg[2,3], Christopher G Mathew[2], Wenlong Carl Chen[2,3,4]

[1] Division of Human Genetics, National Health Laboratory Service; School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2] Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[3] Strengthening Oncology Services Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[4] National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa

Short Title: Analysis of Whole Exome Sequence Data from Black South African Patients With OSCC.

Corresponding Author: Zahraa Dangor

E-mail address: 2208826@students.wits.ac.za / zahraadangor@icloud.com

Keywords: Oesophageal Cancer, Oesophageal Squamous Cell Cancer, Whole Exome Sequencing, Germline Mutations, South Africa

Abstract for Journal Submission

Oesophageal squamous cell cancer is highly prevalent in the Black South African population, yet its germline mutation landscape remains poorly understood. This study aimed to identify pathogenic germline variants in known cancer predisposition genes in a cohort of 16 Black South African patients diagnosed between the ages of 26 and 38. Whole exome sequencing data were generated using the Agilent SureSelect Human All Exome Capture assay (version 8) and sequenced on the Illumina NextSeq 2000 platform. Sequencing data were analysed using a curated set of 109 genes implicated in oesophageal squamous cell cancer, derived from literature, clinical panels, and the Fanconi Anaemia pathway. Variants were filtered and prioritised using a stepwise approach: exonic and splicing variants were retained, while downstream, upstream, intergenic, intronic, and synonymous single nucleotide variants were excluded. Population frequency filtering was applied using gnomAD exomes, removing variants with a minor allele frequency >1%. Variants classified as benign or likely benign based on American College of Medical Genetics and Association for Molecular Pathology guidelines were excluded, while pathogenicity was further assessed using in silico prediction tools. Seventeen candidate variants were identified in 14 genes, including 15 missense variants classified as variants of uncertain significance and two frameshift deletions classified as likely pathogenic. Notably, rs5742973, a “hot VUS,” was identified in *PMS1*, and two frameshift deletions were found in *MSH3* (rs766700228 and rs759756736). The findings highlight the potential involvement of the mismatch repair and Fanconi Anaemia pathways in oesophageal squamous cell cancer . This study highlights the importance of germline studies in underrepresented populations, providing a basis for future research on the genetic aetiology of oesophageal squamous cell cancer and its potential applications in risk prediction, early diagnosis, and targeted therapies.

Keywords: Oesophageal Cancer, Oesophageal Squamous Cell Cancer, Whole Exome Sequencing, Germline Mutations, South Africa

1. Introduction

Oesophageal Cancer (OC) is the seventh most prevalent cancer worldwide, with over 600,000 new cases diagnosed each year. It is responsible for more than 540,000 deaths annually, ranking sixth in cancer mortality (Sung et al., 2021). Oesophageal cancer is characterised by two major histologic subtypes with distinct epidemiological and clinical features: oesophageal adenocarcinoma (OAC) and oesophageal squamous cell cancer (OSCC). Globally, OSCC is the more common type, contributing to approximately 90% of OC cases (Sung et al., 2021). Oesophageal squamous cell cancer is characterised by the growth of malignant tumours in the squamous epithelium that lines the oesophagus (Strickland et al., 2012). The histologic progression of OSCC involves the transition from normal squamous epithelium to hyperproliferative epithelium, low, intermediate and high-grade dysplasia followed by metastasis (Zhang and Guo, 2010).Non-specific clinical features often develop during advanced stages of the disease, with dysphagia being the most frequently reported symptom. Other symptoms include hoarseness, epigastric pain, heartburn and chest pain (Jiang et al., 2015). The largely asymptomatic disease progression and late clinical presentation result in late-stage diagnosis and poor prognosis.

OSCC is endemic in specific regions and exhibits the highest global divergence in cancer prevalence, with high incidence reported in China, as well as southern and eastern Africa (Abnet et al., 2018). Oesophageal cancer hotspots in Africa have been termed the African OSCC corridor and they include Ethiopia, Kenya, Malawi, Tanzania, and South Africa. According to Global Cancer Observatory (GLOBOCAN) 2020, Southern Africa has the second-highest incidence rate worldwide (Sung et al., 2021). Furthermore, the South African National Cancer Registry reported that 720 females and 942 males were histologically diagnosed with OC in 2020 alone, of which 79% (568) of females and 70% (663) of males were Black (NCR Cancer Statistics 2020). This highlights the high prevalence of this cancer in the Black South African population.

The aetiology of OSCC is multifactorial and marked by striking heterogeneity across populations worldwide (Murphy et al., 2017). The risk factors associated with this cancer

include several environmental, lifestyle and genetic factors (Huang and Yu, 2018). An improved understanding of the genetic causes of this cancer in Africa is of great importance.

To date, the genetic aetiology of OSCC in Africa has been primarily investigated through genetic association studies in the form of candidate gene association studies and genome-wide association studies (GWAS). In brief, Simba et al. (2019) reviewed genetic variants associated with OSCC in African populations, including 23 studies published between 1990 and 2019. These studies, consisting of 17 case-control genetic susceptibility studies, investigated 37 genes and identified 25 statistically significant variants across 20 genes (*ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, GSTT2B, MGMT, MLH3, MSH3, NAT2, PLCE1, PMS1, PTGS2, RUNX1, SLC11A, TP53*). Of these, 16 SNPs were considered potentially reliable, with 10 showing associations with increased OSCC risk in genes like *ALDH2, CASP8, CHEK2, CYP2E1, MGMT, MLH3, MSH3, PMS* and *RUNX1.* Chen et al. (2023) conducted the first African OSCC GWAS in a Black South African cohort (1,686 cases, 3,217 controls), discovering a novel risk locus on chromosome 9 near *FAM120A* (rs12379660, P = $4.58x10^{-8}$, OR=1.28, 95% CI=1.22-1.34) and an African-specific locus on chromosome 2 within *MYO1B* (rs142741123, P = $5.49x10^{-8}$, OR= 2.66, 95% CI=2.43–2.89). A meta-analysis with a Chinese OSCC GWAS identified three loci—*FAM120A* (chromosome 9), *PLCE1* (chromosome 10), and *CHEK2* (chromosome 22)—that reached genome-wide significance ($p < 5 \times 10^{-8}$), highlighting new potential genetic risk factors for OSCC.

These studies aimed to identify the common variants associated with this disease to better understand its genetic aetiology. However, association studies cannot directly identify functional causality variants that drive disease aetiology. No candidate gene list for OSCC has been defined to facilitate disease detection. A well-curated gene list is more likely to harbour causal mutations, reducing noise in analyses and increasing statistical power. Furthermore, leveraging prior knowledge through a candidate gene list can facilitate functional validation and improve efficiency in rare variant studies (MacArthur et al., 2014). This highlights the need for germline mutation analysis as the appropriate next step in African OSCC research.

Pathogenic variants in distinct DNA repair pathways increase susceptibility to various cancer types (Torgovnick and Schumacher, 2015).One such pathway is the Fanconi Anaemia (FA)

pathway, a genetically heterogeneous chromosomal instability syndrome associated with multiple congenital abnormalities, aplastic anaemia, and cancer (Morgan et al., 2005).Numerous studies have identified the FA pathway as a contributor to cancer predisposition in OSCC (Akbari et al., 2011; Zeng et al., 2021).Research into the molecular basis of FA in Black sub-Saharan African populations has revealed a founder mutation in the *FANCG* gene (NM_004629.1:c.637_643delTACCGCC), present in >80% of Black patients with FA. This highlights the importance of including FA pathway genes in OSCC studies to better understand this population's genetic mechanisms underlying OSCC susceptibility.

Whole-exome sequencing (WES) enables the investigation of cancer-related genetic aberrations that are predominantly located in the exonic regions of the genome (Bartha and Győrffy, 2019). This makes WES a suitable target and attractive technique for identifying germline mutations in known cancer predisposition genes. Germline mutation analysis in patients with OSCC has the potential to improve the understanding of the underlying mechanisms of OSCC and may provide information for the early detection, diagnosis, and risk management of this cancer (Deng et al., 2019). Identifying pathogenic germline variants is particularly crucial in patients with OSCC who present at an early age, as these individuals are often diagnosed at a younger age and tend to develop more aggressive cancer forms. The presence of germline mutations in known cancer predisposition genes is commonly linked to earlier onset and more severe disease outcomes, highlighting the importance of genetic testing in these cases (Goodwin et al., 2012).

Thus, this study aimed to identify pathogenic variants in the known cancer predisposition genes of Black African patients with OSCC. The objectives of this study included: (1) to conduct a targeted literature review to compile a list of cancer predisposition genes associated with OSCC; (2) to perform variant annotation, prioritisation and classification in the curated set of genes; (3) to identify pathogenic germline variants potentially implicated in the development of OSCC.

2. Materials and methods

2.1 Study participants

The study population for this study included 16 Black South African patients with histologically confirmed OSCC. These patients were selected from the Johannesburg Cancer Study (JCS), a

large case-control study conducted from 1995 through 2016. In brief, the JCS recruited newly diagnosed cancer patients of Black African ancestry to examine whether cancer risk factors identified in Western countries applied to the Black African population in Johannesburg (Chen et al., 2020).Upon enrolment in the JCS, these 16 patients received a clinical and pathological diagnosis of OSCC between the ages of 26-38. Young patients were selected for this study given the association between early age of cancer diagnosis and inherited cancer (Goodwin et al., 2012).

2.2Development of a curated set of cancer predisposition genes associated with OSCC

Germline WES data were generated from patients with histologically confirmed OSCC. Although the sequencing encompassed the entire exome, the analysis focused on a curated subset of cancer predisposition genes previously implicated in OSCC. This targeted approach reduced computational complexity and analysis time while providing flexibility for future investigations to include additional genes (Molina-Ramírez et al., 2022).

The curated gene set was developed by integrating three primary sources. First, a targeted literature review was conducted using PubMed. Search terms included “Oesophageal Squamous Cell Cancer,”, “Esophageal Squamous Cell Cancer” “germline mutations in Oesophageal Squamous Cell Cancer,” “Oesophageal Squamous Cell Cancer in Africa,” “aetiology of Esophageal Squamous Cell Carcinoma in African Populations,” “genetic factors associated with Oesophageal Squamous Cell Cancer,” and “cancer susceptibility genes,” combined with boolean operators to refine the search. Second, established diagnostic cancer gene panels, such as those from Invitae, were consulted to identify genes commonly associated with inherited cancer syndromes. Third, genes integral to the Fanconi Anaemia (FA) pathway, a critical DNA repair pathway linked to cancer predisposition, were specifically investigated, as germline variants in these genes have been associated with an elevated risk of OSCC (Akbari et al., 2011).The integration of these sources allowed for the development of a curated gene set tailored to investigating the germline mutation landscape of OSCC.

2.3 Whole exome sequencing

A typical WES workflow is divided into sample collection, DNA extraction, library preparation, exome capture, sequencing and data analysis (Bartha and Győrffy, 2019). Sample collection, DNA extraction, library preparation, exome capture and sequencing were performed prior to this study. In brief, sample collection and DNA extraction were performed for the JCS. Blood

samples were drawn from patients, and germline DNA was isolated using the Qiagen Flexigene DNA kit (Qiagen,USA) and stored at -30°C (Chen et al., 2018). The exonic regions of the genome were captured using the Agilent SureSelect Human All Exome Capture assay (version 8) (Agilent,USA).Sequencing was then performed on the Illumina NextSeq 2000 instrument (Illumina,USA). The sequencing data were provided for the analysis of variants in selected genes. This approach allowed for detailed bioinformatic analysis, with an emphasis on identifying relevant genetic variants.

2.4 Data processing, variant calling and annotation

Sequencing quality control (QC), alignment, and variant calling were conducted using the Illumina DRAGEN-GATK software (Behera et al., 2024). Following these steps, data preprocessing was performed to extract essential gene information such as gene IDs, gene names, transcript details, chromosomal locations, and other metadata from Ensembl using the biomaRt R package (Drost and Tjeldnes, 2015). The data, now enriched with relevant gene information, were analysed. Variants were annotated with ANNOVAR version (2023Aug30), a command-line driven tool designed for the functional annotation of genetic variants derived from high-throughput sequencing data (Wang et al., 2010).

Variant annotation involves adding auxiliary metadata and knowledge to quality-filtered raw putative variant calls to enhance the assessment of variants likely to impact function (Sefid Dashti and Gamieldien, 2017)Variant annotation integrates multiple databases and tools to provide detailed insights into genetic variants. This process begins with three primary gene annotation databases refGene (last updated 2020-08-22), knownGene (last update 2022-05-15), and ensGene (last update 2024-2024-05-13) (Wang et al., 2010). RefGene, derived from the NCBI RefSeq database, provides highly curated annotations that serve as a reliable reference for clinical and research applications(Pruitt et al., 2012) . KnownGenes, sourced from the UCSC Genome Browser, integrates data from multiple resources for comprehensive annotations and is particularly useful for large-scale analyses (Karolchik et al., 2003). EnsGenes, maintained by Ensembl, employs automated computational analysis to deliver broad, up-to-date annotations suited for comparative genomics and large-scale studies (Hubbard et al., 2002).

The annotation process generates different outputs that offer detailed information about geneticvariants(https://annovar.openbioinformatics.org/en/latest/userguide/startup/#:~:tex

t=The%20rest%20of%20the%20columns,HVAR%20scores%2C%20LRT%20scores%2C%20MutationTaster.) . Basic Information (variant id, chr, start, end, ref, alt) describes the variant's location and nucleotide changes. Functional Annotations were derived from RefSeq (*funcrefgene, generefgene, genedetailrefgene, exonicfuncrefgene, aachangerefgene*) classified the variant's location within the genome (e.g., exonic, intronic) and its functional effects, such as nonsynonymous coding changes or amino acid alterations (Pruitt et al., 2012).

Population Frequency Data (*gnomad40_exome_af, esp6500siv2_all,1000g2015aug_all*) provided allele frequencies from global and population-specific datasets. For example, gnomAD highlights variant frequency in diverse populations, aiding in the identification of rare variants that are more likely to be pathogenic (Chen et al., 2024). Protein Function Prediction Scores (*sift_score, fathmm_score, provean_score*) predict whether variants impact protein function (Ng and Henikoff, 2003; Choi et al., 2012; Shihab et al., 2015), while Meta-prediction Scores (*metasvm_score, metalr_score*) integrate results from multiple tools for a robust pathogenicity assessment (Kim et al., 2017; Chen et al., 2023).

To integrate and refine the outputs of individual in silico predictors, the REVEL (Rare Exome Variant Ensemble Learner) tool was employed. This approach is particularly advantageous, as individual tools often yield conflicting results due to their use of different predictive features. REVEL, an ensembl-based model, combines predictions from multiple established tools including MutPred, Functional Analysis through Hidden Markov Models (FATHMM), Variant Effect Scoring Tool (VEST), Polymorphism Phenotyping v2 (PolyPhen2), Sorting Intolerant from Tolerant (SIFT), Protein Variation Effect Analyzer (PROVEAN), MutationAssessor, MutationTaster and Genomic Evolutionary Rate Profiling (GERP). It is designed to generate a comprehensive meta-score that reflects the likelihood of a variant being pathogenic. By consolidating data from various tools, REVEL provides improved accuracy compared to single-tool assessments (Ioannidis et al., 2016). REVEL scores are calculated using dbNSFP, a database designed for the functional prediction and annotation of all possible non-synonymous single-nucleotide variants in the human genome (Liu et al., 2020) . REVEL produces a score for a missense variant of between 0 and 1, with larger scores indicating a higher chance that the variant is pathogenic (Chen et al., 2024).

Clinical Predictions (*clnsig, clndn*) provide information on variant significance from resources like ClinVar (Landrum et al., 2013). Conservation Scores (*phylop100way_vertebrate,*

phastcons100way_vertebrate) highlight evolutionary constraints, indicating functionally important regions. Additional tools such as GTEx Annotations (*gtex_v8_gene, gtex_v8_tissue*) link variants to tissue-specific gene expression (Lonsdale et al., 2013), while Structural and Functional Domain Information (*interpro_domain*) identifies associations with protein domains (Apweiler et al., 2001). This integrated framework of metadata and tools enabled detailed functional and clinical interpretations of genetic variants.

2.5 Variant filtering and prioritisation

We prioritised variants to generate a shortlist of germline variants in the cancer predisposition genes. A stepwise filtering approach was applied to perform this step as depicted in Figure 1. The proposed strategy that was employed was compiled using previously described strategies (Sefid Dashti and Gamieldien, 2017; Nissen et al., 2018). Each step includes one exclusion criterion that will be applied to filter out variants.

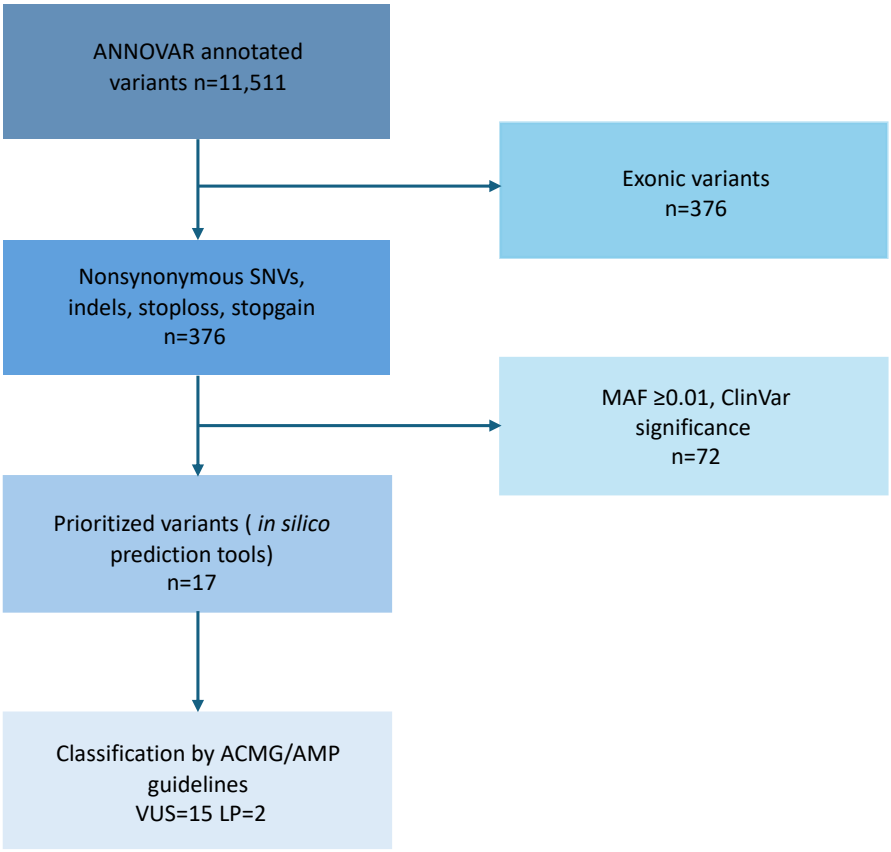


Figure 1. Inclusion and exclusion criteria applied during bioinformatic analysis, with the number of variants shown at each step.

To assess the pathogenicity of the identified variants, a thorough multi-step filtering strategy was applied to prioritize variants that are both functionally significant and clinically relevant,

focusing specifically on the subset of cancer predisposition genes. First, functional categories were used to retain only exonic and splice site variants, as these are more likely to affect gene function by altering protein sequences or disrupting mRNA splicing. Variants in non-coding regions, such as upstream, downstream, untranslated regions (UTRs), introns (except those at splice sites), and intergenic regions, were excluded due to their generally lower likelihood of directly impacting protein function. Next, synonymous single nucleotide variants (SNVs) and non-splicing intronic variants were removed, as synonymous mutations typically do not affect protein structure and function unless they impact splicing or regulatory elements. Variants in introns that do not affect splice sites were similarly de-prioritised.

To focus on rare, potentially pathogenic variants, population allele frequency filtering was applied using gnomAD exomes v4 African population data (Chen et al., 2024). Variants with a Minor Allele Frequency (MAF)>1% in the African population were excluded, as they are more likely to represent benign polymorphisms rather than disease-causing mutations. Given the study population of Black South Africans, this step ensured that only variants with low population frequency, and thus a higher probability of pathogenicity, were retained. Despite the high prevalence of OSCC in this population, rare variants were prioritised as they are more likely to represent high-penetrance risk factors that significantly increase individual susceptibility.

Following this step, missense variants and frameshift and non-frameshift indels were analysed separately to avoid inadvertently filtering out these indels, as frameshift and non-frameshift variants lacked pathogenicity scores.

Further refinement was achieved using ClinVar significance to assess the clinical significance of the variants. Variants classified as likely benign or benign were excluded from further analysis, while variants classified as uncertain significance (VUS), likely pathogenic (LP), or pathogenic (P) were retained for further investigation. The variants that remained after applying the filtering criteria up to this point are presented in Supplementary Table 1. Lastly, variants with a low pathogenicity prediction were excluded using the REVEL score.

By combining these filters based on functional categories, functional impact within exonic regions, population frequency, clinical significance from ClinVar, and computational predictions, a refined list of potentially pathogenic germline variants in OSCC was generated for downstream analysis.

2.6 Variant classification

Once a shortlist of missense variants and indels was generated, they were classified according to the American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) guidelines (Richards et al., 2015) using the InterVar webserver and Varsome, respectively. The InterVar webserver (version last updated 2022-June-13 17:57:28), also known as wInterVar, provides an intuitive graphical user interface for variant interpretation (https://wintervar.wglab.org/) . Missense variants were queried by entering their chromosomal position, dbSNP identifier, or gene name, along with the corresponding nucleotide change. The webserver automatically generates and displays detailed information for each variant, including the classification criteria, supporting evidence, and relevant sub-population data. Additionally, the classification criteria can be manually adjusted, allowing for variant re-interpretation and resubmission for further analysis (Li and Wang, 2017). Varsome is a variant search engine for human genomic variation, aggregates data from multiple genomic databases and reports variant pathogenicity through an automated classification system (https://varsome.com/) (Ioannidis et al., 2016). The frameshift deletions were queried on Varsome version 12.6.1 by entering their chromosomal position or dbSNP identifier to assess their pathogenicity.

3. Results

Seventeen candidate variants were identified in 14 cancer predisposition genes across 16 Black South African OSCC patients. These included 15 missense variants classified as VUSs and two frameshift deletions classified as likely pathogenic. Notably, a “hot VUS” (rs5742973) was identified in *PMS1*, while two frameshift deletions in *MSH3* suggest a role for the mismatch repair and Fanconi Anaemia pathways in OSCC susceptibility.

3.1. Development of curated gene set for Oesophageal Squamous Cell Cancer

Variants in a curated set of cancer predisposition genes were analysed to investigate the germline mutation landscape of OSCC. This set was derived from three sources as shown in Table 1. The Invitae Multi-Cancer Panel contributed 70 genes to the analysis, selected due to Invitae’s reputation as an established clinical diagnostic laboratory with expertise in developing highly validated cancer panels (https://www.invitae.com/us/providers/test-catalog/test-01101).From the targeted literature search, 28 genes were identified as being

associated with or implicated in the development of OSCC. The FA pathway, a critical DNA repair mechanism, contributed 21 genes to the curated set.

Table 1| Curated set of cancer predisposition genes derived from three sources

Source	Genes
Invitae Multi-Cancer Panel	*AIP,ALK,APC,ATM,AXIN2,BAP1,BARD1,BLM,BMPR1A,BRCA1,BRCA2,BRIP1 CDC73, CDH1,CDK4,CDKN1B,CDKN2A,CHEK2, CTNNA1,DICER1,EGFR, EPCAM,FH,FLCN, GREM1,HOXB13,KIT,LZTR1,MAX, MBD4,MEN1,MET, MITF,MLH1,MSH2,MSH3,MSH6, MUTYH,NF1,NF2,NTHL1,PALB2,PDGFRA, PDGFRA,PMS2,POLD1,POLE,POT1, PRKAR1A, PTCH1,PTEN,RAD51C,RAD51D RB1,RET,SDHA,SDHAF2, SDHB,SDHC, SDHD,SMAD4,SMARCA,SMARCB1, SMARCE1,STK11,SUFU, TMEM127, TP53,TSC1,TSC2,VHL*
Literature	*ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, FAM120A, GJB2, GSTT2B, MGMT, MLH3, MSH3, MUTYH, MYO1B, NAT2, PLCE1, PMS, PMS1, PMS11, PMS2, PTGS2, RECQL4, RUNX1, SLC11A, TP53.*
FA pathway	*FANCA,FANCB,FANCC,BRCA2,FANCD2,FANCE,FANCF,FANCG,FANCI, BRIP1, FANCL,FANCM,PALB2,RAD51C,SLX4,ERCC4,RAD51,BRCA1,UBE2T,MAD2L2, RFWD3*

An overlap between the sources accounts for the reduction in the final gene count from 119 to 109 unique genes. Specifically, six genes, *CHEK2*, *MLH3*, *MSH3*, *MUTYH*, *PMS2*, and *TP53*, were shared between the Invitae Multi-Cancer Panel and the targeted literature search. Additionally, five genes, *BRCA1*, *BRCA2*, *BRIP1*, *PALB2*, and *RAD51C*, overlap between the Invitae Multi-Cancer Panel and the FA pathway. *BRIP1* also appears in both the FA pathway and the literature search. Focusing on this curated set of 109 cancer predisposition genes significantly reduced the complexity and computational burden of analysing the entire exome. At the same time, this approach maintained flexibility for future expansions to incorporate additional genes, enabling a tailored and scalable investigation of the genetic landscape of OSCC (Molina-Ramírez et al., 2022).

3.2 Patient characteristics and Whole Exome Sequencing Summary Statistics

A total of 16 Black South African patients, aged 26–38, with histologically confirmed OSCC were included in this study. All patients received a clinical and pathological diagnosis of OSCC, and their samples were anonymized and assigned unique identifiers. No additional clinical features or family history were available for these individuals.

Across the 16 patients, a total of 11,511 genetic variants were identified within the curated set of genes. Of these, 759 variants were located within coding regions. According to ANNOVAR annotations, these coding sequence variants primarily consisted of synonymous single nucleotide variants (SNVs, n=386), followed by missense variants (n=319), and insertions or deletions (indels, n=29).

3.3 Identification and classification of candidate variants

A shortlist of 17 candidate variants in 14 genes was generated and classified according to ACMG/AMP guidelines. Of the 14 genes identified in this study, four (*ATM*, *MSH3*, *POLE*, and *TMEM*) originated from the Invitae Multi-Cancer panel, six (*MARCHF9*, *MUTYH*, *NAT2*, *PMS1*, *PMS2*, and *RUNX1*) were derived from the targeted literature review and the remaining four (*FANCA*, *FANCC*, *FANCI*, and *FANCM*) were classified as part of the FA pathway

3.3.1 Truncating variants classified as likely pathogenic

Two truncating variants were identified in the *MSH3* gene, a known cancer predisposition gene, in a single patient (Patient ID: 11749), as shown in Table 4. rs76700228, is a frameshift deletion located on chromosome 5 at position 80654930. This alteration involves the deletion of four base pairs (TCCC) in the coding region, leading to a frameshift starting at position 705 and resulting in a premature stop codon. It has been classified as likely pathogenic based on the ACMG/AMP evidence code PVS1, which indicates very strong pathogenicity due to the null effect of the frameshift on the gene product. rs75975673, is another frameshift deletion located at position 80654937 on chromosome 5. This variant involves deletion of seven base pairs (TTCCCGC), causing a frameshift and subsequent early termination of the protein. It is also classified as likely pathogenic, supported by PVS1 (very strong pathogenicity) and PM2 (moderate pathogenicity based on its rarity in the population). Both variants have MAFs, derived from gnomAD exomes v4 African population data, significantly below 1% in the population, with MAFs of 0.0034 and 0.0020, respectively, providing additional evidence for their pathogenicity. These two heterozygous (0/1) variants were identified in the same individual (Patient ID: 11749), suggesting compound heterozygosity in *MSH3*. This genetic configuration indicates that each allele of the gene carries a pathogenic mutation, likely leading to a complete loss of protein function.

Table 4 | Likely pathogenic frameshift deletions							
SNP ID	**Gene**	**HGVS_P**	**Location (Hg38)**	**ACMG codes**	**MAF**	**Patient ID**	**Zygosity**
rs766700228	*MSH3*	p.A70Sfs*9	5:80654930	PVS1[1]	0.0034	11749	Heterozygous
rs759756736	*MSH3*	p.F71Pfs*7	5:80654937	PVS1, PM2[2]	0.0020	11749	Heterozygous

[1] *PVS1 Null variant in a gene where LOF is a known mechanism of disease.*

2 PM2 Absent from controls in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

3.3.2 Missense variants classified as variants of uncertain significance

A total of 15 missense variants were identified in 13 cancer predisposition genes across 11 patients, as shown in Table 2. Four of these patients harboured more than one variant, as detailed in Table 3. These variants were classified as variants of uncertain significance (VUSs) according to ACMG/AMP guidelines, with classification performed using InterVar. Supporting evidence included criteria such as PM1 (location in mutational hotspots or functional domains of known importance), PM2 (rarity in population databases), and BP4 (benign computational evidence) (Richards et al., 2015).The MAF of the identified variants, derived from gnomAD exomes v4 African population data, ranged from 0.00000 to 0.00230, highlighting their rarity within this population. REVEL scores, which predict variant pathogenicity, ranged from 0.129 to 0.824. Zygosity information is also provided, with 0/0 (major allele) and 1/1 (minor allele) indicating homozygosity, and 0/1 indicating heterozygosity. Notably, all identified variants were present in the heterozygous state.

Table 2| VUSs identified in cancer predisposition genes								
SNP ID	**Gene**	**HGVS_P**	**Location (Hg38)**	**ACMG codes**	**MAF**	**REVEL**	**Patient ID**	**Zygosity**
rs587782041	*MUTYH*	p.R98W	1:45333297	PM1,PM2,PP3,BP1	0.00003	0.642	9527	Heterozygous
rs5742973	*PMS1*	p.E27Q	2:189791888	PM1[1],PM2[2],PP3[3],BP1[4]	0.00230	0.824	3410	Heterozygous
rs114185660	*PMS2*	p.E202D	7:5992037	PM1,PM2,BP1[5]	0.00040	0.573	16695	Heterozygous
rs45477599	*NAT2*	p.L24I	8:18400073	PM2,BP4	0.00180	0.332	11749	Heterozygous
rs185822330	*FANCC*	p.R535C	9:95101781	PM1,PM2,BP1,BP4	0.00020	0.175	13054	Heterozygous
rs553326913	*FANCC*	p.A238P	9:95135477	PM1,PM2,BP1	0.00006	0.350	13054	Heterozygous
rs776596331	*TMEM*	p.Y57H	11:67469024	PM1,PM2	0.00160	0.255	10115	Heterozygous
rs150143957	*ATM*	p.D44G	11:108227834	PM1,PM2,BP1	0.00080	0.269	10923	Heterozygous
rs774466216	*MARCHF9*	p.T107I	12:57755848	PM1	0.00080	0.141	10115	Heterozygous
rs141519273	*POLE*	p.R1839C	12:132639162	PM1,PM2,BP1	0.00000	0.678	13054	Heterozygous
rs142747831	*FANCM*	p.P266A	14:45148951	PM1	0.00170	0.129	23312	Heterozygous
rs112345761	*FANCM*	p.V960I	14:45175710	PM1,PM2,BP4	0.00000	0.031	9527	Heterozygous

rs544848412	*FANCI*	p.E1243K	15:89315372	PM1,PM2,BP1	0.00009	0.186	16695	Heterozygous
rs376089640	*FANCA*	p.S1087L	16:89748747	PM1,PM2,BP1	0.00006	0.432	9992	Heterozygous
rs759227406	*RUNX1*	p.P251L	21:34799435	PM1,PM2,PP3,BP1	0.00006	0.491	11749	Heterozygous

[1]*PM1 Located in a mutational hot spot and/or critical and well-established functional domain without benign variation.*
[2]*PM2 Absent from controls in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium*
[3]*PP3 Multiple lines of computational evidence support a deleterious effect on the gene or gene product*
[4]*BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease*
[5]*BP4 Multiple lines of computational evidence suggest no impact on gene or gene product*

Table 3| Summary of germline variants and zygosity across patient samples harbouring more than one variant

Patient ID	Gene	Transcript ID	Exon number	HGVS_C	HGVS_P
13054	FANCC	NM_000136	15	c.C1603T	p.R535C
	FANCC	NM_000136	18	c.G712C	p.A238P
10115	*MARCHF9*	NM_138396	1	c.C320T	p.T107I
	FANCI	NM_018193	36	c.G3727A	p.E1243K
11749	*NAT2*	NM_000015	2	c.T70A	p.L24I
	RUNX1	NM_001001890	5	c.C752T	p.P251L
16695	*PMS2*	NM_001322008	7	c.G606C	p.E202D
	FANCI	NM_018193	36	c.G3727A	p.E1243K

A total of 17 variants were identified in 14 genes, including 15 missense variants and two frameshift deletions. The frameshift deletions were classified as likely pathogenic, while the missense variants were categorized as variants of uncertain significance. The identified variants were distributed across genes involved in DNA repair pathways, particularly the mismatch repair and Fanconi Anaemia pathways.

4. Discussion

This study describes a targeted exome analysis of the germline mutation landscape of OSCC in 16 Black South African patients. The high prevalence of OSCC in this population, combined with the early age of diagnosis in the selected cohort, highlights the need to investigate the germline mutation profiles associated with OSCC. Given the complexity of identifying germline variants across numerous cancer predisposition genes, this study focused on a curated set of genes linked to OSCC and other cancers as an initial framework for analysis. This approach identified many variants refined through filtering to 17 candidate variants across 14 genes.

Advances in genomic technologies have made WES faster, more affordable, and widely accessible. These developments have increased the use of sequencing in diverse populations and research settings. In cancer research, such approaches are particularly valuable for identifying germline mutations associated with inherited cancers. However, broader sequencing efforts inevitably detect many VUSs (Amano et al., 2022). Over time, as new data are generated, the clinical relevance of VUSs may be reassessed. Many are eventually reclassified, with most downgraded to benign and a smaller subset upgraded to pathogenic (Mersch et al., 2018).

The 15 missense VUSs identified in the selected cancer predisposition genes, represented a large proportion of these candidate variants. Similar WES OSCC studies have also detected a large proportion of VUSs in cancer predisposition genes (Deng et al., 2019). Variants of uncertain significance are genetic variants identified through sequencing whose impact on protein function and clinical significance remains unclear (Richards et al., 2015). Consequently, these variants cannot currently be used in clinical practice to predict cancer risk or guide treatment decisions (Lucci-Cordisco et al., 2022).

VUSs occur across all populations but are disproportionately prevalent in individuals of African ancestry. For example, a multi-centre study in the United States, which used the Invitae Multi-Cancer Panel to analyse 2,571 patients of European ancestry (EA) and 110 patients of African ancestry (AA) with solid tumours, reported similar rates of pathogenic variants in both groups (13.3% in EAs; 12.7% in AAs). However, the prevalence of VUSs was significantly higher in AAs (65.5%) compared to EAs (46.1%) (Samadder et al., 2021), and this has also recently been demonstrated in a high rate of VUS in the mixed ancestry and African ancestry populations of South Africa (Osler et al., 2024). This disparity highlights the challenges associated with of germline mutation analysis in individuals of African ancestry, as the higher prevalence of VUSs increases the complexity of the actionable insights provided by sequencing results. The difficulties of these analyses likely reflect the underrepresentation of African populations in genomic databases, and their greater genomic diversity, as well as a lack of tools to functionally validate VUSs(McDonald and Ricks-Santi, 2022). Addressing these limitations is essential to improving the relevance and accuracy of genomic testing for populations of African descent.

While the clinical significance of the identified missense variants remains unclear, one variant of particular interest is rs5742973, with a REVEL score of 0.824. This variant was classified as a 'hot VUS,' suggesting that it is potentially pathogenic and requires further investigation. A 'hot VUS' refers to a subclass of variants of uncertain significance considered particularly suspicious. Such variants are of high concern and may be classified as pathogenic if additional evidence supporting their pathogenicity is obtained (Joynt et al., 2021). The *PMS1* gene encodes a protein which forms part of the mismatch repair (MMR) system, responsible for correcting DNA mismatches and small insertions and deletions that may arise during DNA replication and homologous recombination (NCBI,2024). Polymorphisms in *PMS1* have been associated with an increased risk of OSCC in African populations (Vogelsang et al., 2012). Thus, mutations in *PMS1* can disrupt the MMR system, which may be linked to an increased risk of developing OSCC (Zeng et al., 2021).However, further evidence and the application of additional ACMG/AMP criteria are needed to determine this variants' clinical relevance.

Five of the candidate missense variants span multiple genes involved in the FA pathway including *FANCA, FANCC, FANCI* and *FANCM,* which are integral to maintaining genome stability and DNA repair. Genes in the FA pathway are involved in the repair of DNA interstrand cross-links (ICLs) (Mirchandani and D'Andrea, 2006). Pathogenic variants in these genes have been implicated in the development of OSCC (Akbari et al., 2011).Similar studies have reported that the FA pathway was the most significantly enriched pathway among pathogenically mutated cancer predisposition genes in OSCC, suggesting it as a primary pathway for OSCC predisposition (Zeng et al., 2021). Additionally, it is well established that squamous cell carcinomas of the head and neck (HNSCC) and oesophagus are among the most common solid tumours observed in patients with Fanconi anaemia (Alter, 2003). These findings have potential therapeutic implications for OSCC patients. Specifically, patients with heterozygous mutations in FA genes may benefit from treatment with DNA cross-linking chemotherapeutic agents. However, as previously noted, due to the uncertain significance of these variants, no definitive conclusions can be drawn regarding their clinical implications at this stage. Overall, these findings emphasise the importance of rare missense variants in the genetic landscape of OSCC within this cohort, with particular attention to variants like rs5742973 which may have clinical relevance in further studies.

The two frameshift deletions, rs766700228 and rs759756736, identified are protein truncating variants (PVT) which occurred in the same patient. The presence of compound heterozygosity in this patient calls for further investigation into the clinical significance of these variants. The *MSH3* gene encodes a protein that forms part of the MutSβ complex, a key component of the DNA MMR pathway. This pathway is critical for correcting replication errors and maintaining genomic stability (NCBI,2024.). Loss-of-function mutations in *MSH3* are associated with an increased risk of cancer due to elevated mutation rates, which can drive carcinogenesis (Graham et al., 2019). These findings highlight the potential impact of PTVs in the *MSH3* gene on DNA mismatch repair, suggesting an increased risk of developing cancers such as colorectal and breast cancers.

This study had several limitations. The underrepresentation of African genomic data in public databases contributed to the identification of a larger proportion of VUSs, but this is also driven by the greater genomic diversity of African populations . Additionally, given the strong association between a family history of OC and an increased risk for OSCC, an important consideration in inherited cancer studies is familial aggregation (Chen et al., 2015). However, family history was not available for the patients in this study, preventing the application of additional ACMG/AMP criteria, such as PP1 (supporting pathogenicity), which relies on evidence of co-segregation of variants within families.

Future studies can build upon this work to deepen our understanding of the germline mutation landscape of OSCC in African populations. The exome data generated offers the advantage of enabling the analysis of additional cancer predisposition genes, by building a more robust list of candidate genes. This could aid in the identification of pathogenic variants driving the disease aetiology. Additionally, including parents and/or affected family members in future studies will provide critical insights into segregation patterns and inheritance allowing for definitive conclusions regarding the pathogenicity of identified variants. A large-scale exome sequencing study of OSCC cases and controls would help to establish whether specific genes had higher mutation burdens in OSCC patients Finally, functional assays should be employed to clarify the significance of the VUSs identified and their potential role in tumorigenesis (Woods et al., 2016).

5. Conclusion

This study represents an important step toward understanding the germline genetic landscape of OSCC in Black South African patients, a population disproportionately affected by this cancer. By leveraging whole-exome sequencing and focusing on a curated set of cancer predisposition genes, we have identified several candidate variants, including two likely pathogenic frameshift deletions in the *MSH3* gene: rs76700228 and rs759756736. These variants, which disrupt the DNA MMR pathway, may significantly contribute to OSCC risk. Future studies should expand this research by analysing additional cancer predisposition genes, incorporating familial segregation data, and employing functional assays to validate the pathogenicity of the VUSs. These efforts will enhance our ability to interpret the genetic mechanisms underlying the development of OSCC and pave the way for precision medicine approaches in African populations.

Statement of ethics

This research study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, under the clearance certificate M240769. It was conducted as a sub-study under the main protocols M21/11/154 and M22/08/81. The ethical clearance certificate is s provided in Appendix B.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Funding

This study was supported by the National Research Foundation (NRF) Thuthuka funding. The funder had no role in the sign, data collection, data analysis and reporting of this study.

Author contributions

Z Dangor, WC Chen, JT Brandenburg and C Mathew designed the study. Z Dangor, WC Chen, JT Brandenburg and C Mathew contributed to the acquisition and interpretation of the data. Z Dangor analysed the data and drafted the manuscript. Z Dangor, WC Chen, JT Brandenburg and C Mathew revised the manuscript for important intellectual content. WC Chen, JT

Brandenburg and C Mathew supervised the research. All authors approved the final version of the manuscript.

Data availability statement

Data available upon request

References

Journals:

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

Akbari, M. R., Malekzadeh, R., Lepage, P., Roquis, D., Sadjadi, A. R., Aghcheli, K., et al. (2011). Mutations in Fanconi anemia genes and the risk of esophageal cancer. *Human Genetics* 129, 573–582. doi: 10.1007/s00439-011-0951-7

Alter, B. P. (2003). Cancer in Fanconi anemia, 1927–2001. *Cancer* 97, 425–440. doi: 10.1002/cncr.11046

Amano, Y., Raz, A., Timmermans, S., and Shkedi-Rafid, S. (2022). Cancer patients' understandings of genetic variants of uncertain significance in clinical care. *J Community Genetics* 13, 381–388. doi: 10.1007/s12687-022-00594-z

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., et al. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 29, 37–40.

Bartha, Á., and Győrffy, B. (2019). Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers (Basel)* 11, 1725. doi: 10.3390/cancers11111725

Behera, S., Catreux, S., Rossi, M., Truong, S., Huang, Z., Ruehle, M., et al. (2024). Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*, 1–15. doi: 10.1038/s41587-024-02382-1

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., et al. (2024). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* 625, 92–100. doi: 10.1038/s41586-023-06045-0

Chen, T., Cheng, H., Chen, X., Yuan, Z., Yang, X., Zhuang, M., et al. (2015). Family history of esophageal cancer increases the risk of esophageal squamous cell carcinoma. *Sci Rep* 5, 16038. doi: 10.1038/srep16038

Chen, W. C., Kerr, R., May, A., Ndlovu, B., Sobalisa, A., Duze, S. T., et al. (2018). The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term

Storage at −30°C. *Biopreservation and Biobanking* 16, 106–113. doi: 10.1089/bio.2017.0050

Chen, W. C., Singh, E., Muchengeti, M., Bradshaw, D., Mathew, C. G., Babb de Villiers, C., et al. (2020). Johannesburg Cancer Study (JCS): contribution to knowledge and opportunities arising from 20 years of data collection in an African setting. *Cancer Epidemiology* 65, 101701. doi: 10.1016/j.canep.2020.101701

Chen, Y., Liu, L., Li, J., Jiang, H., Ding, C., and Zhou, Z. (2023). MetaLR: Meta-tuning of Learning Rates for Transfer Learning in Medical Imaging. doi: 10.48550/arXiv.2206.01408

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., and Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE* 7, e46688. doi: 10.1371/journal.pone.0046688

Deng, J., Weng, X., Ye, J., Zhou, D., Liu, Y., and Zhao, K. (2019). Identification of the Germline Mutation Profile in Esophageal Squamous Cell Carcinoma by Whole Exome Sequencing. *Frontier in Genetics* 10, 47. doi: 10.3389/fgene.2019.00047

Drost, H.-G., and Tjeldnes, H. (2015). biomartr: Genomic Data Retrieval. 1.0.7. doi: 10.32614/CRAN.package.biomartr

Goodwin, P. J., Phillips, K.-A., West, D. W., Ennis, M., Hopper, J. L., John, E. M., et al. (2012). Breast Cancer Prognosis in BRCA1 and BRCA2 Mutation Carriers: An International Prospective Breast Cancer Family Registry Population-Based Cohort Study. *JCO* 30, 19–26. doi: 10.1200/JCO.2010.33.0068

Hopkins, J. J., Wakeling, M. N., Johnson, M. B., Flanagan, S. E., and Laver, T. W. (2023). REVEL Is Better at Predicting Pathogenicity of Loss-of-Function than Gain-of-Function Variants. *Human Mutation* 2023, 8857940. doi: 10.1155/2023/8857940

Huang, F.-L., and Yu, S.-J. (2018). Esophageal cancer: Risk factors, genetic association, and treatment. *Asian Journal of Surgery* 41, 210–215. doi: 10.1016/j.asjsur.2016.10.005

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002). The Ensembl genome database project. *Nucleic Acids Res* 30, 38–41.

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *American Journal of Human Genetics* 99, 877. doi: 10.1016/j.ajhg.2016.08.016

Jiang, Y.-X., Zhang, D.-W., Chen, Y., Sun, H.-H., Xu, S.-C., and Gao, H.-J. (2015). The characteristics of oesophageal squamous cell carcinoma: an analysis of 1317 cases in southeastern China. *Contemp Oncol (Pozn)* 19, 137–141. doi: 10.5114/wo.2015.51419

Joynt, A. C. M., Axford, M. M., Chad, L., and Costain, G. (2021). Understanding genetic variants of uncertain significance. *Paediatrics & Child Health* 27, 10. doi: 10.1093/pch/pxab070

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., et al. (2003). The UCSC Genome Browser Database. *Nucleic Acids Research* 31, 51–54.

Kim, S., Jhong, J.-H., Lee, J., and Koo, J.-Y. (2017). Meta-analytic support vector machine for integrating multiple omics data. *BioData Min* 10, 2. doi: 10.1186/s13040-017-0126-8

Landrum, M., Lee, J., Riley, G., Jang, W., Rubinstein, W., Church, D., et al. (2013). "ClinVar," in *The NCBI Handbook [Internet]. 2nd edition*, (National Center for Biotechnology Information (US)). Available at: https://www.ncbi.nlm.nih.gov/books/NBK174587/ (Accessed November 26, 2024).

Li, Q., and Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* 100, 267–280. doi: 10.1016/j.ajhg.2017.01.004

Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* 12, 103. doi: 10.1186/s13073-020-00803-9

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45, 580–585. doi: 10.1038/ng.2653

Lucci-Cordisco, E., Amenta, S., Panfili, A., del Valle, J., Capellá, G., Pineda, M., et al. (2022). Variants of uncertain significance (VUS) in cancer predisposing genes: What are we learning from multigene panels? *European Journal of Medical Genetics* 65, 104400. doi: 10.1016/j.ejmg.2021.104400

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature* 508, 469–476. doi: 10.1038/nature13127

McDonald, J. T., and Ricks-Santi, L. J. (2022). Hereditary variants of unknown significance in African American women with breast cancer. *PLOS ONE* 17, e0273835. doi: 10.1371/journal.pone.0273835

Mersch, J., Brown, N., Pirzadeh-Miller, S., Mundt, E., Cox, H. C., Brown, K., et al. (2018). Prevalence of Variant Reclassification Following Hereditary Cancer Genetic Testing. *Journal of the American Medical Association* 320, 1266–1274. doi: 10.1001/jama.2018.13152

Mirchandani, K. D., and D'Andrea, A. D. (2006). The Fanconi anemia/BRCA pathway: A coordinator of cross-link repair. *Experimental Cell Research* 312, 2647–2653. doi: 10.1016/j.yexcr.2006.06.014

Molina-Ramírez, L. P., Kyle, C., Ellingford, J. M., Wright, R., Taylor, A., Bhaskar, S. S., et al. (2022). Personalised virtual gene panels reduce interpretation workload and maintain diagnostic rates of proband-only clinical exome sequencing for rare disorders. *Journal Medical Genetics* 59, 393–398. doi: 10.1136/jmedgenet-2020-107303

Morgan, N. V., Essop, F., Demuth, I., de Ravel, T., Jansen, S., Tischkowitz, M., et al. (2005). A common Fanconi anemia mutation in black populations of sub-Saharan Africa. *Blood* 105, 3542–3544. doi: 10.1182/blood-2004-10-3968

Murphy, G., McCormack, V., Abedi-Ardekani, B., Arnold, M., Camargo, M. C., Dar, N. A., et al. (2017). International cancer seminars: a focus on esophageal squamous cell carcinoma. *Annals of Oncologyl* 28, 2086–2093. doi: 10.1093/annonc/mdx279

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31, 3812–3814.

Nissen, S. K., Christiansen, M., Helleberg, M., Kjær, K., Jørgensen, S., Gerstoft, J., et al. (2018). Whole Exome Sequencing of HIV-1 long-term non-progressors identifies rare variants in genes encoding innate immune sensors and signaling molecules. *Scientific Reports* 8. doi: 10.1038/s41598-018-33481-0

Osler, T. S., Brandenburg, J.-T., Schoeman, M., Chen, W. C., Urban, M. F., and Mathew, C. G. (2024). Prevalence and Reclassification of Genetic Variants in South African Populations with Breast Cancer. *Genes, Chromosomes and Cancer* 63, e23275. doi: 10.1002/gcc.23275

Pruitt, K., Brown, G., Tatusova, T., and Maglott, D. (2012). "The Reference Sequence (RefSeq) Database," in *The NCBI Handbook [Internet]*, (National Center for Biotechnology Information (US)). Available at: https://www.ncbi.nlm.nih.gov/books/NBK21091/ (Accessed November 26, 2024).

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* 17, 405–424. doi: 10.1038/gim.2015.30

Samadder, N. J., Riegert-Johnson, D., Boardman, L., Rhodes, D., Wick, M., Okuno, S., et al. (2021). Comparison of Universal Genetic Testing vs Guideline-Directed Targeted Testing for Patients With Hereditary Cancer Syndrome. *Journal of the American Medical Association Oncology* 7, 230–237. doi: 10.1001/jamaoncol.2020.6252

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., et al. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536–1543. doi: 10.1093/bioinformatics/btv009

Simba, H., Kuivaniemi, H., Lutje, V., Tromp, G., and Sewram, V. (2019). Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in Genetics* 10, 642. doi: 10.3389/fgene.2019.00642

Strickland, N. J., Matsha, T., Erasmus, R. T., and Zaahl, M. G. (2012). Molecular analysis of Ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International Journal of Cancer* 131, 623–632. doi: 10.1002/ijc.26418

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., 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, 209–249. doi: 10.3322/caac.21660

Torgovnick, A., and Schumacher, B. (2015). DNA repair mechanisms in cancer development and therapy. *Frontier in Genetics* 6, 157. doi: 10.3389/fgene.2015.00157

Vogelsang, M., Wang, Y., Veber, N., Mwapagha, L. M., and Parker, M. I. (2012). The Cumulative Effects of Polymorphisms in the DNA Mismatch Repair Genes and Tobacco Smoking in Oesophageal Cancer Risk. *PLoS One* 7, e36962. doi: 10.1371/journal.pone.0036962

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38, e164. doi: 10.1093/nar/gkq603

Woods, N. T., Baskin, R., Golubeva, V., Jhuraney, A., De-Gregoriis, G., Vaclova, T., et al. (2016). Functional assays provide a robust tool for the clinical annotation of genetic variants of uncertain significance. *npj Genomic Medicine* 1, 1–9. doi: 10.1038/npjgenmed.2016.1

Zeng, B., Huang, P., Du, P., Sun, X., Huang, X., Fang, X., et al. (2021). Comprehensive Study of Germline Mutations and Double-Hit Events in Esophageal Squamous Cell Cancer. *Frontier in Oncology* 11, 637431. doi: 10.3389/fonc.2021.637431

Zhang, X.-M., and Guo, M.-Z. (2010). The value of epigenetic markers in esophageal cancer. *Frontier in Medicine. China* 4, 378–384. doi: 10.1007/s11684-010-0230-3

Online references:

Cancer Statistics (n.d.). *NICD*. Available at: https://www.nicd.ac.za/centres/national-cancer-registry/cancer-statistics/ (Accessed April 11, 2024).

Invitae Multi-Cancer Panel | Test catalog | Invitae (n.d.). Available at: https://www.invitae.com/us/providers/test-catalog/test-01101 (Accessed November 19, 2024).

PMS1 PMS1 homolog 1, mismatch repair system component [Homo sapiens (human)] - Gene - NCBI (n.d.). Available at: https://www.ncbi.nlm.nih.gov/gene/5378 (Accessed November 25, 2024).

PubChem (n.d.). MSH3 - mutS homolog 3 (human). Available at: https://pubchem.ncbi.nlm.nih.gov/gene/MSH3/human (Accessed November 19, 2024).

Quick Start-Up Guide - ANNOVAR Documentation (n.d.). Available at: https://annovar.openbioinformatics.org/en/latest/user-guide/startup/ (Accessed November 26, 2024).

Sefid Dashti, M. J., and Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. *BioTechniques* 62, 18–30. doi: 10.2144/000114492

Appendices

Appendix A: Supplementary Tables and Figures

Appendix B: Ethics Clearance Certificate

Appendix C: Approved Research Protocol

Appendix D: Turnitin Report and Plagiarism Declaration

Appendix E: Author Guidelines for *Frontiers in Genetics*

Appendix A: Supplementary Tables and Figures

Supplementary Figure 1. Script for data preprocessing in RStudio

get_gene_info.R

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("biomaRt")

library(biomaRt)
ensembl = useEnsembl(biomart="ensembl")
listDatasets(ensembl)
mart <- useDataset("hsapiens_gene_ensembl", mart=ensembl)

genes <- c("AIP", "ALK", "APC", "ATM", "AXIN2", "BAP1", "BARD1", "BLM", "BMPR1A", "CDC73", "CDH1", "CDK4", "CDKN1B",
"CDKN2A", "CTNNA1", "DICER1", "EGFR", "EPCAM", "FH", "FLCN", "GREM1", "HOXB13", "KIT", "LZTR1", "MAX", "MBD4", "MEN1",
"MET", "MITF", "MLH1", "MSH2", "MSH6", "NF1", "NF2", "NTHL1", "PDGRRA", "POLD1", "POLE", "POT1", "PRKAR1A", "PTCH1",
"PTEN", "RAD51D", "RB1", "RET", "SDHA", "SDHAF2", "SDHB", "SDHC", "SDHD", "SMAD4", "SMARCA4", "SMARCB1", "SMARCE1",
"STK11", "SUFU", "TMEM127", "TSC1", "TSC2", "VHL", "FANCA", "FANCB", "FANCC", "BRCA2", "FANCD2", "FANCE", "FANCF",
"FANCG", "FANCI", "BRIP1", "FANCL", "FANCM", "PALB2", "RAD51C", "SLX4", "ERCC4", "RAD51", "BRCA1", "UBE2T", "MAD2L2",
"RFWD3", "GSTT2B", "ADH1B", "CP", "ADH3", "AR", "CYP3A5", "PTGS2", "SLC11A", "NAT2", "PMS11", "MLH3", "ALDH2", "PMS",
"CASP8", "PMS1", "CYP2E1", "MGMT", "MSH3", "RUNX1", "FAM120A", "MYO1B", "PLCE1", "CHEK2", "TP53", "GJB2", "RECQL4",
"MUTYH", "PMS2")

attributes <- c('chromosome_name', 'start_position', 'end_position', 'strand', 'hgnc_symbol')

gene_info <- getBM(attributes=attributes,
                   filters='hgnc_symbol',
                   values=genes,
                   mart=mart)

print(gene_info)

setwd("E:")

write.csv(gene_info, "oscc_wes_gene_info.csv", row.names = FALSE)

```

Supplementary Table 1. 72 variants remaining prior to pathogenicity prediction and classification.

Chr	Ref	Alt	Gene name	Exonic function	aa Change	SNP ID	ClinVar Sig	MAF
chr1	G	A	*MUTYH*	nonsynonymous SNV	p.R98W	rs587782041	VUS	0.00003
chr1	A	G	*FH*	nonsynonymous SNV	p.Y2H	rs112335468	Conflicting interpretations	0.00080
chr2	C	G	*ALK*	nonsynonymous SNV	p.E859D	rs61754865	Conflicting interpretations	0.00410
chr2	G	C	*EPCAM*	nonsynonymous SNV	p.D241H	.	.	0.00006
chr2	C	T	*MSH6*	nonsynonymous SNV	p.P42S	rs34014629	Conflicting interpretations	0.00210
chr2	G	C	*PMS1*	nonsynonymous SNV	p.E27Q	rs5742973	VUS	0.00230
chr2	A	G	*PMS1*	nonsynonymous SNV	p.N608S	rs55680602	.	0.00060
chr2	A	C	*CASP8*	nonsynonymous SNV	p.K305T	.	VUS	0.00000
chr2	G	T	*FLACC1*	nonsynonymous SNV	p.A198E	rs142608789	.	0.00150
chr2	G	A	*BARD1*	nonsynonymous SNV	p.P262L	rs367890377	Conflicting interpretations	0.00070
chr5	TT	-	*SDHA*	frameshift deletion	p.L568Efs*4	rs112307877	VUS	0.00150
chr5	CTCC	-	*MSH3*	frameshift deletion	p.A70Sfs*9	rs766700228	.	0.00050
chr5	G	C	*MSH3*	nonsynonymous SNV	p.A70P	rs765698219	VUS	0.00007
chr5	CTTCCCG	-	*MSH3*	frameshift deletion	p.F71Pfs*7	rs759756736	.	0.00010
chr5	C	T	*MSH3*	nonsynonymous SNV	p.A438V	rs35121792	Conflicting interpretations	0.00500
chr5	T	G	*APC*	nonsynonymous SNV	p.D2288E	.	.	0.00000
chr7	C	T	*PMS2*	nonsynonymous SNV	p.D678N	rs143340522	Conflicting interpretations	0.00700
chr7	C	G	*PMS2*	nonsynonymous SNV	p.V632L	rs758225108	VUS	0.00060
chr7	A	G	*PMS2*	nonsynonymous SNV	p.Y413H	rs370236216	Conflicting interpretations	0.00030
chr7	C	G	*PMS2*	nonsynonymous SNV	p.E202D	rs114185660	VUS	0.00040
chr7	G	T	*PMS2*	nonsynonymous SNV	p.T171K	rs1805322	Conflicting interpretations	0.00410
chr7	G	A	*PMS2*	nonsynonymous SNV	p.S22L	rs116373169	Conflicting interpretations	0.00490
chr7	G	A	*CYP3A5*	nonsynonymous SNV	p.T110M	rs371669684	.	0.00006
chr7	A	T	*MET*	nonsynonymous SNV	p.M35L	rs375353223	Conflicting interpretations	0.00090
chr8	T	A	*NAT2*	nonsynonymous SNV	p.L24I	rs45477599	.	0.00180

chr8	C	G	*RECQL4*	nonsynonymous SNV	p.G1113R	rs35101495	Conflicting interpretations	0.00900
chr9	G	T	*FAM120A*	nonsynonymous SNV	p.A330S	rs144744529	.	0.00030
chr9	G	A	*FANCC*	nonsynonymous SNV	p.R535C	rs185822330	VUS	0.00020
chr9	C	G	*FANCC*	nonsynonymous SNV	p.A238P	rs553326913	.	0.00006
chr9	G	T	*FANCC*	nonsynonymous SNV	p.P78H	rs138722298	VUS	0.00020
chr9	C	T	*PTCH1*	nonsynonymous SNV	p.G25E	rs774712511	VUS	0.00000
chr10	G	C	*PTEN*	nonsynonymous SNV	p.R30S	.	.	0.00030
chr10	T	G	*PTEN*	nonsynonymous SNV	p.S294R	rs143335584	VUS	0.00100
chr10	G	A	*PLCE1*	nonsynonymous SNV	p.V23M	rs144179807	.	0.00320
chr10	G	A	*PLCE1*	nonsynonymous SNV	p.G891S	rs199781223	Conflicting interpretations	0.00460
chr11	A	G	*TMEM134*	nonsynonymous SNV	p.Y57H	rs776596331	.	0.00160
chr11	A	G	*ATM*	nonsynonymous SNV	p.D44G	rs150143957	VUS	0.00080
chr12	C	T	*MARCHF9*	nonsynonymous SNV	p.T107I	rs774466216	.	0.00080
chr12	G	A	*POLE*	nonsynonymous SNV	p.R1839C	rs141519273	VUS	0.00000
chr12	C	A	*POLE*	nonsynonymous SNV	p.A341S	rs137860861	Conflicting interpretations	0.00100
chr13	A	G	*BRCA2*	nonsynonymous SNV	p.D156G	rs68071147	Conflicting interpretations	0.00050
chr13	A	G	*BRCA2*	nonsynonymous SNV	p.N1501S,	rs545444016	Conflicting interpretations	0.00006
chr14	C	G	*FANCM*	nonsynonymous SNV	p.P266A	rs142747831	VUS	0.00170
chr14	C	T	*FANCM*	nonsynonymous SNV	p.R628W	rs112784867	VUS	0.00040
chr14	G	A	*FANCM*	nonsynonymous SNV	p.V960I	rs112345761	VUS	0.00000
chr14	GAA	-	*FANCM*	no frameshift deletion	p.E1596del	rs765421461	Conflicting interpretations	0.00380
chr14	A	T	*DICER1*	nonsynonymous SNV	p.S1618T	.	VUS	0.00000
chr15	C	T	*GREM1*	nonsynonymous SNV	p.H41Y	.	.	0.00010
chr15	G	A	*RAD51*	nonsynonymous SNV	p.R151Q+H13	rs121917739	Conflicting interpretations	0.00520
chr15	G	A	*FANCI*	nonsynonymous SNV	p.C156Y	rs112387610	VUS	0.00050
chr15	A	C	*FANCI*	nonsynonymous SNV	p.T681P	rs554876309	VUS	0.00070
chr15	A	G	*FANCI*	nonsynonymous SNV	p.N757S	rs111883566	VUS	0.00060
chr15	C	T	*FANCI*	nonsynonymous SNV	p.T927M	rs138432305	VUS	0.00020
chr15	G	A	*FANCI*	nonsynonymous SNV	p.E1243K	rs544848412	VUS	0.00009
chr15	G	A	*FANCI*	nonsynonymous SNV	p.G1256E	rs138461165	Conflicting interpretations	0.00140

chr15	T	A	*BLM*	nonsynonymous SNV	p.V1205E	.	.	0.00003
chr16	G	A	*NTHL1*	nonsynonymous SNV	p.A38V	rs202082304	Conflicting interpretations	0.00980
chr16	G	A	*SLX4*	nonsynonymous SNV	p.R1226W,	rs142008398	Conflicting interpretations	0.00420
chr16	C	T	*SLX4*	nonsynonymous SNV	p.S393N	rs112301713	VUS	0.00030
chr16	A	T	*ERCC4*	nonsynonymous SNV	p.Q496H	rs146601373	Conflicting interpretations	0.00960
chr16	C	A	*ERCC4*	nonsynonymous SNV	p.A860D	rs4986933	Conflicting interpretations	0.00920
chr16	G	T	*PALB2*	nonsynonymous SNV	p.P117T	.	VUS	0.00000
chr16	G	A	*PALB2*	nonsynonymous SNV	p.P8L	rs150390726	Conflicting interpretations	0.00150
chr16	C	G	*CDH1*	nonsynonymous SNV	p.A394G	rs532576241	VUS	0.00009
chr16	G	A	*FANCA*	nonsynonymous SNV	p.S1087L	rs376089640	VUS	0.00006
chr16	C	T	*FANCA*	nonsynonymous SNV	p.E420K	rs760352719	.	0.00000
chr16	T	C	*FANCA*	nonsynonymous SNV	p.Q254R	rs13336566	Conflicting interpretations	0.00890
chr17	C	T	*NF1*	nonsynonymous SNV	p.T1627I	rs376655102	Conflicting interpretations	0.00030
chr17	C	T	*RAD51D*	nonsynonymous SNV	p.A78T	rs80116829	Conflicting interpretations	0.00530
chr17	C	T	*AXIN2*	nonsynonymous SNV	p.D745N	rs140344858	Conflicting interpretations	0.00540
chr21	G	A	*RUNX1*	nonsynonymous SNV	p.P251L	.	VUS	0.00006
chrX	C	T	*AR*	nonsynonymous SNV	p.P648S,	.	.	0.00030

Appendix B: Ethics Clearance Certificate

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

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Ms Z Dangor
School of Pathology
Division of Human Genetics
Faculty of Health Sciences

E-mail: 2208826@students.wits.ac.za

CC: Supervisor: Drs WC Chen & J-T Brandenburg; Prof C Mathew
Carl.Chen@wits.ac.za
and <HREC-Medical Research Office@wits.ac.za>

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

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

DATE: 29 August 2024

REF: R14/49

PROTOCOL NO: **M240769** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Germline mutation landscape of oesophageal squamous cell cancer in South African patients*

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

MSWorks2000/Iain0007/Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Ms Z Dangor

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M240769

NAME: (Principal Investigator)	Ms Z Dangor **DEGREE:** MSc (Med)
DEPARTMENT:	School of Pathology Division of Human Genetics Faculty of Health Sciences
PROJECT TITLE:	*Germline mutation landscape of oesophageal squamous cell cancer in South African patients*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M21/11/154 and M22/08/81
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Drs WC Chen and J-T Brandenburg; Professor C Mathew
APPROVED BY:	Professor P Ruff, Chairperson, HREC (Medical)
DATE OF APPROVAL:	2024/08/29 **EXPIRY DATE:** 2029/08/28

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

DECLARATION OF INVESTIGATORS

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

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report** in the format available at https://wits.ac.za/research/researcher-support/research-ethics/ethics-committees/. This report is due annually, for the duration of the Clearance Certificate, beginning on the first anniversary of Date of Approval above. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

Date

Appendix C: Approved Research Protocol

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Private Bag 3 Wits, 2050
Fax: 027117172119
Tel: 02711 7172076

Reference: Mrs Sandra Benn
E-mail: sandra.benn@wits.ac.za

27 June 2024
Person No: 2208826
PAG

Ms Z Dangor
8 Katjiepiering Street
Weltevredenpark Ext 5
1709
South Africa

Dear Ms Zahraa Dangor

Master of Science in Medicine: Approval of Title

We have pleasure in advising that your proposal entitled *Germline mutation landscape of oesophageal squamous cell cancer in South African patients* has been approved. Please note that any amendments to this title have to be endorsed by the Faculty's higher degrees committee and formally approved.

Yours sincerely

Mrs Sandra Benn
Faculty Registrar
Faculty of Health Sciences

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

CANDIDATE'S SURNAME: Dangor	FIRST NAME/S: Zahraa	STUDENT NUMBER: 2208826
CURRENT QUALIFICATIONS: BSc Human physiology, Genetics and Psychology (UP) BHSc (Hons) Human Genetics (Wits)		

TEL: 0114752344	CELL:0605333802	E-MAIL: 2208826@students.wits.ac.za	FAX: N/A

DEGREE FOR WHICH PROTOCOL IS BEING SUBMITTED:		
PART-TIME OR FULL-TIME: Full-time		
FIRST REGISTERED FOR THIS DEGREE:	TERM: One	YEAR: 2024
DEPARTMENT: Human Genetics		

- TITLE OF PROPOSED RESEARCH: Germline mutation landscape of oesophageal squamous cell cancer in South African patients

CANDIDATE'S SIGNATURE:	DATE:2024/04/11
SUPERVISOR 1 (NAME & SURNAME): Dr Wenlong Carl Chen	50% Supervision
SUPERVISOR'S QUALIFICATIONS: PhD	
SUPERVISOR'S DEPARTMENT: SBIMB, Wits	
SUPERVISOR'S ADDRESS / TEL / E-MAIL: carl.chen@wits.ac.za	
SUPERVISOR 2 (NAME & SURNAME): Dr Jean-Tristan Brandenburg	25% Supervision
SUPERVISOR'S QUALIFICATIONS: PhD	
SUPERVISOR'S ADDRESS / TEL / E-MAIL: jean-tristan.brandenburg@wits.ac.za	
SUPERVISOR 3 (NAME & SURNAME): Prof Christopher Mathew	25% Supervision
SUPERVISOR'S QUALIFICATIONS: PhD	
SUPERVISOR'S ADDRESS / TEL / E-MAIL: Sydney Brenner Institute for Molecular Bioscience, 9 Jubilee Road, Parktown 2193. Tel. 011 7176631, Email: christopher.mathew@kcl.ac.uk; chris.g.mathew@wits.ac.za	

SYNOPSIS OF RESEARCH:*(Brief summary of proposed research project; between 200-300 words only; with sub-headings: an introduction and justification for study, aim/s, proposed methodology and expected outcome/s)*

Introduction

Oesophageal Cancer is the seventh most prevalent cancer worldwide and ranks sixth in cancer mortality. Oesophageal cancer is characterised by two major histologic subtypes with distinct epidemiological and clinical features: oesophageal adenocarcinoma (OAC) and oesophageal squamous cell cancer (OSCC). Globally, OSCC is the more common type, contributing to approximately 90% of cases.

Study rationale.

The genetic aetiology of OSCC in Africa has been primarily investigated through multiple case-control association studies identifying SNPs in candidate genes and a single GWAS. These studies identify the common variants associated with OSCC. However, these studies are unable to identify functional causality variants that drive disease aetiology. Currently, there are no germline gene sequencing studies for OSCC conducted in Africa. This highlights the need for germline mutation analysis as the appropriate next step in African OSCC research.

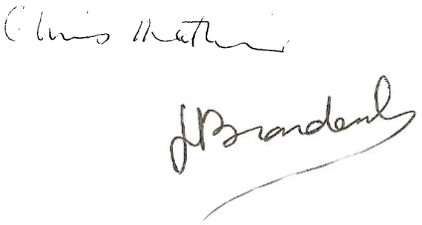
Aim

The aim of this study is to analyse the germline mutation landscape in the known cancer predisposition genes in Black African patients with OSCC.

WITS ETHICS NOT REQUIRED:	Yes	No
WITS ETHICS PENDING:	(Yes)	No
WITS ETHICS APPROVED:	Yes	No
(circle appropriate symbol)*		

***Please note the final human ethics clearance certificate or animal ethics certificate must be available prior to starting research**

IF Y SUPPLY ETHICS CLEARANCE CERTIFICATE AS ATTACHMENT AND INCLUDE ETHICS NUMBER HERE:

As supervisor/s, I/we confirm that I have read the protocol which has been submitted for assessment.		
SIGNATURE OF SUPERVISOR/S:		
SIGNATURE PG OFFICE STAFF …… …..	REGISTERED YES….. NO…..	STAMP

SYNOPSIS OF RESEARCH CONTINUED

Proposed methodology

The methodology of this study involves the construction of a virtual gene panel of known cancer predisposition genes as well as the bioinformatic analysis of Whole Exome Sequencing data obtained from 16 Black South African patients with OSCC. This workflow will be divided into variant annotation, variant classification, and variant prioritisation. A list of pathogenic variants will be generated for each sample file.

Expected outcome

Pathogenic variants are present in some of the known cancer predisposition genes in Black South African patients with OSCC.

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Germline mutation landscape of oesophageal squamous cell cancer in South African patients

Research proposal.

12/04/2024

To the faculty of Health Science, University of the Witwatersrand, in partial fulfilment of the requirements for the Degree of

MSc (Med) Genomic Medicine

By

Zahraa Dangor (2208826)

Supervisor 1: Dr Wenlong Carl Chen
Supervisor 2: Dr Jean-Tristan Brandenburg
Supervisor 3: Prof Christopher Mathew

NATIONAL HEALTH
LABORATORY SERVICE

SDIMB
SYDNEY BRENNER INSTITUTE
FOR MOLECULAR BIOSCIENCE

Table of Contents

1. Introduction 1

1.1 Oesophageal squamous cell cancer 1

1.2 Oesophageal squamous cell cancer in Africa 2

1.3 Disease aetiology 2

1.3.1 Lifestyle and Environmental Risk Factors 2

1.3.2 Genetic Risk Factors 3

1.4 Germline mutation analysis using Germline Whole Exome Sequencing 5

2. Aim 6

3. Objectives 6

1. Methods 7

4.1 Study participants. 7

4.2 Whole exome sequencing 7

4.3 Compiling a list of cancer predisposition genes 8

4.3 Data Analysis 9

4 Ethics 11

5 Timing 12

6 Funding 12

7 Limitations 13

8 References 13

List of abbreviations

ANNOVAR	ANNOtate VARiation
ACMG/AMP	American College of Medical Genetics and Genomics and Association for Molecular Pathology
CI	Confidence Interval
InterVar	Interpretation of variants
JCS	Johannesburg Cancer Study
GWAS	Genome-wide association studies
OAC	Oesophageal Adenocarcinoma
OC	Oesophageal Cancer
OR	Odds ratio
OSCC	Oesophageal squamous cell carcinoma
RR	Relative risk
SNP	Single nucleotide polymorphisms
VCF	Variant Call Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

List of figures

Figure one: Stepwise filtering approach to prioritize variants and generate a shortlist of variants……...10

1. Introduction

1.1 Oesophageal squamous cell cancer

Oesophageal Cancer (OC) is the seventh most prevalent cancer worldwide, with over 600,000 new cases diagnosed each year, and is responsible for more than 540,000 deaths annually, ranking sixth in cancer mortality (Sung et al., 2021). Oesophageal cancer is characterised by two major histologic subtypes with distinct epidemiological and clinical features: oesophageal adenocarcinoma (OAC) and oesophageal squamous cell cancer (OSCC). Oesophageal adenocarcinoma arises through the uncontrolled, abnormal proliferation of glandular tissue, whereas OSCC arises through the rapid proliferation and dysplasia of squamous epithelial cells (Alema & Iva, 2014). Globally, OSCC is the more common type, contributing to approximately 90% of OC cases (Sung et al., 2021).

OSCC is characterised by the growth of malignant tumours in the squamous epithelium that lines the oesophagus (Strickland et al., 2012). The histologic progression of OSCC involves the transition from normal squamous epithelium to hyperproliferative epithelium, low, intermediate and high-grade dysplasia followed by metastasis (Zhang & Guo, 2010). Non-specific clinical features often develop during advanced stages of the disease, with dysphagia being the most frequently reported symptom. Other symptoms include hoarseness, epigastric pain, heartburn and chest pain (Jiang et al., 2015). The largely asymptomatic disease progression and late clinical presentation result in late-stage diagnosis and poor prognosis.

1.2 Oesophageal squamous cell cancer in Africa

OSCC is endemic in specific regions and exhibits the highest global divergence in cancer prevalence, with high incidence reported in China, as well as southern and eastern Africa (Abnet et al., 2018). Oesophageal cancer hotspots in Africa have been termed the African OSCC corridor which includes Ethiopia, Kenya, Malawi, Tanzania, and South Africa. According to GLOBOCAN 2020, Southern Africa has the second-highest incidence rate worldwide (Sung et al., 2021). Furthermore, the South African National Cancer Registry reported that 720 females and 942 males were histologically diagnosed with OC in 2020 alone, of which 568 and 663 were black, respectively (NCR Cancer Statistics, 2023). This highlights the high prevalence of this cancer in the Black South African population.

1.3 Disease aetiology

The aetiology of OSCC is multifactorial and marked by striking heterogeneity across populations worldwide (Murphy et al., 2017). The risk factors associated with this cancer include several environmental, lifestyle and genetic factors (Huang & Yu, 2018). An improved understanding of the causes of this cancer in Africa, particularly the genetic causes, is of great importance.

1.3.1 Lifestyle and Environmental Risk Factors

Lifestyle and environmental risk factors postulated to be linked to OSCC in Africa include dietary deficiencies, tobacco smoking, domestic smoke exposure from wood fires, alcohol consumption, poor oral hygiene and drinking hot beverages (Munishi et al., 2015; Simba et al., 2023). Several case-control studies have been performed to elucidate the lifestyle risk factors for this cancer in Black Africans. (Simba et al., 2023) systematically assessed all environmental and lifestyle risk factors reported in relevant African literature and tested for an association with the

progression of OSCC. A meta-analysis was performed on studies that reported on odds ratios (OR) or relative risk (RR) and 95% confidence intervals (CI). This meta-analysis generated a pooled analysis for each of the seven risk factors' effect sizes, as shown in **Table 1**.

Table 1. Pooled analysis of lifestyle and environmental risk factors associated with OSCC in Africa from (Simba et al., 2023).

Risk factor	**Odds-ratio**	**95% CI**
Tobacco use	3.01	2.37-3.83
Alcohol consumption	1.79	1.35-2.37
Combined alcohol and tobacco use	3.85	2.49-5.93
Hot food and beverage exposure	1.68	1.13-2.49
Poor oral hygiene	3.52	2.64-4.69
PAH exposure	2.08	1.27-3.42
Fruit and vegetable consumption	0.51	0.37-0.71

1.3.2 Genetic risk factors

To date, the genetic aetiology of OSCC in Africa has been primarily investigated through genetic association studies in the form of candidate gene association studies and genome-wide association studies (GWAS) (Simba et al., 2019). These studies aim to identify the common variants associated with this disease for an improved understanding of its genetic aetiology.

Simba et al. (2019) performed a comprehensive review of existing evidence of variants associated with OSCC in Africa. The primary aim of this was to bridge the knowledge gaps in the OSCC genetic aetiology and outline the role of genetic risk factors in the molecular pathogenesis of this cancer in African populations. Twenty-three genetic studies were included in their review, all of which were carried out on African populations between 1990 and 2019. These genetic studies comprised 17 genetic susceptibility studies, all of which were case-control studies which collectively studied 37 genes (Bye et al., 2011, 2012; Chelule et al., 2006; Chen et al., 2019;

Dandara et al., 2005, 2006; Dietzsch et al., 2003; Eltahir et al., 2012; D. Li et al., 2005, 2010; Li et al., 2008; Matejcic et al., 2011, 2015; Strickland et al., 2012; Vogelsang et al., 2012; Vos et al., 2003; Zaahl et al., 2005). The sample sizes of these studies ranged between 18-880 patients, with a mere six studies having over 200 samples. Findings included 25 statistically significant variants in 20 genes (*GSTT2B, ADH1B, CP, ALDH2, CASP8, ADH3, CHEK2, CYP2E1, AR, CYP3A5, MGMT, MLH3, MSH3, PTGS2*, *SLC11A, RUNX1, PLCE1, NAT2, PMS11*, and *TP53*). However, only the findings from 16 SNPs were potentially reliable, given the small sample sizes. Of the 16 SNPs, 10 were associated with an increased risk in the genes *MLH3, ALDH2* (ALDH2*1/*2), *PMS* (c.-21+639G > A; rs5742938), *CHEK2* (rs4822983 C > T, and rs1033667, C > T,) *CASP8* (Asp302His; rs1045485)*, PMS1, CYP2E1* (7632T > A), *MGMT* (Leu84Phe; rs12917*)*, *MSH3* (Ala1045Thr; rs26279) and *RUNX1* (rs2014300).

Chen et al. (2023) conducted the first African OSCC GWAS to further investigate genetic risk loci associated with OSCC. This study was conducted in the Black South African population, comprising 1,686 cases and 3,217 controls. Samples were genotyped using an H3Africa array, an array designed to enrich African genomic diversity. The findings of this study included one novel risk locus located on chromosome 9, upstream of the *FAM120A* gene (rs12379660, P = 4.58x10-8, OR=1.28, 95% CI=1.22-1.34). One African-specific risk locus was discovered on chromosome 2 (rs142741123, P = 5.49x10-8, OR= 2.66, 95% CI=2.43–2.89) within the *MYO1B* gene. Next, a meta-analysis was carried out using the African OSCC GWAS and a Chinese OSCC GWAS. This revealed three associated loci on chromosome 9 at *FAM120A*, chromosome 10 at *PLCE1* and chromosome 22 at *CHEK2*, all of which reached genome-wide significance. The findings of this study provide evidence and motivation for further investigation into the functional effects of these risk loci.

1.4 Germline mutation analysis using Germline Whole Exome Sequencing

Whole Exome Sequencing (WES) involves the analysis of protein-coding regions in the human genome. This technology enables the investigation of cancer-related genetic aberrations that are predominantly located in the exonic regions (Bartha & Győrffy, 2019). This makes WES a suitable and attractive technique for the identification of germline mutations in known cancer predisposition genes.

Little to no WES has been conducted for OSCC in Africa compared to other high-incident regions such as China (Chang et al., 2017). Larger research efforts have been made to identify the risk factors present in Chinese patients. (Zeng et al., 2021) employed WES to identify pathogenic germline variants in known cancer susceptibility genes. The rationale behind this study was that the profile of pathogenic germline variants in OSCC remains unclear, and the discovery of oncogenic alternations associated with OSCC is urgently needed. In order to conduct this study, the researchers utilised Whole Genome Sequencing (WGS) and WES data from 592 OSCC patients. A shortlist of pathogenic germline variants was generated following a bioinformatic analysis pipeline, including sequence alignment, variant calling variant annotation, pathogenicity evaluation, identification of potential double-hit events, and various statistical analyses. This study's findings include identifying 157 variants in cancer predisposition genes, accounting for 25.0% of the OSCC cases. More than five variants were identified in six of the genes included in the analysis: *TP53* (n = 15 mutations), *GJB2* (n = 8), *BRCA2* (n = 6), *RECQL4* (n = 6), *MUTYH* (n = 6), and *PMS2* (n = 5). Additionally, 84 double-hit events were identified in 18 tumour suppressor genes from 83 patients. Genes in the Fanconi anaemia pathway had the highest number of pathogenic variants compared to other pathways included (Zeng et al., 2021).

Overall, germline mutation analysis in OSCC improves the understanding of the underlying

mechanisms of OSCC and may provide information for the early detection, diagnosis, and risk management of this cancer (Deng et al., 2019). Similar studies are needed in Africa to improve or understanding of the germline mutational landscape in OSCC.

1.6 Study Rationale

As detailed, the genetic aetiology of OSCC in Africa has been primarily investigated through multiple case-control association studies identifying SNPs in candidate genes (Simba et al., 2019), and a single GWAS (Chen et al., 2023). These studies identify the common variants associated with OSCC. However, these studies are unable to identify functional causality variants that drive disease aetiology. Currently, there are no germline mutation studies for OSCC conducted in Africa. This highlights the need for germline mutation analysis as the appropriate next step in African OSCC research.

2. Aim

The aim of this study is to identify pathogenic variants in the known cancer predisposition genes of Black African patients with OSCC.

3. Objectives

1. To conduct an extensive literature review to compile a list of cancer predisposition genes associated with OSCC.
2. To perform variant annotation, classification and prioritisation in the genes of interest.
3. To identify pathogenic germline variants potentially associated with OSCC.

1. Methods

4.1 Study participants.

The study population for this project includes 16 Black South African patients with histologically confirmed OSCC. These patients have been selected from the Johannesburg Cancer Study (JCS), a large case-control study conducted from 1995 through 2016 (PI: Dr M Muchengeti & Dr WC Chen). In brief, the JCS recruited newly diagnosed cancer patients of Black African ancestry to examine whether cancer risk factors identified in Western countries applied to the Black African population in Johannesburg (Chen et al., 2020). Upon enrolment in the JCS, these 16 patients received a clinical and pathological diagnosis of OSCC between the ages of 26-38. The patients have been de-identified and assigned a unique sample ID for the purpose of this study.

Young patients were selected given the association between early age of cancer diagnosis and inherited cancer (Goodwin et al., 2012). An example of this association was demonstrated in a study conducted by (Goodwin et al., 2012) which compared the age of diagnosis of breast cancer in individuals who were heterozygous for germline mutations in the known breast cancer predisposition genes *BRCA1* and *BRCA2* with that in patients with sporadic disease. Findings included that patients carrying germline mutations in *BRCA1* and *BRCA2* had a younger age of diagnosis than those with sporadic disease (39.8 and 42.2 v 45.7 years). These germline mutation carriers were also more likely to have high-grade tumours. This study showed that individuals with germline mutations in known cancer predisposition genes tend to be diagnosed at an earlier age and often develop more aggressive forms of cancer.

4.2 Whole exome sequencing

A typical WES workflow is divided into sample collection, DNA extraction, library preparation,

exome capture, sequencing and data analysis (Bartha & Győrffy, 2019). Sample collection, DNA extraction, library preparation, exome capture and sequencing were performed prior to this study. In brief, sample collection and DNA extraction were performed for the JCS. Blood samples were drawn from patients, and germline DNA was isolated using the Qiagen Flexigene DNA kit and stored at -80C. The exonic regions of the genome were captured using the Agilent SureSelect Human All Exome Capture assay (version 8). Next, sequencing was performed on the Illumina NextSeq 2000 instrument. Sequencing quality control (QC), alignment and variant calling was performed using the Illumina DRAGEN-GATK software. The sequencing data will be provided for bioinformatic analysis in the form of raw Fastq files and Variant Call Format (VCF) files.

4.3 Compiling a list of cancer predisposition genes

In our study, germline WES data was generated from patients diagnosed with histologically confirmed OSCC. Although sequence data from the full complement of protein-coding genes will be available, in the limited scope of this study the exome dataset will be mined for a list of cancer predisposition genes associated with OSCC to reduce the time required and complexity of the analysis. The literature search that will be conducted to generate this list of genes forms part of the study's methodology. This list will include genes from the Invitae Multi Cancer panel, previously described genes known to be associated with OSCC, and genes that form part of the Fanconi anaemia (FA) pathway. The Invitae Multi Cancer panel, containing 70 genes, was selected as Invitae is an established clinical diagnostic laboratory that performs genetic testing and has done extensive research to develop these panels. Genes associated with OSCC will be selected following a literature search, including some of the genes identified by the genetic susceptibility studies summarised above. Lastly, genes that form part of the FA pathway will be included based on data that has linked variants in these genes with an increased predisposition to OSCC (Akbari et al.,

2011).

4.3 Data analysis

The data analysis of this project involves the bioinformatic analysis of WES data. This analysis will be divided into a three-step workflow: variant annotation, variant classification, and variant prioritisation. Variant annotation will be conducted using ANNOVAR (ANNOtate VARiation). ANNOVAR is a command-line-driven software tool designed for the annotation of genetic variants identified from high throughput sequencing data, including WES data. This tool can perform gene-based annotation to identify whether sequence variants result in amino acid changes, region-based annotation to identify variants in specific regions and provide genomic coordinates, and filter-based annotation to identify variants that are reported in databases such as the 1000 genomes project and gnomAD. Additionally, the functional consequence of variants is reported using *in silico* prediction tools such as SIFT, PolyPhen-2 and CADD, (Wang et al., 2010).

Next, variants will be classified using InterVar, a software tool designed for the clinical interpretation of genetic variants based on the American College of Medical Genetics and Genomics and Association for Molecular Pathology guidelines (ACMG/AMP) guideline (Richards et al., 2015). The input for this tool is the annotated file generated from ANNOVAR, while the output of InterVar is the variant class. It integrates the criteria outlined in these guidelines to provide an automated classification of variants (Li & Wang, 2017). Variants are classified as benign, likely benign, variant of uncertain significance (VUS), likely pathogenic and pathogenic. The ACMG/AMP classification will form part of the variant prioritisation strategy to filter out likely benign and benign variants.

One of the primary obstacles in disease gene discovery lies in the vast number of variants present in individual exomes. Approximately 30,000 variants are harboured in an individual exome,

10,000 of which are predicted to be nonsynonymous, splice site variants and small indels. (Robinson et al., 2011). Therefore, variants must be prioritised to generate a shortlist of germline variants in the cancer predisposition genes. A stepwise filtering approach will be applied to perform this step, as depicted in **Figure 1**. The proposed strategy that will be employed was compiled using previously described strategies (Nissen et al., 2018; Sefid Dashti & Gamieldien, 2017). Each step includes one exclusion criterion that will be applied to filter out variants.

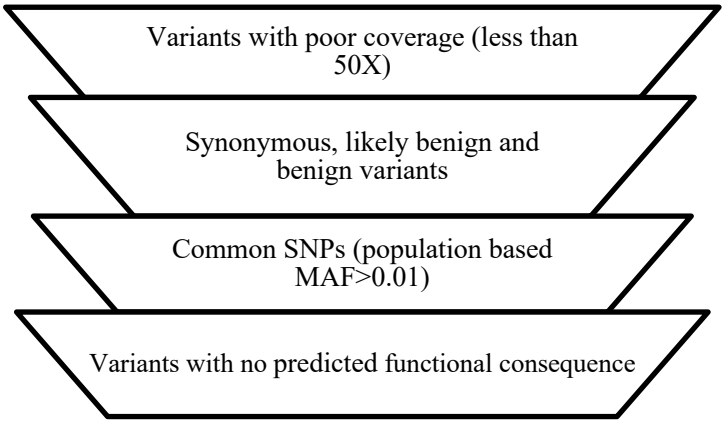


Figure 1: Stepwise filtering approach to prioritise variants and generate a shortlist of variants.

Following variant prioritisation, a shortlist of germline variants will be obtained for each of the 16 patients. Further analyses are required to interpret this data and determine which of the variants are pathogenic. The variant class provided by the variant classification step will be utilised to identify pathogenic variants specifically. Given the limited scope of this study, VUS will not be analysed further. Existing literature and databases such as ClinVar will be utilised to provide additional lines of evidence supporting the pathogenicity of the remaining variants.

Given that this study is concerned with the germline mutation landscape of OSCC, Knudson's two-hit hypothesis must be considered. In contrast to sporadic cancers, up to 10% of all cancers occur due to inherited genetic mutations in cancer predisposition genes. According to Knudson's

two-hit hypothesis, these genes follow a loss-of-function model. Therefore, an individual may inherit one mutant allele of a predisposition gene (Knudson, 1971). These germline heterozygote mutations are usually not disease-causing because their normal function is partially compensated by the presence of one wild-type allele. However, if a somatic mutation occurs in the second allele as a result of environmental or endogenous events, the given predisposition gene becomes inactivated, leading to the total loss of function and resultant carcinogenesis (Knudson, 1971). Therefore, in contrast to sporadic cancers in which two somatic events are required to inactivate a key cancer predisposition gene, germline mutations increase the carrier's risk for cancer as a single somatic event is required (Knudson, 1971; Wang, 2016). In the context of this study, this knowledge translates to specifically identifying protein-truncating mutations such as nonsense mutations, frameshift mutations and splice site mutations, as these disrupt the normal function of cancer predisposition genes, increasing the risk of cancer development.

Given the strong association between a family history of OC and an increased risk for OSCC, an important consideration in inherited cancer studies is familial aggregation (Chen et al., 2015). However, family history is not available for the patients in this study.

4 Ethics

This study will analyse germline whole exome sequencing data from patients with OSCC enrolled in the JCS. Patient data has already been de-identified, and no patients will be contacted during this study. The data will be analysed on the ZA-Wits-Core computing cluster.

This sub-study will fall under the parent studies M220881 (The Johannesburg Cancer Study, PI: Dr M Muchengeti & Dr WC Chen) and M2111154 (Investigation of genetic risk factors for African breast, cervical and oesophageal cancer, PI: Prof C Mathew). The germline WES dataset comprises

data from 16 histologically confirmed OSCC patients enrolled under M220881, and the whole exome data will be generated as part of M2111154. This MSc research project is under consideration by the Wits Human Research Ethics Committee (Medical).

5 Timing

	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov
Literature review										
Ethics application										
Preparing protocol										
Protocol assessment										
Construction of virtual gene panel										
Data analysis										
Research report writing										
Submission and presentation of research report										

6 Funding

The WES data was generated using the parent study's NRF Thuthuka funding (PI: Dr WC Chen). Bioinformatic analysis of this data does not require any funding as (1) the computational resources needed for the study are accessible via the ZA-Wits-Core Cluster (2) all the required bioinformatics tools such as ANNOVAR and InterVar open-source. In addition, the student will be supported by the Wits Postgraduate Merit Award.

7 Limitations

The limitations of this study include the fact that variants outside of the list of cancer predisposition genes panel will be missed. Given the technical limitations of WES, this study will also be unable to detect *de novo* CNVs. Additionally, the small study population of 16 individuals limits variant discovery as the chances of finding pathogenic variants in all 16 individuals are low. Lastly, the lack of representation of Southern African genomes in publicly available databases such as 1000 genomes and gnomAD (Ramsay, 2022) limits the estimation of robust variant allele frequencies, which forms part of the filtering steps in the analysis.

8 References

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

Akbari, M. R., Malekzadeh, R., Lepage, P., Roquis, D., Sadjadi, A. R., Aghcheli, K., Yazdanbod, A., Shakeri, R., Bashiri, J., Sotoudeh, M., Pourshams, A., Ghadirian, P., & Narod, S. A. (2011). Mutations in Fanconi anemia genes and the risk of esophageal cancer. *Human Genetics*, *129*(5), 573–582. https://doi.org/10.1007/s00439-011-0951-7

Alema, O., & Iva, B. (2014). Cancer of the esophagus: Histopathological sub-types in northern Uganda. *African Health Sciences*, *14*(1), 17–21. https://doi.org/10.4314/ahs.v14i1.4

Bartha, Á., & Győrffy, B. (2019). Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology. *Cancers*, *11*(11), 1725.

https://doi.org/10.3390/cancers11111725

Bye, H., Prescott, N. J., Lewis, C. M., Matejcic, M., Moodley, L., Robertson, B., van Rensburg, C., Parker, M. I., & Mathew, C. G. (2012). Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis*, *33*(11), 2155–2161. https://doi.org/10.1093/carcin/bgs262

Bye, H., Prescott, N. J., Matejcic, M., Rose, E., Lewis, C. M., Parker, M. I., & Mathew, C. G. (2011). Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis*, *32*(12), 1855–1861. https://doi.org/10.1093/carcin/bgr211

Cancer Statistics. (n.d.). *NICD*. Retrieved April 11, 2024, from https://www.nicd.ac.za/centres/national-cancer-registry/cancer-statistics/

Chang, J., Tan, W., Ling, Z., Xi, R., Shao, M., Chen, M., Luo, Y., Zhao, Y., Liu, Y., Huang, X., Xia, Y., Hu, J., Parker, J. S., Marron, D., Cui, Q., Peng, L., Chu, J., Li, H., Du, Z., … Lin, D. (2017). Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nature Communications*, *8*(1), 15290. https://doi.org/10.1038/ncomms15290

Chelule, P. K., Pegoraro, R. J., Gqaleni, N., & Dutton, M. F. (2006). The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. *Disease Markers*, *22*(5–6), 351–354. https://doi.org/10.1155/2006/980392

Chen, T., Cheng, H., Chen, X., Yuan, Z., Yang, X., Zhuang, M., Lu, M., Jin, L., & Ye, W. (2015). Family history of esophageal cancer increases the risk of esophageal squamous cell carcinoma. *Scientific Reports*, *5*(1), 16038. https://doi.org/10.1038/srep16038

Chen, W. C., Brandenburg, J.-T., Choudhury, A., Hayat, M., Sengupta, D., Swiel, Y., Babb de

Villiers, C., Ferndale, L., Aldous, C., Soo, C. C., Lee, S., Curtis, C., Newton, R., Waterboer, T., Sitas, F., Bradshaw, D., Abnet, C. C., Ramsay, M., Parker, M. I., … Mathew, C. G. (2023). Genome-wide association study of esophageal squamous cell cancer identifies shared and distinct risk variants in African and Chinese populations. *American Journal of Human Genetics*, *110*(10), 1690–1703. https://doi.org/10.1016/j.ajhg.2023.08.007

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

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

Dandara, C., Ballo, R., & Parker, M. I. (2005). CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer Letters*, *225*(2), 275–282. https://doi.org/10.1016/j.canlet.2004.11.004

Dandara, C., Li, D.-P., Walther, G., & Parker, M. I. (2006). Gene-environment interaction: The role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis*, *27*(4), 791–797. https://doi.org/10.1093/carcin/bgi257

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

Eltahir, H. A., Adam, A. A. M., Yahia, Z. A., Ali, N. F., Mursi, D. M., Higazi, A. M., Eid, N. A., Elhassan, A. M., Mohammed, H. S., & Ibrahim, M. E. (2012). P53 Codon 72 arginine/proline polymorphism and cancer in Sudan. *Molecular Biology Reports*, *39*(12), 10833–10836. https://doi.org/10.1007/s11033-012-1978-0

Goodwin, P. J., Phillips, K.-A., West, D. W., Ennis, M., Hopper, J. L., John, E. M., O'Malley, F. P., Milne, R. L., Andrulis, I. L., Friedlander, M. L., Southey, M. C., Apicella, C., Giles, G. G., & Longacre, T. A. (2012). Breast Cancer Prognosis in BRCA1 and BRCA2 Mutation Carriers: An International Prospective Breast Cancer Family Registry Population-Based Cohort Study. *Journal of Clinical Oncology*, *30*(1), 19–26. https://doi.org/10.1200/JCO.2010.33.0068

Huang, F.-L., & Yu, S.-J. (2018). Esophageal cancer: Risk factors, genetic association, and treatment. *Asian Journal of Surgery*, *41*(3), 210–215. https://doi.org/10.1016/j.asjsur.2016.10.005

Jiang, Y.-X., Zhang, D.-W., Chen, Y., Sun, H.-H., Xu, S.-C., & Gao, H.-J. (2015). The characteristics of oesophageal squamous cell carcinoma: An analysis of 1317 cases in southeastern China. *Contemporary Oncology*, *19*(2), 137–141. https://doi.org/10.5114/wo.2015.51419

Kaz, A. M., & Grady, W. M. (2014). Epigenetic Biomarkers in Esophageal Cancer. *Cancer Letters*, *342*(2), 10.1016/j.canlet.2012.02.036. https://doi.org/10.1016/j.canlet.2012.02.036

Knudson, A. G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*, *68*(4), 820–823.

Li, D., Dandara, C., & Parker, M. I. (2005). Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clinical Chemistry and Laboratory Medicine*, *43*(4), 370–375. https://doi.org/10.1515/CCLM.2005.067

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

Li, D.-P., Dandara, C., Walther, G., & Parker, M. I. (2008). Genetic polymorphisms of alcohol metabolising enzymes: Their role in susceptibility to oesophageal cancer. *Clinical Chemistry and Laboratory Medicine*, *46*(3), 323–328. https://doi.org/10.1515/CCLM.2008.073

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, *100*(2), 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004

Matejcic, M., Li, D., Prescott, N. J., Lewis, C. M., Mathew, C. G., & Parker, M. I. (2011). Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: A case-control study. *PloS One*, *6*(12), e29366. https://doi.org/10.1371/journal.pone.0029366

Matejcic, M., Vogelsang, M., Wang, Y., & Parker, M. I. (2015). Erratum to: NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC Cancer*, *15*, 658. https://doi.org/10.1186/s12885-015-1681-3

Munishi, M. O., Hanisch, R., Mapunda, O., Ndyetabura, T., Ndaro, A., Schüz, J., Kibiki, G., & McCormack, V. (2015). Africa's oesophageal cancer corridor—Do hot beverages contribute? *Cancer Causes & Control : CCC*, *26*(10), 1477–1486. https://doi.org/10.1007/s10552-015-0646-9

Murphy, G., McCormack, V., Abedi-Ardekani, B., Arnold, M., Camargo, M. C., Dar, N. A., Dawsey, S. M., Etemadi, A., Fitzgerald, R. C., Fleischer, D. E., Freedman, N. D., Goldstein, A. M., Gopal, S., Hashemian, M., Hu, N., Hyland, P. L., Kaimila, B., Kamangar, F., Malekzadeh, R., … Brennan, P. (2017). International cancer seminars: A focus on esophageal squamous cell carcinoma. *Annals of Oncology*, *28*(9), 2086–2093. https://doi.org/10.1093/annonc/mdx279

Nissen, S. K., Christiansen, M., Helleberg, M., Kjær, K., Jørgensen, S., Gerstoft, J., Katzenstein, T., Benfield, T., Kronborg, G., Larsen, C., Laursen, A., Pedersen, G., Jakobsen, M., Tolstrup, M., & Mogensen, T. (2018). Whole Exome Sequencing of HIV-1 long-term non-progressors identifies rare variants in genes encoding innate immune sensors and signaling molecules. *Scientific Reports*, *8*. https://doi.org/10.1038/s41598-018-33481-0

Ramsay, M. (2022). African genomic data sharing and the struggle for equitable benefit. *Patterns*, *3*(1), 100412. https://doi.org/10.1016/j.patter.2021.100412

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Robinson, P., Krawitz, P., & Mundlos, S. (2011). Strategies for exome and genome sequence data analysis in disease-gene discovery projects. *Clinical Genetics*, *80*(2), 127–132. https://doi.org/10.1111/j.1399-0004.2011.01713.x

Sefid Dashti, M. J., & Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. *BioTechniques*, *62*(1), 18–30. https://doi.org/10.2144/000114492

Simba, H., Kuivaniemi, H., Abnet, C. C., Tromp, G., & Sewram, V. (2023). Environmental and life-style risk factors for esophageal squamous cell carcinoma in Africa: A systematic review and meta-analysis. *BMC Public Health*, *23*(1), 1782. https://doi.org/10.1186/s12889-023-16629-0

Simba, H., Kuivaniemi, H., Lutje, V., Tromp, G., & Sewram, V. (2019). Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in Genetics*, *10*, 642. https://doi.org/10.3389/fgene.2019.00642

Strickland, N. J., Matsha, T., Erasmus, R. T., & Zaahl, M. G. (2012). Molecular analysis of Ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International Journal of Cancer*, *131*(3), 623–632. https://doi.org/10.1002/ijc.26418

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Vogelsang, M., Wang, Y., Veber, N., Mwapagha, L. M., & Parker, M. I. (2012). The Cumulative Effects of Polymorphisms in the DNA Mismatch Repair Genes and Tobacco Smoking in Oesophageal Cancer Risk. *PLoS ONE*, *7*(5), e36962. https://doi.org/10.1371/journal.pone.0036962

Vos, M., Adams, C. H., Victor, T. C., & van Helden, P. D. (2003). Polymorphisms and mutations found in the regions flanking exons 5 to 8 of the TP53 gene in a population at high risk for esophageal cancer in South Africa. *Cancer Genetics and Cytogenetics*, *140*(1), 23–30. https://doi.org/10.1016/s0165-4608(02)00638-6

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q. (2016). Cancer predisposition genes: Molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes. *Acta Pharmacologica Sinica*, *37*(2), 143–149. https://doi.org/10.1038/aps.2015.89

Zaahl, M. G., Warnich, L., Victor, T. C., & Kotze, M. J. (2005). Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer Genetics and Cytogenetics*, *159*(1), 48–52. https://doi.org/10.1016/j.cancergencyto.2004.09.017

Zeng, B., Huang, P., Du, P., Sun, X., Huang, X., Fang, X., & Li, L. (2021). Comprehensive Study of Germline Mutations and Double-Hit Events in Esophageal Squamous Cell Cancer. *Frontiers in Oncology*, *11*, 637431. https://doi.org/10.3389/fonc.2021.637431

Zhang, X.-M., & Guo, M.-Z. (2010). The value of epigenetic markers in esophageal cancer. *Frontiers of Medicine in China*, *4*(4), 378–384. https://doi.org/10.1007/s11684-010-0230-3

Appendix D: Plagiarism Declaration and Turnitin Report

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Zahraa Dangor (Student number: 2208826) am a student registered for the degree of MSc (Med) Genomic Medicine in the academic year 2024.

I hereby declare the following:

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

Signature: _____ Date: 30/11/2024

1

to submit to turnitin .pdf

ORIGINALITY REPORT

9%	6%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.helix.com Internet Source	1%
2	hdl.handle.net Internet Source	1%
3	www.imrpress.com Internet Source	1%
4	link.springer.com Internet Source	1%
5	www.frontiersin.org Internet Source	<1%
6	Submitted to National School of Healthcare Science Student Paper	<1%
7	actaneurocomms.biomedcentral.com Internet Source	<1%
8	Jasmin J Hopkins, Matthew N Wakeling, Matthew B Johnson, Sarah E Flanagan, Thomas W Laver. "REVEL is better at predicting pathogenicity of loss-of-function	<1%

	than gain-of-function variants", Cold Spring Harbor Laboratory, 2023 Publication	
9	Submitted to University of Sydney Student Paper	<1 %
10	www.future-science.com Internet Source	<1 %
11	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1 %
12	Submitted to University of Witwatersrand Student Paper	<1 %
13	Submitted to University of Melbourne Student Paper	<1 %
14	hal.science Internet Source	<1 %
15	Bartha, Győrffy. "Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology", Cancers, 2019 Publication	<1 %
16	Wenlong Carl Chen, Elvira Singh, Mazvita Muchengeti, Debbie Bradshaw et al. "Johannesburg Cancer Study (JCS): contribution to knowledge and opportunities arising from 20 years of data collection in an African setting", Cancer Epidemiology, 2020 Publication	<1 %

17	www.idexlab.com Internet Source	<1%
18	Gil-Rodríguez, María Concepción, Matthew A. Deardorff, Morad Ansari, Christopher A. Tan, Ilaria Parenti, Carolina Baquero-Montoya, Lilian B. Ousager, Beatriz Puisac, María Hernández-Marcos, María Esperanza Teresa-Rodrigo, Iñigo Marcos-Alcalde, Jan-Jaap Wesselink, Silvia Lusa-Bernal, Emilia K. Bijlsma, Diana Braunholz, Inés Bueno-Martinez, Dinah Clark, Nicola S. Cooper, Cynthia J. Curry, Richard Fisher, Alan Fryer, Jaya Ganesh, Cristina Gervasini, Gabriele Gillessen-Kaesbach, Yiran Guo, Hakon Hakonarson, Robert J. Hopkin, Maninder Kaur, Brendan J. Keating, María Kibaek, Esther Kinning, Tjitske Kleefstra, Antonie D. Kline, Ekaterina Kuchinskaya, Lidia Larizza, Yun R. Li, Xuanzhu Liu, Milena Mariani, Jonathan D. Picker, Ángeles Pié, Jelena Pozojevic, Ethel Queralt, Julie Richer, Elizabeth Roeder, Anubha Sinha, Richard H. Scott, Joyce So, Katherine A. Wusik, Louise Wilson, Jianguo Zhang, Paulino Gómez-Puertas, César H. Casale, Lena Ström, Angelo Selicorni, Feliciano J. Ramos, Laird G. Jackson et al. "De novo Heterozygous Mutations in SMC3 Cause a Range of Cornelia	<1%

Appendix E: Author guidelines

Frontiers

Writing and formatting

Title

The title should be concise, omitting terms that are implicit and, where possible, be a statement of the main result or conclusion presented in the manuscript. Abbreviations should be avoided within the title.

Witty and creative titles are welcome, but only if relevant and within measure. Consider if a title meant to be thought-provoking might be misinterpreted as offensive or alarming. In extreme cases, the editorial office may veto a title and propose an alternative.

Authors should avoid:

- titles that are a mere question without giving the answer
- unambitious titles, for example starting with 'Towards,' 'A description of,' 'A characterization of' or 'Preliminary study on'
- vague titles, for example starting with 'Role of', 'Link between', or 'Effect of' that do not specify the role, link, or effect
- including terms that are out of place, for example the taxonomic affiliation apart from species name.

Abstract

As a primary goal, the abstract should make the general significance and conceptual advance of the work clearly accessible to a broad readership. The abstract should be no longer than a single paragraph and should be structured, for example, according to the **IMRAD format**. For the specific structure of the abstract, authors should follow the requirements of the article type or journal to which they're submitting. Minimize the use of abbreviations and do not cite references, figures or tables.

For clinical trial articles, please include the unique identifier and the URL of the publicly accessible website on which the trial is registered.

Manuscript length

We encourage you to closely follow the article word count lengths given in the 'Article types' page of the journals. The manuscript length includes only the main body of the text, footnotes, and all citations within it, and excludes the abstract, section titles, figure and table captions, funding statement, acknowledgments, and references in the bibliography. Please indicate the number of words and the number of figures and tables included in your manuscript on the first page.

Sections

The manuscript is organized by headings and subheadings. The section headings should be those appropriate for your field and the research itself. You may insert up to 5 heading levels into your manuscript (i.e.,: 3.2.2.1.2 Heading Title).

For Original Research articles, it is recommended to organize your manuscript in the following sections or their equivalents for your field.

Introduction
Succinct, with no subheadings.

Materials and methods
This section may be divided by subheadings and should contain sufficient detail so that when read in conjunction with cited references, all procedures can be repeated. For experiments reporting results on animal or human subject research, an ethics approval statement should be included in this section (for further information, see the 'Bioethics' section of our **policies and publication ethics**.)

Results
This section may be divided by subheadings. Footnotes should not be used and must be transferred to the main text.

Discussion
This section may be divided by subheadings. Discussions should cover the key findings of the study: discuss any prior research related to the subject to place the novelty of the discovery in the appropriate context, discuss the potential shortcomings and limitations on their interpretations, discuss their integration into the current understanding of the problem and how this advances the current views, speculate on the future direction of the research, and freely postulate theories that could be tested in the future.

For further information, please check the descriptions defined in the journal's 'Article types' page, in the 'For authors' menu on every journal page.

Language editing

Frontiers requires manuscripts submitted to meet international English language standards to be considered for publication.

For authors who would like their manuscript to receive language editing or proofreading to improve the clarity of the manuscript and help highlight their research, we recommend the language-editing services provided by the following external partners.

Note that sending your manuscript for language editing does not imply or guarantee that it will be accepted for publication by a Frontiers journal. Editorial decisions on the scientific content of a manuscript are independent of whether it has received language editing or proofreading by these partner services or other services.

Editage
We recommend the language-editing service provided by our external partner Editage. These services may be particularly useful for researchers for whom English is not the primary language. They can help to improve the grammar, syntax, and flow of your manuscript prior to submission. Frontiers' authors will receive a 10% discount by visiting the following link: **editage.com/frontiers**.

The Charlesworth Group
We recommend the Charlesworth Group's author services, who has a long-standing track record in language editing and proofreading. This is a third-party service for which Frontiers' authors will receive a 10% discount by visiting the following link: **www.cwauthors.com/frontiers**.

Frontiers推荐您使用在英语语言编辑和校对领域具有悠久历史和良好口碑的查尔斯沃思作者服务。此项服务由第三方为您提供，Frontiers中国作者通过此链接提交稿件时可获得10%的特别优惠: **www.cwauthors.com.cn/frontiers**.

Language style

The default language style at Frontiers is American English. If you prefer your article to be formatted in British English, please specify this on the first page of your manuscript. For any questions regarding style, we recommend authors to consult the **Chicago Manual of Style**.

Inclusive language guidelines

Frontiers is an inclusive publisher and we ask that all submissions are in line with our inclusive language policy. When preparing your manuscript for submission, take a mindful approach towards personal biases and a concerted effort to limit their influence. Authors should remove any suggestion or implication of superiority or inferiority of one person over another based on age, gender, race, ethnicity, culture, sexual orientation, disability, religion, or socio-economic class. We ask authors to use inclusive language practices and awareness of diversity, equity, and inclusion into their research and keep it at the forefront during the composition of their findings.

External guidance that may be useful is available at C4DISC's **Guidelines on Inclusive Language and Images in Scholarly Communication**.

Furthermore, when drafting your work, please take into account the following considerations

In general, seek to avoid

- language that could be deemed insulting, profane, or derogatory.
- descriptors that identify personal attributes such as age, gender, race, ethnicity, culture, sexual orientation, disability, or health conditions, where they are not critically relevant to the discussion.
- any form of language that suggests a particular culture or group as the default or standard.

And where feasible:

- proactively ask individuals or groups how they would prefer to be referenced.
- adhere to the SAGER guidelines for reference to sex and gender in research.

Remember, the language we use can influence perceptions, evoke emotions, and shape perspectives. Let's work together to nurture an inclusive, respectful, and empowering discourse in science.

Guidelines for artificial intelligence and related technologies

These guidelines cover acceptable uses of generative AI technologies such as Large Language Models (ChatGPT, Jasper) and text-to-image generators (DALL-E 2, Midjourney, Stable Diffusion) in the writing or editing of manuscripts submitted to Frontiers.

Use of AI to create written or visual content

Authors should not list a generative AI technology as a co-author or author of any submitted manuscript. Generative AI technologies cannot be held accountable for all aspects of a manuscript and consequently do not meet the criteria required for authorship.

If the author of a submitted manuscript has used written or visual content produced by or edited using a generative AI technology, this use must follow all Frontiers guidelines and policies. Specifically, the author is responsible for checking the factual accuracy of any content created by the generative AI technology. This includes, but is not limited to, any quotes, citations or references. Figures produced by or edited using a generative AI technology must be checked to ensure they accurately reflect the data presented in the manuscript. Authors must also check that any written or visual content produced by or edited using a generative AI technology is free from plagiarism.

If the author of a submitted manuscript has used written or visual content produced by or edited using a generative AI technology, such use must be acknowledged in the acknowledgements section of the manuscript and the methods section if applicable. This explanation must list the name, version, model, and source of the generative AI technology. We encourage authors to upload all input prompts provided to a generative AI technology and outputs received from a generative AI technology in the supplementary files for the manuscript.

The entire document should be single-spaced and must contain page and line numbers in order to facilitate the review process. The manuscript should be written using either Word or LaTeX. See above for templates.

Abbreviations and nomenclatures

The use of abbreviations should be kept to a minimum. Non-standard abbreviations should be avoided unless they appear at least four times, and must be defined upon first use in the main text. Consider also giving a list of non-standard abbreviations at the end, immediately before the acknowledgments.

Equations should be inserted in editable format from the equation editor.

Italicize gene symbols and use the approved gene nomenclature where it is available. For human genes, please refer to the HUGO Gene Nomenclature Committee (**HGNC**). New symbols for human genes should be submitted to the HGNC **here**. Common alternative gene aliases may also be reported, but should not be used alone in place of the HGNC symbol. Nomenclature committees for other species are listed **here**. Protein products are not italicized.

We encourage the use of Standard International Units in all manuscripts.

Chemical compounds and biomolecules should be referred to using systematic nomenclature, preferably using the recommendations by the International Union of Pure and Applied Chemistry (IUPAC).

Astronomical objects should be referred to using the nomenclature given by the International Astronomical Union (IAU) provided **here**.

Life Science Identifiers (LSIDs) for ZOOBANK registered names or nomenclatural acts should be listed in the manuscript before the keywords. An LSID is represented as a uniform resource name (URN) with the following format:
urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]

For more information on LSIDs please see the 'Code' section of our **policies and publication ethics**.

Enhancing search engine optimization (SEO)

There are a few simple ways to maximize your article's discoverability and search results.

- Include a few of your article's keywords in the title of the article
- Do not use long article titles
- Pick 5-8 keywords using a mix of generic and more specific terms on the article subject(s)
- Use the maximum amount of keywords in the first two sentences of the abstract
- Use some of the keywords in level 1 headings

References

Preparing and formatting references

Submissions to Frontiers must be grounded in relevant and up to date peer-reviewed, academic research, and this should be reflected in the accompanying reference lists.

Authors are welcome to use online referencing tools in preparation of their manuscript. Some useful resources include **RefMe**, **Zotero**, and **Mendeley**.

- The citation of non-academic and non-peer-reviewed sources (e.g. blog posts, website content), as well as anonymous sources or commercial websites should be avoided or kept to a minimum
- Authors should avoid citing content that is not directly relevant to the scope of the article and the journal
- Reference lists should reflect the current status of knowledge in the field, avoid bias, and not include a high proportion of citations to the same authors or sources, school of thought, etc.
- The length of the reference list should be appropriate depending on the article type, covering the relevant literature through sufficient referencing
- Authors should ensure that references are accurate, that all links are accessible, and that the citations/references adhere to the reference styles outlined below

Frontiers' journals use one of two reference styles, either Harvard (author-date) or Vancouver (numbered). These formats should be adhered to for the in-text citations and the reference lists. Please check our **help center** to find the correct style for the journal to which you're submitting.

- All citations of published works in the text, figures, or tables must be in the reference list and vice-versa.

- The names of the first six authors followed by et al. and the DOI (when available) should be provided.
- Given names of authors should be abbreviated to initials (e.g. Smith, J., Lewis, C.S., etc.).
- The reference list should only include articles that are published or accepted.
- Unpublished data, submitted manuscripts, or personal communications should be cited within the text only, for article types that allow such inclusions. Where additional details are available, these will be included as footnotes.
- For accepted but unpublished works use 'in press' instead of page numbers.
- Data sets that have been deposited to an online repository should be included in the reference list. Include the version and unique identifier when available.
- Personal communications should be documented by a letter of permission.
- Website URLs should be included as footnotes.
- Any inclusion of verbatim text must be contained in quotation marks and should clearly reference the original source.
- Preprints can be cited provided that a DOI or archive URL is available, and the citation clearly mentions that the contribution is a preprint. If a peer-reviewed journal publication for the same preprint exists, the official journal publication is the preferred source. See the preprints section for each reference style below for more information.

Acknowledgements

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors. Should the content of the manuscript have previously appeared online, such as in a thesis or preprint, this should be mentioned here, in addition to listing the source within the reference list.

Acknowledgements are designed to recognize individuals directly involved in the research. We request that authors avoid personal expressions of gratitude toward figures or entities unrelated or indirectly related to the production of the manuscript. While such acknowledgements are understood to be significant to the authors, they are considered indirectly involved in the research unless they pertain to specific research activities. It remains the responsibility of the authors to ensure that these guidelines are adhered to in the final publication.

Templates

If working with Word please use our **Word templates**. If you wish to submit your article as LaTeX, we recommend our **LaTeX templates**.

For LaTeX files, please ensure all relevant manuscript files are uploaded: .tex file, PDF, and .bib file (if the bibliography is not already included in the .tex file).

During the **interactive review**, authors are encouraged to upload versions using track changes. Editors and reviewers can only download the PDF file of the submitted manuscript.

Figures, tables, and images

Figures, tables, and images: rights and permissions

All figures, tables, and images will be published under a Creative Commons **CC-BY license**, and permission must be obtained for use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

For additional information, please see the 'Image manipulation' section of our **policies and publication ethics**.

Figures and images: style guidelines

We require figures to be submitted individually, in the same order as they are referred to in the manuscript; the figures will then be automatically embedded at the end of the submitted manuscript. Ensure that each figure is mentioned in the text and in numerical order.

For figures with more than one panel, panels should be clearly indicated using labels (A), (B), (C), (D), etc. However, do not embed the part labels over any part of the image. These labels will be replaced during typesetting according to Frontiers' journal style. For graphs, there must be a self-explanatory label (including units) along each axis.

For LaTeX files, figures should be included in the provided PDF. In case of acceptance, our production office might require high-resolution files of the figures included in the manuscript in EPS, JPEG or TIF/TIFF format.

To upload more than one figure at a time, save the figures (labeled in order of appearance in the manuscript) in a zip file and upload them as 'Supplementary material presentation.'

Please note that figures not in accordance with the guidelines will cause substantial delay during the production process.

Captions

Captions should be preceded by the appropriate label, for example 'Figure 1.' Figure captions should be placed at the end of the manuscript. Figure panels are referred to by bold capital letters in brackets: (A), (B), (C), (D), etc.

Image size and resolution requirements

Figures should be prepared with the PDF layout in mind. Individual figures should not be longer than one page and with a width that corresponds to one column (85 mm) or two columns (180 mm).

All images must have a resolution of 300 dpi at final size. Check the resolution of your figure by enlarging it to 150%. If the image appears blurry, jagged, or has a stair-stepped effect, the resolution is too low.

The text should be legible and of high quality. The smallest visible text should be no less than eight points in height when viewed at actual size.

Solid lines should not be broken up. Any lines in the graphic should be no smaller than two points wide.

Please note that saving a figure directly as an image file (JPEG, TIF) can greatly affect the resolution of your image. To avoid this, one option is to export the file as PDF, then convert into TIFF or EPS using a graphics software.

Format and color image mode

The following formats are accepted: TIF/TIFF (.tif/.tiff), JPEG (.jpg), and EPS (.eps) (upon acceptance). Images must be submitted in the color mode RGB.

Images of chemical structures

Chemical structures should be prepared using ChemDraw or a similar program. If working with ChemDraw please use our **ChemDraw template**. If working with another program please follow the guidelines below.

- Drawing settings: chain angle, 120° bond spacing, 18% width; fixed length, 14.4 pt; bold width, 2.0 pt; line width, 0.6 pt; margin width, 1.6 pt; hash spacing, 2.5 pt. Scale 100% Atom Label settings: font, Arial; size, 8 pt
- Assign all chemical compounds a bold, Arabic numeral in the order in which the compounds are presented in the manuscript text.

Table requirements and style guidelines

Tables should be inserted at the end of the manuscript in an editable format. If you use a word processor, build your table in Word. If you use a LaTeX processor, build your table in LaTeX. An empty line should be left before and after the table.

Table captions must be placed immediately before the table. Captions should be preceded by the appropriate label, for example 'Table 1.' Please use only a single paragraph for the caption.

Ensure that each table is mentioned in the text and in numerical order.

Large tables covering several pages cannot be included in the final PDF for formatting reasons. These tables will be published as supplementary material.

Tables which are not according to the above guidelines will cause substantial delay during the production process.

During production, tables will be formatted according to Frontiers' house style. Here is an example of a formatted table.

This is a heading	This is a heading
This is a subheading	
This is an example text	This is an example text
This is an example text	This is an example text
This is an example text	This is an example text

Accessibility

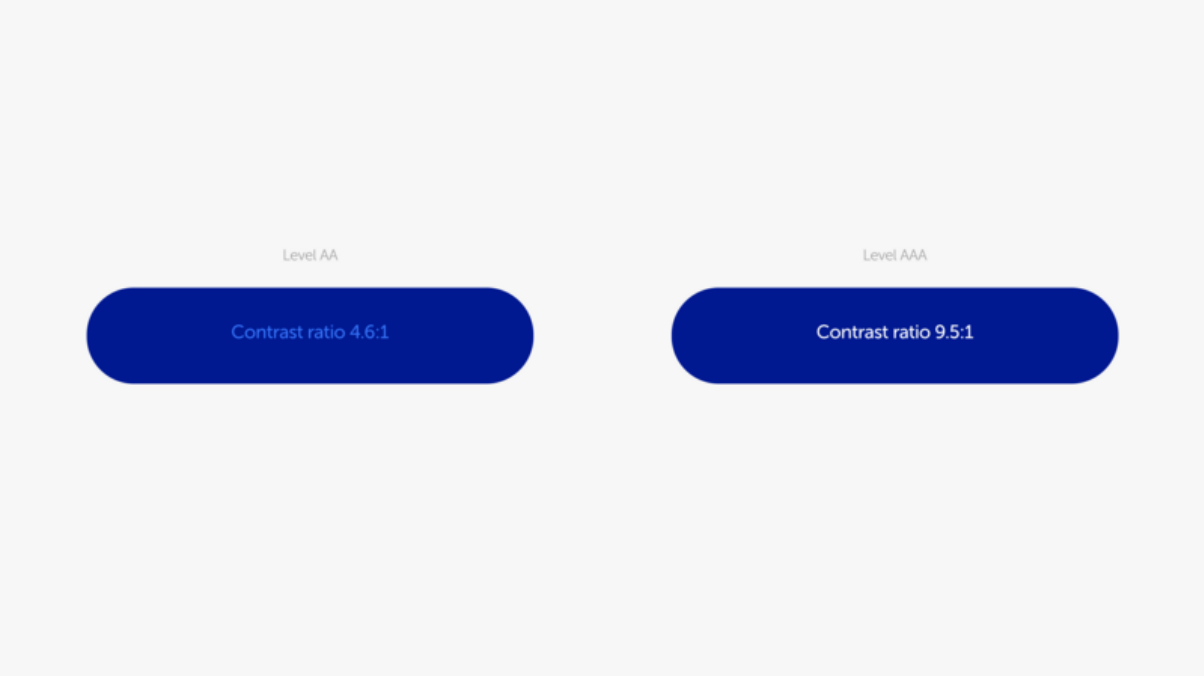
We encourage authors to make the figures and visual elements of their articles accessible for the visually impaired. Effective use of color can help people with low visual acuity, or color blindness, understand all the content of an article.

These guidelines are easy to implement and are in accordance with the W3C Web Content Accessibility Guidelines (**WCAG 2.1**), the standard for web accessibility best practices.

Ensure sufficient contrast between text and its background
People who have low visual acuity or color blindness could find it difficult to read text with low contrast background color. Try using colors that provide maximum contrast.

WC3 recommends the following contrast ratio levels:

- Level AA, contrast ratio of at least 4.5:1
- Level AAA, contrast ratio of at least 7:1

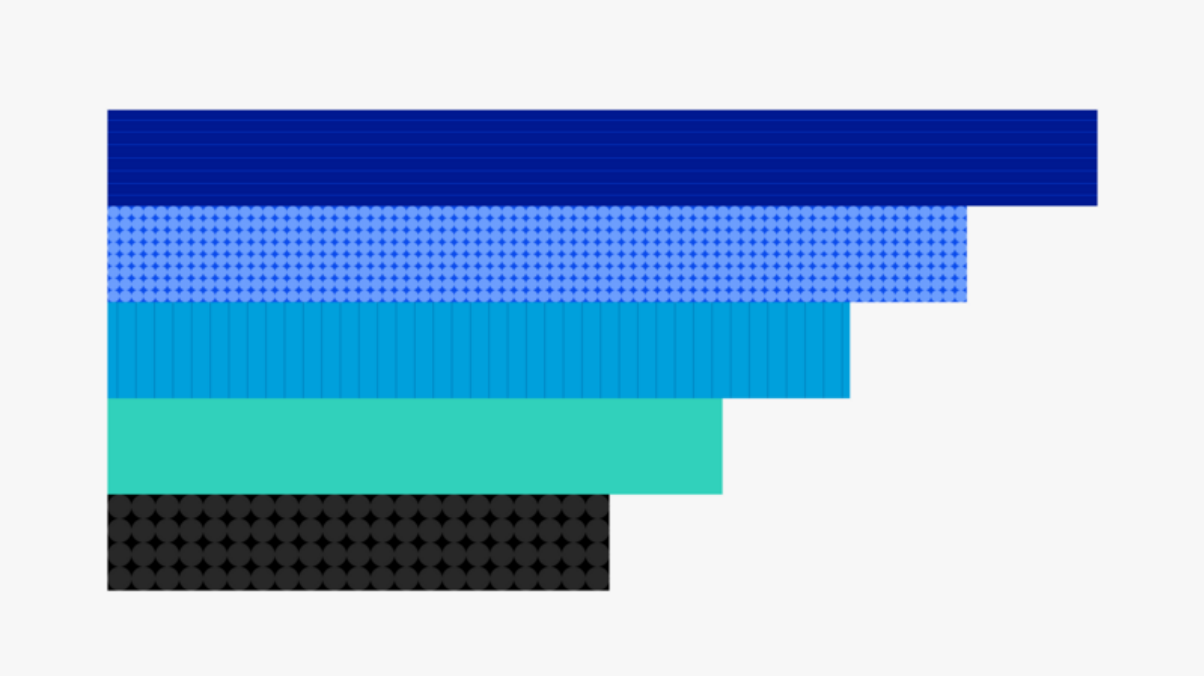


You can verify the contrast ratio of your palette with these online ratio checkers:

- WebAIM
- Color Safe

Avoid using red or green indicators
More than 99% of color-blind people have a red-green color vision deficiency.

Avoid using only color to communicate information
Elements with complex information like charts and graphs can be hard to read when only color is used to distinguish the data. Try to use other visual aspects to communicate information, such as shape, labels, and size. Incorporating patterns into the shape fills also make differences clearer; for an example please see below:



Supplementary material

Selecting supplementary material

Data that are not of primary importance to the text, or which cannot be included in the article because they are too large or the current format does not permit it (such as videos, raw data traces, and PowerPoint presentations), can be uploaded as supplementary material during the submission procedure and will be displayed along with the published article. All supplementary files are deposited to figshare for permanent storage and receive a DOI.

Supplementary material is not typeset, so please ensure that all information is clearly presented without tracked changes/highlighted text/line numbers, and the appropriate caption is included in the file. To avoid discrepancies between the published article and the supplementary material, please do not add the title, author list, affiliations or correspondence in the supplementary files.

File type requirements

The supplementary material can be uploaded as:

- data sheet (Word, Excel, CSV, CDX, FASTA, PDF or Zip files)
- presentation (PowerPoint, PDF or Zip files)
- image (CDX, EPS, JPEG, PDF, PNG or TIF/TIFF),
- table (Word, Excel, CSV or PDF)
- audio (MP3, WAV or WMA)
- video (AVI, DIVX, FLV, MOV, MP4, MPEG, MPG or WMV).

Technical requirements for supplementary images:

- 300 DPIs
- RGB color mode.

For supplementary material templates (LaTeX and Word), see our **supplementary material templates**.

Submitting information

Choosing where to submit

Open access and copyright

All Frontiers' articles from July 2012 onwards are published with open access under the Creative Commons **CC-BY license** (the current version is CC-BY, version 4.0). This means that the author(s) retains copyright, but the content is free to download, distribute, and adapt for commercial or non-commercial purposes, given appropriate attribution to the original article.

Upon submission, the author(s) grants Frontiers a license to publish, including to display, store, copy, and reuse the content. The CC-BY Creative Commons attribution license enables anyone to use the publication freely, given appropriate attribution to the author(s) and citing Frontiers as the original publisher. The CC-BY Creative Commons attribution license does not apply to third-party materials that display a copyright notice to prohibit copying. Unless the third-party content is also subject to a CC-BY Creative Commons attribution license, or an equally permissive license, the author(s) must comply with any third-party copyright notices.

Registration with Frontiers

Corresponding and all submitting authors must **register** with Frontiers before submitting an article. You must be logged in to your personal Frontiers account to submit an article.

For any co-author who would like their name on the article abstract page and PDF to be linked to a Frontiers profile on the **Loop network**, please **register** before the final publication of the paper.

Article type

We require authors to select the appropriate article type for their manuscript and to comply with the article type descriptions defined in the journal's 'Article types' page, which can be found under the 'About journal' menu in 'For authors' on every Frontiers journal page. Please pay close attention to the word count limits.

Keywords

All article types require a minimum of five and a maximum of eight keywords.

CrossMark policy

CrossMark is a multi-publisher initiative to provide a standard way for readers to locate the current version of a piece of content. By applying the CrossMark logo Frontiers is committed to maintaining the content it publishes and to alerting readers to changes if and when they occur.

Clicking on the CrossMark logo will tell you the current status of a document and may also give you additional publication record information about the document.

For Corrigenda, General Commentaries, and Editorials, the title of your manuscript should have the following format.

- 'Corrigendum: [Title of original article]'
- General Commentaries:
  'Commentary: [Title of original article]'
  'Response: Commentary: [Title of original article]'
- 'Editorial: [Title of Research Topic]'

Authors and affiliations

All names are listed together and separated by commas. Provide exact and correct author names as these will be indexed in official archives. Affiliations should be keyed to the author's name with superscript numbers and be listed as follows:

- Laboratory, Institute, Department, Organization, City, State abbreviation (only for United States, Canada, and Australia), and Country (without detailed address information such as city zip codes or street names).

Example: Max Maximus1
1 Department of Excellence, International University of Science, New York, NY, United States.

Correspondence

The corresponding author(s) should be marked with an asterisk in the author list. Provide the exact contact email address of the corresponding author(s) in a separate section.

Example: Max Maximus*
maximus@iuscience.edu

If any authors wish to include a change of address, list the present address(es) below the

correspondence details using a unique superscript symbol keyed to the author(s) in the author list.

Equal contributions

The authors who have contributed equally should be marked with a symbol (†) in the author list of the doc/latex and pdf files of the manuscript uploaded at submission.

Please use the appropriate standard statement(s) to indicate equal contributions:

- **Equal contribution:** These authors contributed equally to this work
- **First authorship:** These authors share first authorship
- **Senior authorship:** These authors share senior authorship
- **Last authorship:** These authors share last authorship
- **Equal contribution and first authorship:** These authors contributed equally to this work and share first authorship
- **Equal contribution and senior authorship:** These authors contributed equally to this work and share senior authorship
- **Equal contribution and last authorship:** These authors contributed equally to this work and share last authorship

Example: Max Maximus 1†, John Smith2† and Barbara Smith1
†These authors contributed equally to this work and share first authorship

Consortium/group and collaborative authors

Consortium/group authorship should be listed in the manuscript with the other author(s).

In cases where authorship is retained by the consortium/group, the consortium/group should be listed as an author separated by a comma or 'and'. The consortium/group name will appear in the author list, in the citation, and in the copyright. If provided, the consortium/group members will be listed in a separate section at the end of the article.

For the collaborators of the consortium/group to be indexed in PubMed, they do not have to be inserted in the Frontiers submission system individually. However, in the manuscript itself, provide a section with the name of the consortium/group as the heading followed by the list of collaborators, so they can be tagged accordingly and indexed properly.

Example: John Smith, Barbara Smith and The Collaborative Working Group.

In cases where work is presented by the author(s) on behalf of a consortium/group, it should

be included in the author list separated with the wording 'for' or 'on behalf of.' The consortium/group will not retain authorship and will only appear in the author list.

Example: John Smith and Barbara Smith on behalf of The Collaborative Working Group.

These guidelines cover acceptable uses of generative AI technologies such as Large Language Models (ChatGPT, Jasper) and text-to-image generators (DALL-E 2, Midjourney, Stable Diffusion) in the writing or editing of manuscripts submitted to Frontiers.

Scope statement

When you submit your manuscript, you will be required to summarize in 200 words your manuscript's scope and its relevance to the journal and/or specialty section you're submitting to. The aim is to convey to editors and reviewers how the contents of your manuscript fit within the selected journal's scope.

This statement will not be published with your article if it is accepted for publication. The information will be used during the initial validation and review processes to assess whether the manuscript is a suitable fit for the chosen journal and specialty.

We encourage you to consider carefully where to submit your manuscript, as submissions to an unsuitable journal or specialty will result in delays and increase the likelihood of manuscript rejection.

If you are submitting to a Research Topic, please also clarify how your submission is suited to the specific topic.