The Clinical Genomics of African Oesophageal Cancer

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine

Johannesburg, 2024

Nerija Lamantha Ngundu

Student number: 2461584

Supervisors:

Prof Christopher Mathew, Distinguished Professor in Human Genetics, BSc (Hons), PhD, FRCPath

Prof Clement Penny, Professor, BSc (Hons), PhD

Declaration

I, **Nerija Lamantha Ngundu** declare that this dissertation is my own, unaided work. It is being submitted for a degree of Master of Science in Medicine to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other university.

(Signature of Candidate)

26th day of September 2024 in Johannesburg, South Africa

This dissertation is dedicated to the memory of my beloved grandmother

Mrs Helene (Koko) Kilola

[12 December 1930 – 07 February 2024]

Publications

No publication as of yet.

A paper on which the candidate is a co-author has been submitted for publication to Molecular Genetics & Genomic Medicine Journal entitled: Somatic mutation profiles in non-tobacco smoking and non-alcohol drinking South African female oesophageal squamous cell carcinoma patients of African ancestry

Lucien Ferndale, Wenlong Carl Chen, Phelelani Thokozani Mpangese, Jean-Tristan Brandenburg, Lamantha Nerija Ngundu, Mishalan Moodley, Reubina Wadee, Colleen A. Wright, M. Iqbal Parker, Pascale Willem, Colleen Aldous, Christopher G. Mathew

Abstract

In South Africa, oesophageal cancer is responsible for the 6th highest cancer-related deaths, with oesophageal squamous cell carcinoma (OSCC) being the most prevalent type at an incidence rate of 8.6/cases/100,000 for males and 4.7/cases/100,000 for females. It is often diagnosed at a late stage due to its asymptomatic nature, making it too late for any form of therapeutic interventions to be introduced. Information regarding the genetics of this disease on the African continent is limited, even more so in South Africa. This makes the task of identifying molecular markers, development of diagnostic and monitoring tools quite difficult. Therefore, the aim of this study was to identify the somatic mutation profiles of African patients with OSCC, thereby assisting in the expansion of knowledge regarding the genomic landscape as well fostering the development of molecular markers that could be useful in the diagnosis and treatment of this cancer. This was done by isolating DNA from blood, saliva and tumour samples and conducting whole exome sequencing (WES) on 21 matched blood/OSCC tumour samples. Data analysis was performed using R. The WES from 21 matched blood/tumour pairs revealed that somatic single nucleotide variants (SNV) were much more prevalent in comparison to somatic insertions or deletions (indels). The tumour mutation burden (TMB) was ~2 mutations per Mb. Tumour Protein 53 (*TP53*) was the most commonly mutated gene with 11 of the 12 mutations occurring in the DNA binding domain of the protein. Mutations in *TP53*, Titin (*TTN*) and Mucin 19 (*MUC19*) suggested that these genes were involved in relatively early events in the development of the tumours. Analysis of copy number alterations revealed a high degree of complexity in the tumour genome, with frequent amplification detected on chromosomes 1p33, 11q23.3 and 21q22.2, and common regions of deletion on chromosomes 5q31.2 and 7q32.3. Three mutational signatures were identified and the molecular pathway analysis showed that the *NOTCH, RTK-KAS*, and *TP53* pathways were the most significantly altered pathways. The alterations discovered in this study have contributed to the greater scheme of the molecular landscape of OSCC.

Acknowledgements

To my supervisor Prof Christopher Mathew, I am forever grateful for your patience, unwavering support and understanding. I do not take being granted the opportunity to conduct this project as part of my Master's degree lightly. I am thankful for the wealth of knowledge that you have imparted on me and for pushing me to be and do my absolute best. You are a phenomenal supervisor and role model.

To Prof Clement Penny, thank you for your caring personality and words of encouragement as my co-supervisor. Thank you for constantly availing yourself to attend to each and have query I had and doing it with so much kindness and grace.

A special thanks to Dr Carl Chen for your assistance with the technical aspect of this project and for facilitating all the training that needed to be done. A huge thank you to Mr Phelelani Mpangese for your collaboration in setting up the bioinformatics pipeline and additional support.

Thank you to Mrs Natalie Smyth and Ms Busisiwe Mthembu at the SBIMB Biobank and Dr Mishalan Moodley at Wits Medical School for the vocational education and training you all so easily facilitated. I'd also want to thank the project's funders, without whom this initiative would not be feasible and all the various stakeholders that were crucial in funding my MSc studies.

To my parents, Joseph and Taty, your support, sacrifices and prayers helped me make it through. Thank you for being the driving force, consistently encouraging me to be the best and to give my all always. Everything I am is because of you. To my siblings Jemima and Jennita, thank you for always bringing laughter into my life. Knowing that when things got tough, I had you both to always be there for me in whatever shape or form. To my late grandmother, Helene Kilola, the pursuit to further my studies was because of you. Even though you didn't understand the nature of my project, you were always curious and consistently asked questions about it. Your prayers and encouragement are what I carry with me always. Thank you for being my good luck charm.

To my colleagues at the Sydney Brenner Institute for Molecular Bioscience, thank you for your friendship and eagerness to lend a helping hand. It was comforting to know that one is not alone. To my close circle of friends, thank you for taking up an interest in my project and your constant support. I wouldn't have made it this far without any of you.

Funding

This study was fully funded by:

1. Newton/SAMRC/UCT grant to Professor Christopher Mathew for the genomic analysis of African Oesophageal cancer
2. SAMRC grant to Professor Paul Ruff for the WITS/MRC Common Epithelial Cancer Research Centre
3. National Research Foundation (NRF) grant

Table of Contents

Declaration i

Publications iii

Abstract iv

Acknowledgements v

Funding vi

Abbreviations and Symbols x

List of Figures xiii

List of Tables xv

1 Introduction 1

1.1 Risk factors in Sub-Saharan Africa: 3

1.1.1 Poor oral health 3

1.1.2 Alcohol and Tobacco 4

1.1.3 Consumption of hot beverages and high food temperatures 4

1.1.4 Socioeconomic status (SES) 4

1.1.5 Polycyclic aromatic hydrocarbons (PAH) 5

1.2 Molecular Genetics: 5

1.3 Genomics of OSCC tumours: 8

1.4 Treatment modalities in OSCC: 11

1.4.1 Epidermal Growth Factor Receptor: (EGFR) 11

1.4.2 Fibroblast Growth Factor Receptor (FGFR) 12

1.4.3 Phosphoinositide 3-kinase – α (PI3Kα) 12

1.4.4 WNT signalling pathway 12

1.5 Rationale: 12

1.6 Aim: 13

1.7 Study objectives 13

1.8 Preliminary work: 15

1.8.1 Patient recruitment 15
1.8.2 Sample collection 16
2 Methods 17
2.1 Ethical clearance: 17
2.2 Sample processing and storage: 17
2.2.1 Blood, Saliva and Tissue DNA isolation 18
2.2.2 Tissue DNA quality and yields assessment 19
2.2.3 Quality control of paired blood/OSCC tumour DNA samples 19
2.3 Whole Exome Sequencing: 20
2.3.1 WES protocol used by Novogene 20
2.3.2 Processing of DNA sequence data 20
2.4 Bioinformatic analyses: 22
3 Results 26
3.1 Clinical and demographic data and potential association with molecular data 26
3.2 Extraction of DNA from OSCC tumour tissue for sequencing by the Sanger Institute 27
3.3 Whole exome sequencing and bioinformatic analysis of 21 OSCC/blood tumour pairs 33
3.3.1 DNA sequence quality in tumour samples 33
3.3.2 Bioinformatic analysis of somatic mutations in OSCC tumour tissue: 37
4 Discussion 48
4.1 Purpose of study: 48
4.2 Clinical characteristics: 48
4.3 Tumour tissue quality control: 50
4.4 Quality and coverage of whole exome sequencing of the 21 blood/tumour pairs: .. 51
4.5 Detection of somatic mutations in the 21 OSCC blood/tumour pairs: 52
4.6 Project limitations 56
4.7 Conclusion and Future directions: 57

5 References 60

Appendices 68

WITS UNIVERSITY 71

STATEMENT OF CONSENT AND SIGNATURE 71

Abbreviations and Symbols

ADH1B	Alcohol Dehydrogenase 1B (class), Beta Polypeptide
AKT	Protein Kinase B
ALDH2	Aldehyde Dehydrogenase 2 Family member
ALSCR12	Amyotrophic Lateral Sclerosis Chromosome Region 12
ANKLE1	Ankyrin Repeat and LEM Domain Containing 1
APOBEC	Apolipoprotein B mRNA Editing Enzyme, Catalytic Polypetide-like
ATP1B2	ATPase, Na+/K+ Transporting, Beta 2 Polypeptide
ATXN3	Ataxin 3
BRCA2	Breast cancer gene 2
CASP8	Caspase 8
CCDC117	Coiled-coil Domain Containing 117
CCND1	Cyclin D1
CDK4/6	Cyclin dependent kinase 4/6
CDKN2A	Cyclin-dependent Kinase Inhibitor 2A
CHBAH	Chris Hani Baragwanath Academic Hospital
CHEK1/2	Checkpoint kinase 1/2
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
CLMN	Calmin
CNV	Copy Number Variant
COSMIC	Catalogue of Somatic Mutations in Cancer
CPCC	Cophenetic correlation coefficient
CROSS	Chemoradiotherapy and Radiotherapy by Operable Surgery for Squamous cell carcinoma
CSMD3	CUB and Sushi Multiple Domains 3
CYP2E1	Cytochrome P450 Family 2 Subfamily E member 1
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
DNAH5	Dyenin, Axonemal, Heavy Chain 5
DQA1	Major Histocompatibility Complex class II
DRB1	Dopamine Receptor Binding 1
EDTA	Ethylenediaminetetraacetic acid tube
EGFR	Epidermal growth factor receptor
ERG	ETS (E26 Transformation-specific)- related Gene
ESCA	Oesophageal Cancer
FAM120A	Family with Sequence Similarity 120 Member A
FAT1	FAT Atypical Cadherin 1
FBXW7	F-box and WD Repeat Domain Containing 7
FDR	False Discovery Rate
FGFR	Fibroblast Growth Factor Receptor
FOXP2	Forkhead box P2
GATK	Genome Analysis Toolkit
gDNA	Genomic DNA
GISTIC	Genomic Identification of Significant Targets in Cancer

GWAS	Genome-wide Association Studies
HIV	Human Immunodeficiency Virus
HLA-II	Major Histocompatibility Complex class II
HRCTI	Histidine Rich Carboxyl Terminus 1
ID	Identification
IGFBP-3	Insulin-like Growth Factor-binding Protein 3
IKSCC	Infiltrating Keratinizing Squamous Cell Carcinoma
INDEL	Insertion/Deletion
KEGG	Kyoto Encyclopedia of Genes and Genomes
KMT2A/C/D	Lysine Methyltransferase 2A/C/D
LRRTM2	Leucine-rich Repeat Transmembrane Neuronal 2
MAF	Mutation Annotation Format
MAPK	Mitogen-activated Protein Kinase
Mb	Mega-base
MDM2	MDM2 Proto-oncogene
MGMT	O-6-Methylguanine DNA Methyltransferase
MLH3	MutL homolog 3
MUC19	Mucin 19
MYC	Myelocytomatosis Viral Oncogene
MYO1B	Myosin 1B
NBPF1	Neuroblastoma Breakpoint Family, Member 1
NFE2L2	Nuclear factor, Erythroid 2-like 2
NMF	Non-negative Factorization Matrix
NOTCH	Notch Receptor
OAC	Oesophageal Adenocarcinoma
OBSCN	Obscurin, Cytoskeletal Calmodulin and Titin-interacting RhoGEF
OSCC	Oesophageal Squamous Cell Carcinoma
PAH	Polycyclic Aromatic Hydrocarbons
PCLO	Piccolo Presynaptic Cytomatrix Protein
PCR	Polymerase Chain Reaction
PDE4D	Phosphodiesterase 4D
PI3K/PIK3CA	Phosphatidylinositol-4,5-biphosphate 3-kinase catalytic subunit alpha
PLCE1	Phospholipase C Epsilon 1
PMS1	PMS1 Homolog 1, Mismatch Repair System Component
QC	Quality Control
RB1	Retinoblastoma 1
RHBDF2	Rhomboid 5 homolog 2
RNA	Ribonucleic acid
RTK-RAS	Receptor Tyrosine Kinase - Rat Sarcoma Acid Virus (Protein Kinase B)
RUNX1	RUNX Family Transcription Factor 1
SAMRC	South African Medical Research Council
SBIMB	Sydney Brenner Institite for Molecular Bioscience
SERPINA3	Serpin Peptidase Inhibitor, Clade A (Alpha-1 Antiproteinase, Antitrypsin), Member 3
SES	Socioeconomic Status
SNP	Single Nucleotide Polymorphism

SNV	Single Nucleotide Variant
SOP	Standard Operating Procedure
SPATA31D1	SPATA31 subfamily D member 1
SSA	Sub-Saharan Africa
SYNE1	Spectrin Repeat Containing, Nuclear Envelope 1
TBP	Tata Box Binding Protein
TCGA	The Cancer Genome Atlas
TMB	Tumour Mutation Burden
TMEM173	Transmembrane Protein 173
TMPRSS13	Transmembrane Protease, Serine 13
TP53	Tumour Protein 53
TTN	Titin
UBXN11	UBX Domain Protein 11
UCT	University of Cape Town
UKZN	University of Kwa-Zulu Natal
USP17L8	Ubiquitin Specific Peptidase 17-like Family Member 8
VAF	Variant Allele Frequency
VCF	Variant Call Format
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organisation
WNT	Wingless-type MMTV Integration Site Family
XBP1	X-box Binding Protein 1

List of Figures

Figure 1.1 Estimated cumulative incidence. 2
Figure 1.2 Age-standardised incidence and mortality rates for South Africa according to the National Cancer Registry. 3
Figure 1.3: Significantly dysregulated pathways in OSCC 9
Figure 1.4: Sample processing workflow. The green stars indicate preparation and analysis steps directly performed by the candidate.. 14
Figure 2.1: Novogene project workflow.. 20
Figure 2.2: Novogene bioinformatics analysis pipeline.. 21
Figure 2.3: Summary indicating Maftools 24
Figure 3.1: The average sequencing depth (bar plot) and the coverage (dot-line plot) present in each chromosome for the 6 retrospectively collected samples as received in the sequencing report from Novogene.. 36
Figure 3.2: The average sequencing depth (bar plot) and the coverage (dot-line plot) present in each chromosome for the 15 more recently collected samples as received in the sequencing report from Novogene. 36
Figure 3.3: A comparison of the TMB in TCGA datasets. 38
Figure 3.4: The 22 most frequently mutated genes in the OSCC tumours generated using R-Studio.. 39
Figure 3.5: The distribution of the SNVs in OSCC is displayed in the transition and transversion plot generated using R-Studio. 40
Figure 3.6: Distribution of TP53 mutations detected in this study in relation to p53 protein domains (TA = transactivation domain, DNA binding = DNA binding domain, Tet = tetramerisation domain) generated in R-Studio. 41
Figure 3.7: Variant allele frequency indicating the clonal status of the top mutated genes.... 41
Figure 3.8: GISTIC Plot of the 21 tumours identifying gains (red peaks) and losses (blue peaks) regions derived through using R-Studio. 42
Figure 3.9: NMF Rank for the OSCC sample set (generated using Maftools on R-Studio)... 43
Figure 3.10: Signature profile of the three established mutational signatures.. 44
Figure 3.11: APOBEC-enrichment analysis. Out of the 21 samples, 6 were enriched (N=6) and 15 were not enriched (N=15). 45
Figure 3.12: Molecular pathways (taken from the KEGG database) implicated in OSCC summarised as bar plots.. 46

List of Tables

Table 1.1 SNPs associated with OSCC risk according to GWAS 7
Table 3.1: Tumour DNA and RNA concentration and absorbance measured with Nanodrop 28
Table 3.2: Tumour DNA yields measured with the Qubit 29
Table 3.3: DNA yields and Genotype match from the Sanger Institute 30
Table 3.4: Sequencing depth and coverage of target regions for the 21 samples 34
Table 3.5: The total number of somatic SNVs and indels for the 21 OSCC tumour samples ..... **Error! Bookmark not defined.**
Table 3.6: Genes mutated in molecular pathways implicated in OSCC 47
Table 3.7: Demographic and clinical data from OSCC patients **Error! Bookmark not defined.**

1 Introduction

Oesophageal cancer describes a group of diseases that are primarily derived from the epithelial, neuroendocrine, lymphoid or mesenchymal tissues (1). It is considered as one of the deadliest forms of human malignancy and is the 6th most common fatal cancer worldwide (2) consisting of two main histological types: oesophageal adenocarcinoma (OAC) and oesophageal squamous cell carcinoma (OSCC) (1). The development of OAC stems from the glandular columnar tissue (3) due to the occurrence of the columnar metaplasia of the distal oesophagus which is a consequence of prolonged chronic gastro-oesophageal reflux disease as well as Barrett's oesophagus (4). On the other hand, OSCC originates from the multi-layered epithelium comprised of the innermost lining of the oesophagus, with the occurrence of dysplasia as a precursor (3). Amongst the two subtypes of oesophageal cancer, 85% of cases are of the OSCC subtype (5). Additionally, OAC has mostly been attributed to smoking and obesity, with OSCC linked to alcohol and cigarette smoking (6). According to the Globocan Oesophagus data sheet (2020), oesophageal cancer accounts for approximately 3.1% of new cases of cancer worldwide with the global cumulative risk of incidence (Figure 1.1) (7) highlighting that the rates of incidence vary between different regions in the world. Taking the variation in the incidence into account, the literature is indicative of the prevalence of OSCC being predominant in developing countries, with a high incidence occurring in eastern Asian, southern Europe, southern South America as well as eastern and southern Africa, whilst OAC is more dominant in developed regions such as Europe and North America (8,9).

OSCC tends to develop without the presence of clinically visible symptoms until it has reached an advanced stage giving the disease a poor prognosis with a 5-year survival rate of 15-30% (10), (11). It is found to be more common in men (mostly aged 50-70 years) than it is women, with the ratio between the two genders varying greatly and can be anywhere between 1:1 or 4:1 depending on the country of interest. The aetiology of OSCC can be defined as multifactorial, in that, it is a combination of environmental exposures, genetic traits and lifestyle factors and is also relative to the population of interest, meaning that certain aspects will only be seen in some populations and not others. A review of the development of OSCC in Asia, more specifically in the Chinese population, found that the risk of developing the disease was mainly attributable to a family history of cancer, drinking well and surface water, as well as a salty diet (11). The same cannot be said for African populations, for which studies are currently focused on Sub-Saharan Africa (SSA). The regions of interest in relation to OSCC

include Ethiopia, Rwanda, Burundi, Malawi, Kenya, Uganda, Tanzania, Zimbabwe, Zambia, Madagascar and South Africa (12).

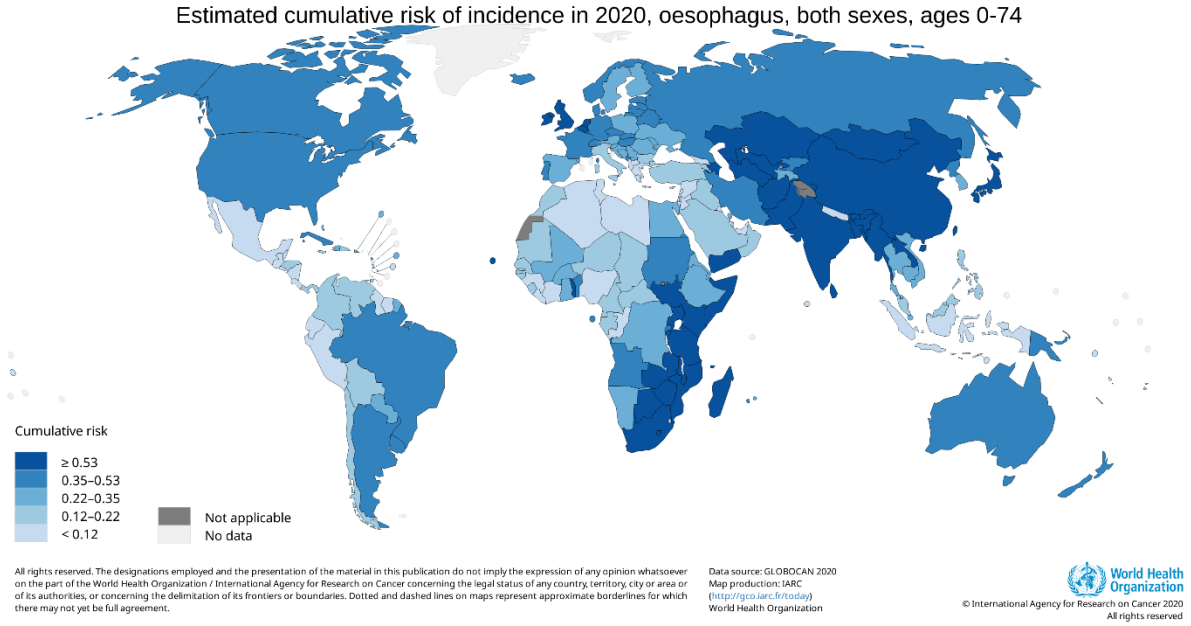

Figure 1.1 Estimated cumulative incidence (7). South eastern Asia, eastern and southern Africa display a high cumulative incidence of oesophageal cancer and that is where the rates of OSCC tend to be the highest.

In South Africa, Oesophageal cancer ranks 7th when it comes to newly diagnosed cancer cases and is responsible for the 6th highest cancer-related deaths in the country, with OSCC being the most prevalent type at an incidence rate of 6.4/cases/100,000 overall (13). When assessed according to sex, the incidence rate is 8.6/cases/100,000 for males and 4.7/cases/100,000 for females (13).

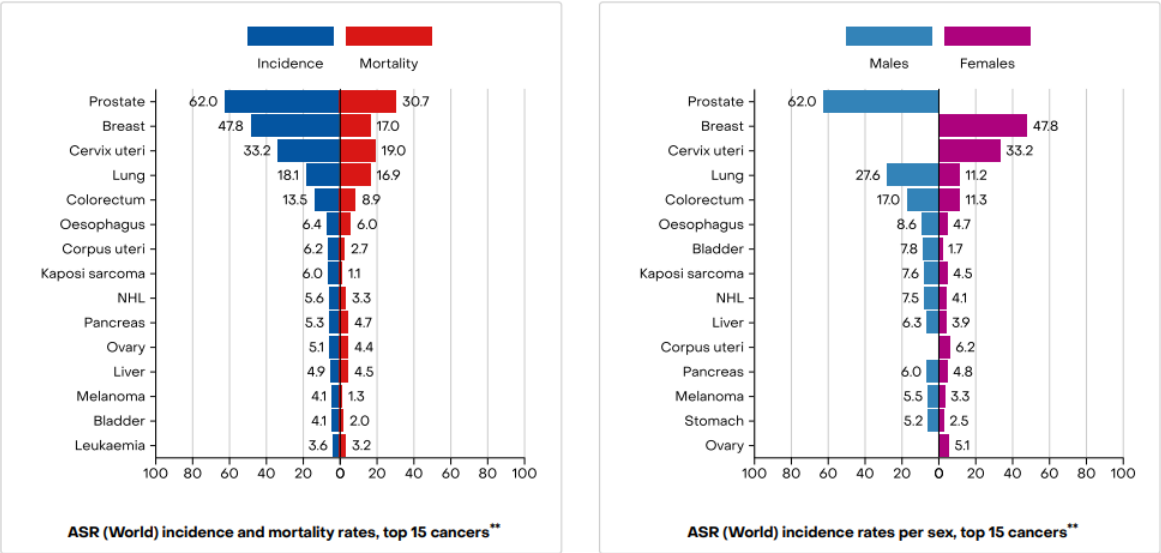

Figure 1.2 Age-standardised incidence and mortality rates for South Africa according to the National Cancer Registry (13).

1.1 Risk factors in Sub-Saharan Africa:

In Sub-Saharan Africa (SSA), OSCC tends to develop in younger patients in comparison to other regions in the world (12). According to a systematic review by R. Kachala published in 2010, the average gender ratio was reported to be 2:1 male to female in SSA (14). However a study conducted by Ferndale *et al.* (2020) on the gender differences in OSCC in South Africa found the ratio to be 1.4:1 male to female (15). OSCC is often diagnosed at a late stage due to its asymptomatic nature, making it late for any form of therapeutic interventions (other than palliative care) to be introduced (12) thereby maintaining a poor prognosis and a low chance of survival. The symptoms that are identifiable at the late stage diagnosis include odynophagia, dysphagia, weight loss and in some instances haematemesis (16). Possible risk factors that have been attributed to the development of OSCC in the African context include: poor oral health, consumption of alcohol and hot beverages, high food temperatures, lower socioeconomic status nutritional deficiencies, consumption of *Fusarium*- or *Aflatoxin*-contaminated maize and exposure to polycyclic aromatic hydrocarbons (PAH) (17,18).

1.1.1 Poor oral health

The mechanisms that are thought to be behind the association of poor oral health and the development of OSCC include the possibility that an inflammatory response as a consequence of periodontal disease together with poor oral hygiene can result in the development of

secondary metabolites such as acetaldehyde and reactive nitrite, which are known carcinogens that are linked to OSCC (19).

1.1.2 Alcohol and Tobacco

A meta-analysis into tobacco smoking in African populations established that partaking in smoking increased the risk of OSCC development (20). To further corroborate this, a study performed in South Africa found that the use of tobacco, regardless of the form it is in, increased the risk of developing oesophageal cancer by at least four times (21). Regarding the role of alcohol, a case-control study performed in Kenya focusing on endoscopy-confirmed cases revealed that the consumption of the traditional beers, busaa and chang'aa, was associated with half of the cases of OSCC that were seen in that area as a result of the carcinogenic substances (high concentrations of ethanol or acetaldehyde) that were found and are attributable to the methods used to prepare these beverages (22).

1.1.3 Consumption of hot beverages and high food temperatures

These two factors are associated with the development of OSCC by means of their potential to cause a thermal irritation to the epithelium of the oesophagus (20,23). This in turn results in the production of inflammatory heat shock proteins, cytokines and chemokines that exacerbate carcinogenesis (22). The thermal injury can also result in the formation of nitrosamines which are known carcinogens (22).

1.1.4 Socioeconomic status (SES)

Factors such as income, education and employment are all widely accepted indices of socioeconomic status (24). SES is multidimensional in nature, and the attempt at interpreting this construct is rather difficult. A study conducted in Taiwan linked the risk of OSCC to wealth and occupation type (24). Various case-control studies have attributed low SES to the risk of OSCC development in African populations. One study in South Africa associated lower salaries with OSCC risk, whereas a study in Zimbabwe suggested that low occupational statuses and mining as a profession placed individuals at a greater risk (20). Another study done in Kenya found that substandard housing increased the incidence of OSCC (20). The proposed reasoning behind this is that an individual within a lower social class does not have the financial capacity to do regular health check-ups, most likely has poor living conditions and in most cases is employed in a sector that could pose great risks to their health (surface and underground mining). As a result of these environmental exposures, their disease risk is influenced.

1.1.5 Polycyclic aromatic hydrocarbons (PAH)

The formation of a PAH is as a result of the incomplete combustion of organic compounds, releasing noxious and carcinogenic compounds as a consequence (22,25). The majority of African populations are considered to be in low-income settings, and are thus, at a significantly greater risk of being exposed to PAHs through certain practices that they partake in. Some of the practices include the burning of forests or crops, making use of cooking and heating methods that involve burning biomass fuels such as wood, coal or dung, as well consuming certain foods and beverages, including mursik (19,20,22). In a study performed in Kenya, it was found that high concentrations of metabolites (specifically 2-hydroxy-napthalene) were associated with the presence of advanced oesophageal dysplasia which is believed to increase the risk of developing OSCC by 28 times (25).

The risk factors mentioned above describe the possibility of developing OSCC from an environmental and lifestyle perspective. However, further studies are warranted in Sub-Saharan Africa given that the factors already mentioned tend to be population-specific and cannot be generalized throughout in order to be able to fully define the risk factors for this disease.

1.2 Molecular Genetics:

From a molecular perspective, it is thought that mutations and epigenetic changes are responsible for the development of the tumours seen in OSCC, with the genetic basis being both inherited (germline) and at a somatic level (12).

From a heritability perspective, rare disorders such as Tylosis and Fanconi anaemia influence the risk of developing OSCC (12). Tylosis is a rare disorder characterised by the thickening of the skin on the palms and soles of the feet, known as palmoplantar keratoderma. The version of this disease (Type A Tylosis) that is linked to the risk of developing oesophageal squamous cell oesophageal cancer is high. It is characterised by a young age of onset and a penetrance estimation of as high as 95% by the time an individual reaches the age of 65 (11). Type A Tylosis results from autosomal dominant mutations in the *RHBDF2* gene on chromosome *17q25* (19). A gain-of-function mutation results in the sustained signalling of the epidermal growth factor receptor (EGFR) within the cells leading to a hyper-proliferative epithelium as well as the dysregulation of wound healing that is observed in OSCC (11).

Fanconi anaemia is a rare, mostly autosomal recessive genetic disease (26) which is characterised by bone marrow failure, susceptibility to cancer and cellular hypersensitivity to DNA cross-linking agents (26,27). It has been established for >20 years that the risk of OSCC in patients with Fanconi anaemia (i.e. homozygotes) is more than 1000x higher than in the general population (28). A study done on Turk and Chinese men found that the occurrence of heterozygous insertion/deletion mutations, as well as homozygous missense mutations in the Fanconi anaemia pathway genes were responsible for the increased risk of OSCC development in those populations (26,29).

Genetic association studies, in which genetic variants are genotyped in a number of OSCC cases and controls, have been used to search for genes which are associated with susceptibility to OSCC. Candidate gene studies in African populations have shown that single nucleotide polymorphisms (SNPs) in several genes are associated with the risk of disease development. These genes include: *ALDH2* (ALDH2*1/*2), *CASP8* rs1045485, *CHEK2* – rs482293 and rs1033667, *CYP2E1* (7632T>A), *MGMT* rs12917, *MLH3* rs26279, *PMS1* rs5742938 and *RUNX1* rs2014300 (12). Their functions range from being involved in the metabolism of alcohol, to DNA repair in order to maintain the integrity of the genome (12). The largest candidate gene study in Africa conducted by Chen *et al.* (2019) investigated seven susceptibility loci in the South African black populations located in the Western Cape and Gauteng provinces. The cohort consisted of 1471 cases and 1791 controls (17). Out of the seven loci that were investigated, only one SNP (rs1033667*)* in the *CHEK2* gene yielded an association with OSCC (17). The key role of this serine-threonine kinase is in responding to DNA damage by means of leading the activation of DNA repair, cell-cycle arrest and senescence (17). The fact that variants from the other six loci showed no significant association with OSCC underlined the need to expand studies of genetic risk factors in the South African population.

Regarding more broadly based germline studies, high-throughput genotyping has allowed for the analysis of SNPs to identify associations between common genetic variants and risk of disease development (in this instance OSCC), with the undertaking of this process known as genome-wide association studies (GWAS) (19). GWAS has greatly influenced the manner in which information about the genetics of complex diseases has come about by providing an understanding of the architecture of disease susceptibility through the identification of the genes that cause disease as well as the mechanisms behind it. This knowledge can provide leads

that assist with the clinical care of those diagnosed with disease and assist in tailoring the approaches that guide the prevention and treatment thereof (30).

A substantial number of GWAS have been carried out for OSCC, particularly in Chinese populations, which have identified multiple loci associated with OSCC risk. A pooled analysis of three GWAS identified rs7447927 at *5q31.2* located in *TMEM173* and rs1642764 at *17p13.1* located in *ATP1B2* (telomeric to *TP53*) in both high- and low-risk incidence regions (31). In addition to these loci, another SNP, rs35597309 situated in the *HLA* Class II gene region between *HLA-DRB1* and *HLA-DQA1* was also identified, however it was mostly in high-incidence regions (31). The table below illustrates SNPs that have been identified as increasing the risk of development of OSCC based on studies performed in East Asia.

Table 1.1: SNPs associated with OSCC risk according to GWAS done in Eastern Asia (19,31)

Loci	**Nearest gene**	**rs-ID**
10q23.33	***PLCE1, KIAA1516***	rs2274223
2q33.1	***CASP8, ALSCR12***	rs13016963
21q22.12	***RUNX1***	rs2014300
22q12.1	***CHEK2, CCDC117, XBP1***	rs2239815
5q11.2	***PDE4D, PDE4DN2***	rs10052657
5q31.2	***TMEM173***	rs7447927
17p13.1	***ATP1B2, TP53***	rs1642764
6p21.32	***HLA class II genes***	rs35597309
12q24.1	***ALDH2***	rs671
4q23	***ADH1B***	

The first GWAS to be conducted in an African population was published recently [Chen *et al.* (2023)] which detected associations at *FAM120A, MYO1B* and *CHEK2* in the black South African population. Interestingly, in this study a meta-analysis of the African GWAS data with a Chinese GWAS identified both shared and distinct risk loci between the two populations (32).

It is crucial to note that GWAS does not generally identify the causal variant at a particular locus, but follow up fine-mapping and functional analysis can be used to investigate the genes and variants that are driving the association of the locus. It is likely that both genetic and environmental risk factors may be shared or different between populations, so this must be taken into consideration when investigating genetic risk factors in OSCC.

1.3 Genomics of OSCC tumours:

With regards to the detection of somatic mutations in OSCC, the approaches that are commonly employed include whole exome sequencing (WES) and whole genome sequencing (WGS).

Whole exome sequencing is a reasonably affordable method that allows for the detection of sub-clonal mutations as well as performing sample profiling in instances where there is low tumour content (33). It is a targeted approach, in the sense that the area of focus is on the part of the genome that codes for proteins and possesses robust analysis tools and databases. Despite the advantages that this approach has, it also has a few disadvantages. Firstly, there is the potential for the exome enrichment kits to introduce a variety of artefacts and biases that can impair copy number calling, despite newer technologies being introduced to reduce these difficulties, thereby resulting in higher false-positive rates (33). Secondly, if there are any chromosomal breakpoints that are outside of the exonic areas, WES may overlook the chromosomal rearrangements that result in fusion genes (33). Lastly, it tends to have a low sensitivity for structural variations (34,35).

On the other hand, whole genome sequencing offers a more comprehensive view as it can detect variations across the genome including non-coding regions (36). It is, however more costly than WES and generates larger datasets which then results in extended turnaround periods to obtain the results (33). WGS is more effective in detecting large-scale structural variants such as copy number variants (CNVs) and chromosomal rearrangements and, apart from being able to detect variants, it can also be used to analyse gene expression, epigenetic modifications and perform GWAS (36). Both of these approaches are competent in detecting single base mutations, indels and large copy number variants, making the use of either appropriate when performing OSCC studies.

Many whole genome sequencing and whole exome sequencing studies of matched OSCC blood/tumour pairs have been done in European and Asian populations to identify somatic mutations that may be driving tumour development. Ohashi *et al.* (2015) reviewed four major studies conducted in relation to genes associated with the development of OSCC in China and found that the most frequently mutated gene throughout the studies was *TP53* at a rate of between 59-93% (37). Simultaneously, a review by Yang *et al.* (2020) included the review by Ohashi and then established that mutations in genes that were involved in either the regulation of the cell cycle such as *CDKN2A, RB1, NFE2L2, CHEK1 and CHEK2* or of cell differentiation

such as *NOTCH1 and NOTCH3* were also affected at frequencies ranging from 2-10% (23) with significant amplification. These genes included *CCND1* (46.4%), *CDK4-CDK6* (23.6%) *and MDM2* (5.7%) (23). Apart from the regulation of the cell cycle, other genes that were altered in OSCC were grouped into different signalling pathways, including *EGFR*, a trans-membrane glycoprotein involved in the control of cellular proliferation was found to be overexpressed in OSCC. This also resulted in the amplification of the *RAS* and *AKT* pathways (23,38). A more recent review by Zhang *et al.* (2022) alluded to the occurrence of genomic alterations in *TP53, KMT2C/D, NFE2L2, NOTCH1, PIK3CA, MYC, CDKN2A/B, CCND1, FGF3/4/19* and *EGFR* being involved in the development of OSCC (39). A schematic taken from this review illustrates some of the significantly dysregulated pathways and their associated genes (Figure 1.3).

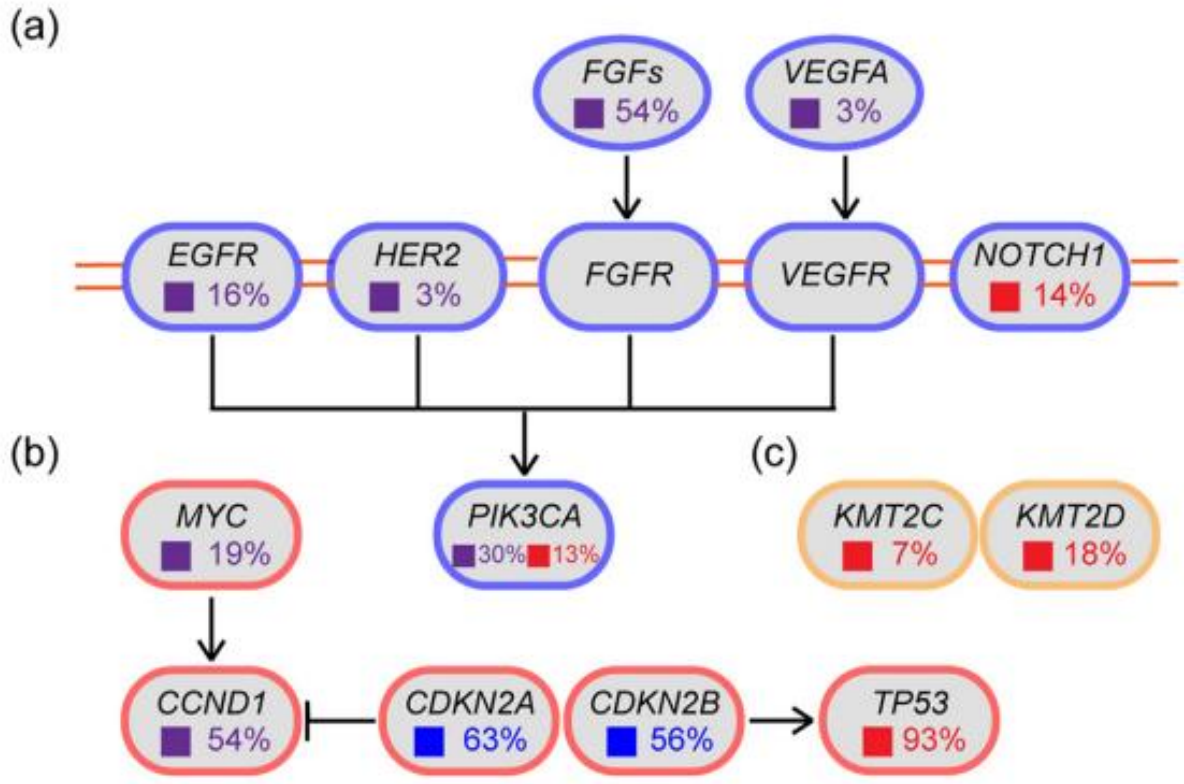

Figure 1.3: Significantly dysregulated pathways in OSCC. (a) RTK-MAPK-PI3K signalling, (b) Cell cycle regulation, (c) Epigenetic modulation (39). [Permission obtained from publisher to reproduce]

Very few studies have been conducted on African populations in relation to the identification of somatic mutations that are associated with the development of OSCC due to inadequate infrastructure and funding needed to conduct the research. The lack of information pertaining

to the detailed systematic review validations of the available data has created a significant gap in the knowledge of the disease along the OSCC corridor.

A recent comprehensive review into the research on the genetics of OSCC in African populations found that only 23 studies were performed, for which the vast majority were candidate gene studies and only one study [Liu *et al.* (2016)] carried out in Malawi, made use of WES (40). In the 59 OSCC patients that were studied, somatic mutations were detected in *TP53, CDKN2A, NFE2L2, CHEK2, NOTCH1, FAT1* and *FBXW7* (16). The authors suggested that the driver gene activation and tumour suppressor gene loss patterns were similar to those reported in Asian and North American patients. They also carried out an analysis of mutation signatures. These signatures are the product of different mutational processes such as defective DNA repair or exposure to different mutagens which produce specific combinations of mutation types (41). The Malawian study revealed a novel mutation signature that was similar to one seen in oral squamous cell carcinoma from individuals known to chew tobacco (16).

A more recent study conducted by Van Loon *et al.* (2023) sought to assess the somatic mutation rate, mutational patterns, copy number profiles and frequently mutated genes in the tumour tissue of 61 Tanzanian OSCC patients (40). The authors also compared their results with those reported by Liu *et al.* (2016) in an attempt to find the commonalities and distinctions between the two representative sites and liken them to the molecular features of OSCC in the cancer genome atlas (TCGA), as well as assess any potential heterogeneity of the disease across eastern Africa (40). Approximately nine genes were found to be significantly mutated (*TP53, CDKN2A, HRCTI, ANKLE1, UBXN11, ATXN3, TMPRSS13* and *TBP*) (40). *TP53* was the most commonly mutated gene in the Tanzania cohort, as was the cases in the Malawian cohort and the Japanese and Chinese cohorts. There was a subset of tumours that did not have mutations in *TP53* (40). The most common mutation signatures found in the Tanzanian cohort were signature 5 (associated with smoking and age), signature 8 (unknown aetiology) and signature 9 (associated with lymphoid malignancies), and the cytidine deaminase activity (APOBEC) pathway (signature 13), but no particular mutagenic insult (40). Whilst a unique signature was found (signature 29) in a small subset of the Malawian tumours, the Tanzanian cohort exhibited no such signature (40). Furthermore, signature 17 was previously noted as common in the Malawian cohort, yet it was also not found in the Tanzanian tumours, with the reasoning behind these differences believed to be as a result of technical variations between input data of the two sources (40).

The genomics of OSCC tumours in different populations has illustrated how variable the molecular characteristics of the disease can be. Despite similar genes being mutated in tumours from Asian and African populations, a one-size-fits-all approach cannot be applied, as patients in different regions may have different risk exposures and there may even be differences in similar regions.

1.4 Treatment modalities in OSCC:

According to the World Health Organization (WHO), surgery, chemoradiotherapry, immunotherapy and chemotherapy are the recommended treatment modalities for OSCC depending on the stage of the disease (42).

Surgery is often the preferred treatment for early-stage OSCC, however, as result of the asymptomatic nature of the disease, patients are at a more advanced stage when diagnosed and thus, surgery tends to have no valuable clinical benefit. Chemoradiotherapy combines the effects of chemotherapy and radiation therapy through the CROSS regimen, which includes incorporating carboplatin and paclitaxel with concurrent radiotherapy (43). Recently, modalities such as targeted drug therapy and immunotherapy are increasingly being investigated and applied. Targeted drug therapy makes use of the patient's tumour genotype in order to provide treatment with specific combinations of chemotherapeutics, whereas, immunotherapy stimulates the immune system by improving the ability of the immune system to identify and destroy the cancer cells (44). The process is facilitated by the ability of the immune cells to identify certain antigens that are linked to cancer on the surface of the tumour, which in turn triggers the immune system (44).

On the basis of OSCC studies performed in other regions, a number of potential therapeutic targets that warrant further investigation have been identified. These include tyrosine kinases (*EGFR, FGFR, PI3K/AKT*), as well as the *WNT* pathway (19).

1.4.1 Epidermal Growth Factor Receptor: (EGFR)

Since the overexpression and amplification of *EGFR* are common occurrences in OSCC and are strongly linked to tumour invasion and progression, it is a candidate target for therapeutic interventions. Gefitnib – a selective inhibitor – has been found to increase the effectiveness of tumour necrosis factor-related apoptosis inducing ligands in OSCC cells by inhibiting the intracellular phosphorylation of *EGFR* (39). Lapatinib – a tyrosine kinase inhibitor- has been found to inhibit the growth of OSCC patient-derived xenograft when used in conjunction with

5-fluorouracil by also inhibiting the intracellular phosphorylation of *EGFR* (45) whilst Nimotuzumab has the ability to promote radiosensitivity in *EGFR* overexpressing OSCC cells by upregulating *IGFBP-3* (39). It achieves this by binding to *EGFR* and blocks the binding of its natural ligands which then disrupts EGFR signalling (46).

1.4.2 Fibroblast Growth Factor Receptor (FGFR)

H3B-6527, a selective inhibitor for *FGFR4* has been found to drastically hinder the growth of OSCC cells by restricting the *PI3K* and *MAPK* signalling pathways (39). It also inhibits FGFR signalling by blocking *FGFR4* kinase activity, thereby preventing phosphorylation and activation of the downstream targets of *FGFR* (47). The inhibitor AZD4245 has also been shown to increase the sensitivity of OSCC cells that possess a high level of active *FGFR1* to Gefitnib thereby proposing a form of combination therapy for OSCC (39). AZD4245 indirectly inhibits *FGFR* signalling by blocking the activation of PI3K-β-mediated AKT which then reduces FGFR-dependent phosphorylation and activation of FGFR substrate 2 (48).

1.4.3 Phosphoinositide 3-kinase – α (PI3Kα)

CYH33, the highly selective PI3Kα inhibitor, has displayed some significant activity in reducing the proliferation of OSCC cells, as well as the growth of xenografts (39). It enhances the sensitivity of OSCC cells to radiation by rejecting the survival signals in the tumour cells and its resultant microenvironment (39).

1.4.4 WNT signalling pathway

Porcupine (PORCN) is a key enzyme in the WNT pathway. It is responsible for the palmitoylation of WNT proteins, thereby enhancing its secretion and also activating signalling of WNT/β (49). A porcupine inhibitor, LGK974, is known to inhibit the secretion of WNT and its signal by disabling the palmitoylation thereof (39,49).

Most of the inhibitors mentioned above are undergoing clinical trials in order to establish their safety and efficacy as useful therapies for the treatment of OSCC. They are pointers for the direction in which the search for treatment therapies to alleviate OSCC is headed.

1.5 Rationale:

Overall, there is limited information on the genetics of OSCC on the African continent, making the task of developing diagnostic and monitoring tools, as well identifying molecular markers for this disease rather difficult. In the context of South African patients, information on the

status of somatic mutations in the context of driver genes and their accompanying mutations is sparse.

Developing more knowledge in these aspects will assist in establishing the genomic landscape of OSCC in South Africa, as well as foster the development of molecular markers that might be able to give an indication for the association of lifestyle risk factors and clinical features with facets such as prognostic value and potential candidates for targeted therapies.

1.6 Aim:

The aim of this study was to identify the somatic mutation profiles of African patients with OSCC using whole exome sequencing, and ultimately establish clinically useful molecular markers for this disease and possibly identify treatment types that may be effective in African populations with OSCC.

1.7 Study objectives

The main objectives of this study could be separated into two components:

A wet-lab component that consisted of:

- Extracting DNA from blood, saliva and tumour tissue samples of patients that were recruited with histologically confirmed oesophageal squamous cell carcinoma. DNA from these samples was sent to the Wellcome Sanger Institute for WES as part of a larger collaborative project with Professor Parker at the University of Cape Town and the Wellcome Sanger Institute, Cambridge UK.
- Ensuring that a thorough quality control process was undertaken to assess the integrity and yields of the DNA.

An analytical component that was dedicated to:

- Determining the occurrence of somatic mutations in the tumour tissue of DNA samples from matched blood/OSCC tumour pairs that were submitted for whole exome sequencing.
- Identifying the prevalent driver genes and mutations, including copy number variants that are involved in the development of this form of cancer in a representative sample of the South African population.

- Establishing the molecular pathways most commonly involved in disease aetiology.
- Using the genomic profile data to assess whether therapies for OSCC which are effective in other populations may also be applicable in African populations.

In order to achieve the aim and objectives of this study, it was split into two discrete projects. Project 1, which formed part of the collaboration with Professor Parker at the University of Cape Town and the Wellcome Sanger Institute in Cambridge, UK referred to above, and project 2 which constituted DNA from 21 OSCC/blood paired samples previously extracted by Mr Mishalan Moodley, a colleague in the Somatic Cell Genetics Unit, Department of Haematology and Molecular Medicine at the University of the Witwatersrand. A summarised schematic showcasing the processes that would be involved in the attempt to achieve the aims and objectives of this study are illustrated below (Figure 1.4).

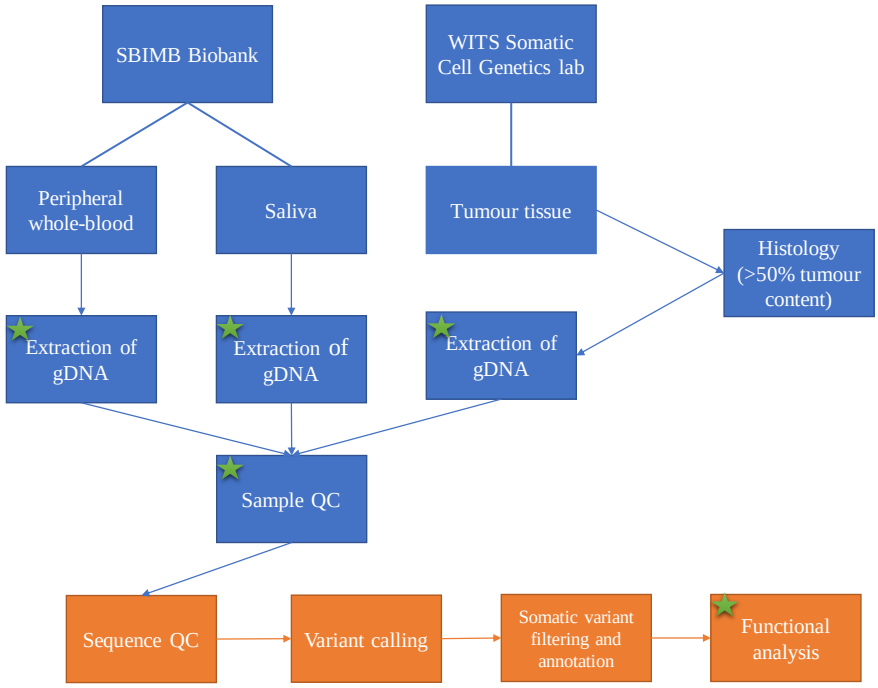

Figure 1.4: Sample processing workflow. The green stars indicate preparation and analysis steps directly performed by the candidate. These included extraction of genomic DNA (gDNA) from the various sample types, as well as enforcing quality control measures and conducting the functional analysis.

1.8 Preliminary work:

Considering that this particular study was a subset of a larger study, certain components were not directly performed by the candidate and are described below.

1.8.1 Patient recruitment

The patient samples used to conduct this study consisted of both retrospectively and prospectively recruited patients.

In terms of the retrospective patients, matched blood/tumour DNA samples were derived from six OSCC patients. Three of the patients were from Charlotte Maxeke Johannesburg Academic Hospital, Gauteng and the other three from Grey's Hospital, Kwa-Zulu Natal. The samples were already submitted for the whole exome sequencing (WES) at Novogene, Cambridge, United Kingdom prior to the start of this study and the data was available on the WITS Bioinformatics Cluster Server.

The 15 prospective patients consisted of potential OSCC patients of female gender arriving for an endoscopy at Grey's Hospital in Pietermaritzburg, Kwa-Zulu Natal, and were recruited by Dr Lucien Ferndale, Head of the Department of Gastro-intestinal Surgery. The inclusion criteria consisted of each patient being at least 18 years of age, presenting with a histologically confirmed diagnosis of OSCC and having no history of alcohol or tobacco use in order to be enrolled into the study. The motivation for restricting this particular cohort to non-smoking, non-drinking females was to investigate the molecular basis for OSCC in participants not proposed to traditional major risk factors for OSCC. Written informed consent was obtained by means of providing patients with a patient information sheet and consent form (Appendix 1). Participants who agreed to take part in the study were then interviewed to obtain information for a questionnaire (see below and Appendix 2). In addition to the patients recruited in Kwa-Zulu Natal, patient enrolment and consenting was done at two other sites in the Gauteng province. This was facilitated by clinicians at Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and Chris Hani Baragwanath Hospital (CHBAH) and carried out by a trained research assistant, Ms Zanele Ndlovu. Similar data and consent were also previously obtained from the retrospective patient samples.

1.8.1.1 Questionnaire

The study questionnaire established the type of cancer with which the patient was diagnosed, in this instance OSCC. It included demographic information such as the ages of the participants, their ethnicity, gender, marital status, income status, the level of formal education

attained and employment status. It also included what type of fuel was used in cooking processes as well as to keep warm during the colder months and the duration for which fuel was used (for the past 20 years). The questionnaire also took into account exposure to the potential risk factors, such as smoking and alcohol consumption, prior to becoming ill. Any medical history that the participants were able to provide was also noted. This included any history of being diagnosed with other forms of cancer, HIV status, and any other co-morbidities. The role that the participants' occupation contributed to their health status was also included. Occupational questions sought to establish what type of setting a participant worked in, whether it was a mine or factory or in an industry that would make it likely for them to be exposed to some of the proposed risk factors for the development of OSCC. All of the abovementioned data was subsequently loaded onto a RedCap database so as to safeguard and facilitate better management of the information.

1.8.2 Sample collection

Whole-blood samples were collected in BD Vacutainer Ethylenediaminetetraacetic acid tubes (Becton, Dickinson and Company, USA). Two 3 mL vials of the peripheral whole-blood samples were collected. The saliva samples were collected using the Oragene OG-500 (DNA Genotek Inc., Canada) collection kit and the tumour biopsy samples were placed in 1 mL of RNALater.

The blood, saliva and tumour biopsies were transported to the Sydney Brenner Institute for Molecular Biosciences (SBIMB) Biobank for processing and storage.

2 Methods

This following chapter will describe the methods that were used to carry out the aims and objectives of the study. It includes the ethical clearance that was obtained, the processing and storage of samples, DNA isolation from the various sample types as well as the quality control (QC) measures taken to assess the yield and quality of the DNA performed by the candidate. Also included in this chapter is the whole exome sequencing process as well as the analysis of the resultant data using various bioinformatic tools.

2.1 Ethical clearance:

For this study, a sub-study ethics application was made to the Human Research Ethics Committee (Medical) of the University of the Witwatersrand and was granted in October 2021 with the certificate number: **M2111143**, with a validity of five years, see Appendix 3. The larger study for which this study falls under is entitled "*The investigation of genetic biomarkers for African oesophageal cancer*" and the certificate number accompanying this clearance is **M180306**, which was renewed in 2023 under **M2310111**, see Appendix 4. Recruitment of OSCC patients from Grey's Hospital in Pietermaritzburg was approved by the Biomedical Research Ethics Committee of the University of Kwa-Zulu Natal (UKZN), Durban, South Africa (Certificate number: **BF270/15**).

2.2 Sample processing and storage:

For the processing of the blood samples, a SBIMB standard operating procedure (SOP) was done as follows: the EDTA tubes were placed in the swing bucket centrifuge and spun at 1,900 x g for 10 minutes at room temperature. For each EDTA tube, there were two cryovials, one for the plasma and another for the buffy coat. The plasma was collected and then placed in a hand-labelled cryovial for further centrifugation. After collecting the plasma, the buffy coat was collected and placed in its designated cryovial and then stored. The plasma-containing cryovial was then centrifuged at 10,000 x g for 10 minutes and a small red pellet would be visible at the completion of this process. The supernatant was then removed and placed in the final cryovial labelled for the plasma samples, making sure to not touch the pellet in the process. The blood samples were then stored in a -80 °C freezer, with the saliva samples being stored at 5 °C and the tumour biopsies stored at -20 °C until DNA isolation.

2.2.1 Blood, Saliva and Tissue DNA isolation

DNA isolation from the blood samples was done using the QIAGen FlexiGene® DNA kit (QIAGen, USA) and a modified protocol described by Chen *et al.* (2018) was followed (50). The procedure involved the lysis of the cells (done using buffer FG1) followed by centrifugation and resuspension in a denaturation buffer that contained protease (buffer FG2/QIAGEN Protease). The DNA was purified by means of incorporating ethanol precipitation and then suspending it in buffer FG3 (10mM TRIS-Cl solution at pH 8.5).

DNA isolation from the saliva samples was performed using the DNA Genotek's PrepIT L2P protocol (DNA Genotek, USA). Prior to DNA isolation, the samples were incubated in an air incubator at 50°C overnight. The entire samples' content was then transferred to 15 mL centrifuge tubes. This was followed by adding 40 μL of prepIT·L2P to the samples, in order to lyse the cells and then incubating the samples at room temperature. The samples were centrifuged at 3,500 x g for 10 minutes and the resultant clear supernatant was transferred to a new 15 mL centrifuge tube. Approximately 1.2 ml of 95% to 100% ethanol was added and mixed through inversion for 10 minutes. The samples were allowed to stand at room temperature for 10 minutes and then centrifuged in the same manner (3,500 x g) and the supernatant was carefully discarded without disturbing the visible DNA pellet. 1 mL of 70% ethanol was then added to wash the pellet and subsequently discarded (without disturbing the precipitated DNA). Between 0.2 – 1 mL of TE buffer was added to the sample, vortexed for 30 seconds and left to incubate overnight. After incubation, the rehydrated DNA was transferred to a 1.5 mL cryovial.

Selection of the tumour samples for DNA extraction was based on a diagnosis of oesophageal squamous cell carcinoma confirmed by histopathology. Tumour tissue biopsies were cut into two equal parts with one part used for histology to determine the % tumour cell content of the sample and the other for DNA extraction (51). Only samples with >50% tumour content were used for DNA sequencing. The work of selecting the tumour samples was carried out by Professor Wadee and colleagues in the Department of Anatomical Pathology of the National Health Laboratory Service. Nucleic acid extraction from tumour biopsies was done using the QIAGen AllPrep ® Universal kit (QIAGen, USA) following the protocol instructions. This procedure consisted of lysing and homogenizing the tissue in a high denaturation guanidineisothiocynate-containing buffer which instantly inactivated the DNases and RNases to facilitate extraction of intact DNA. This was done by adding 350 μL buffer RLT (QIAGen, USA) to each tissue specimen and lysing it at 30 MHz for 45 seconds by making use of the

TissueLyser II (QIAGen®) (51). The resultant lysate was then pipetted onto an AllPrep® DNA spin column (QIAGen, USA) to allow for the selective and efficient binding of genomic DNA and then centrifuged at 20,000 x g for 30 seconds (51). The DNA spin column was then placed in a new 2 mL collection tube to carry out DNA purification (51). 350 μL of buffer AW1 was added to the DNA spin column and then centrifuged for 15 seconds (51). In a separate tube, 20 μL of Proteinase K was added to 60 μL of buffer AW1, mixed gently, applied to the DNA spin column and allowed to incubate for 5 minutes at room temperature (51). 350 μL of buffer AW1 was added once again and centrifuged for 15 seconds (51). After this, 500 μL of buffer AW2 was added and centrifuged for 2 minutes (51). The spin column was then placed into a new 1.5 mL Eppendorf tube (51). Subsequently, the tumour DNA was eluted in 65μL of elution buffer (51).

2.2.2 Tissue DNA quality and yields assessment

The quality and the yield of the tumour tissue DNA were assessed by making use of the Nanodrop™ 1000 (Thermo Scientific, USA) spectrophotometer, as well as the Qubit (Thermo Scientific, USA) fluorometer (51,52). The Nanodrop 1000 was used to establish the 260/280 and 260/230 ratios for which DNA purity was assessed by an absorbance at 260nm and 280nm (51). A 260/280 ratio of >1.8 for DNA was indicative of the samples possessing a good quality (51). For greater accuracy, the Qubit Fluorometer was used to measure the concentrations of the extracted DNA, as this uses a selective dye that only fluoresces when attached to its target nucleic acid, thus making the measurements highly sensitive and selective.

2.2.3 Quality control of paired blood/OSCC tumour DNA samples

To ensure that the integrity of the DNA samples was not compromised, the DNA samples were stored at -80 °C. In preparation for shipping for DNA sequencing, the DNA samples were placed in 2 mL safe lock microcentrifuge tubes and these were individually sealed with parafilm and placed into a freezer box (53). The freezer box was then placed in a 3.81 cm thick Styrofoam box filled with dry ice to keep the samples intact for the duration of the shipping process (53).

2.3 Whole Exome Sequencing:

2.3.1 WES protocol used by Novogene

In total, DNA samples from 21 matched blood/OSCC tumour samples were submitted to Novogene to conduct next-generation sequencing using their WES protocol. The 21 samples consisted of the 6 retrospective patients and the 15 prospective patients. The workflow that they used to carry out this analysis is illustrated below.

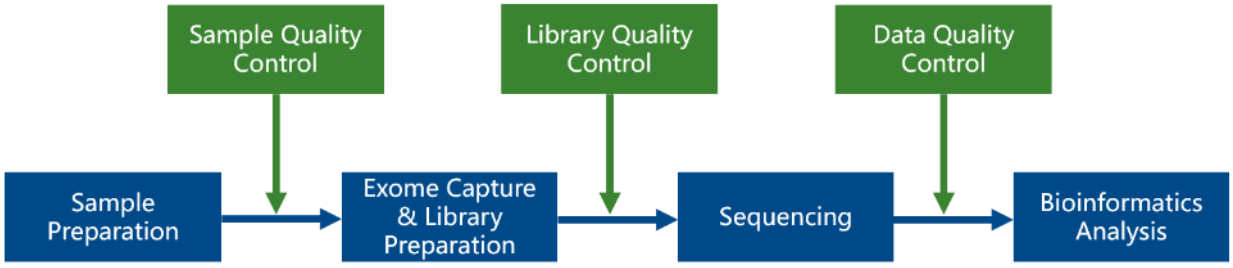

Figure 2.1: Novogene project workflow. After sample preparation, exome capture and library preparation took place. This was followed by sequencing and subsequently bioinformatic analysis. Quality control steps included sample, library and data assessments, respectively.

Genomic DNA was randomly fragmented into short fragments of the size 180-280 bp (Covaris, Massachusetts, USA). Oligonucleotide adapters were ligated to the 3' end of the resultant fragments after blunt-end conversion and 3' adenylation. The DNA fragments with adaptors were PCR enriched, size selected and purified. The prepped libraries were then hybridized in the buffer that had biotin-labelled probes, followed up by exonic capture using magnetic beads tagged with streptavidin. The captured library was then washed, PCR enriched, indexed and purified, using the AMPure XP system (Beckman Coulter, Beverly, USA) and then quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system. Libraries that passed quality control were sequenced on an Illumina NextSeq instrument, making use of the PE150 strategy in accordance with the required library concentration and data amounts.

2.3.2 Processing of DNA sequence data

The flowchart below (Figure 2.2) illustrates the pipeline of data quality control and bioinformatics analysis applied by Novogene.

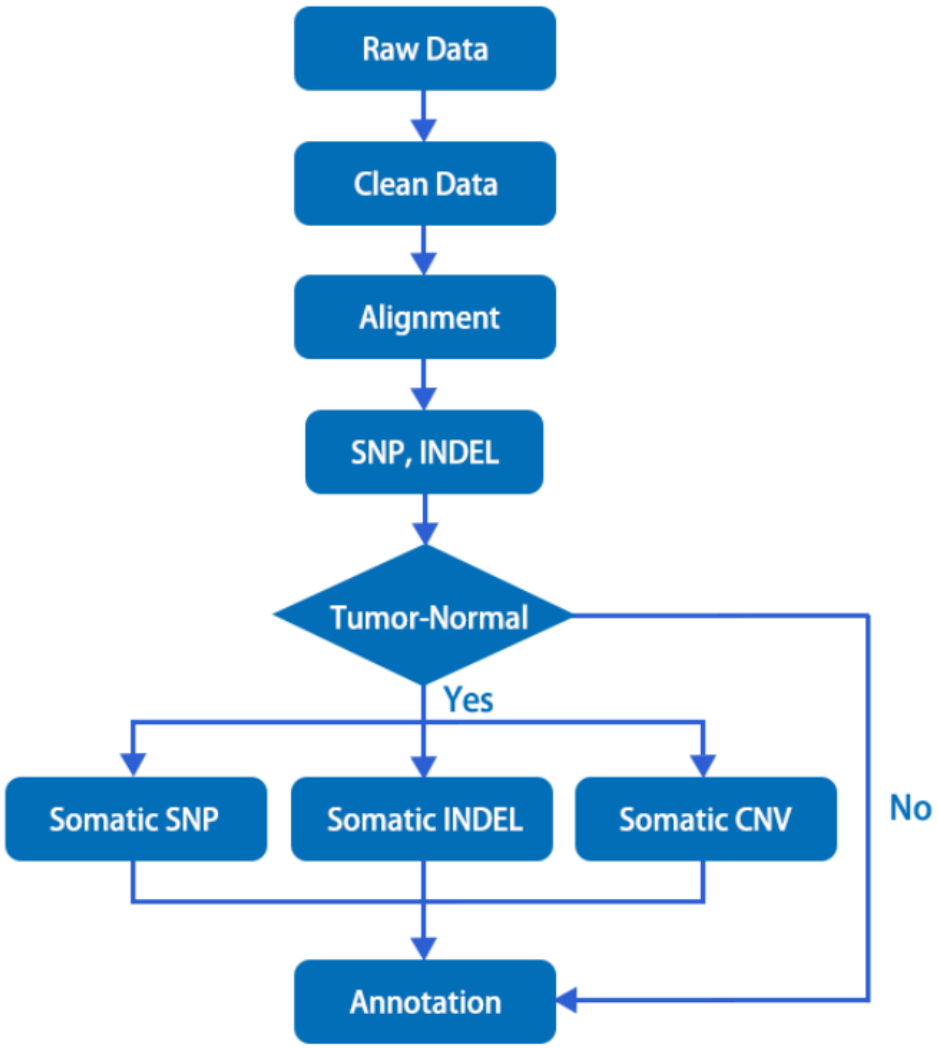

Figure 2.2: Novogene bioinformatics analysis pipeline. The quality of the raw data was assessed and reads containing adapters or with low quality were filtered out. This was followed by alignment with the reference genome and variant calling. If a variant was detected, 'Yes' it would be classified into its respective category and if not, 'No' it would simply be noted.

The detection of somatic mutations in the tumours was based on using the germline DNA from each paired sample as the control such that a variant which was present in the tumour DNA but not in the germline DNA was defined as a somatic mutation.

Novogene filtered the raw sequences for sequencing artefacts, sequence adaptors, PCR duplicates, unrecognizable nucleotides and low-quality nucleotides. This was done through three main quality control procedures that they incorporated. This included discarding the paired-end reads if either one read had an adapter contamination, that is >10 nucleotides aligned to the adapter, allowing ≤ 10% mismatches; if more than 10% of the bases were uncertain in

either one read; and if the proportion of low quality (Phred quality < 5) bases was over 50% in either one read. The sequences were then aligned to the human genome build 38 using Burrows Wheel Aligner, with PCR duplicates and variants possessing low coverage being removed. The remaining variants were then called using the Genome Analysis Toolkit (GATK) Variants and then annotated using ANNOVAR.

2.4 Bioinformatic analyses:

The GATK can be considered an industry standard for detecting SNPs and indels in germline DNA and RNAseq datasets, given that it grew to encompass somatic short variant calling, copy number and structural variation analysis (54). In addition to being variant callers, this toolkit provides various tools for processing and performing quality control of high-throughput sequencing data (54).

MuTect (v1.1.14) was incorporated for carrying out the quality control of the sequencing data as well as the calling of the mutation through the local assembly of haplotypes (55). It does so by accepting the sequence data from a matched tumour and normal sample after read alignment, duplicate read detection, base quality score recalibration and local realignment by acting on each genomic location separately through four main steps, which include: (i) removing low-quality sequence data; (ii) detecting variants in the tumour sample using a Bayesian classifier; (iii) filtration to remove false positives that resulted from the correlated sequencing artefacts that are not capture by the error model; and (iv) designating the variants as either somatic or germline by a second Bayesian classifier (56). Assessing the sequencing error rate, GC content distribution, and sequencing quality assisted in reducing the rate of both false positives and negatives.

In addition to MuTect, Strelka2 (v2.9.4), a quick and precise small variant caller for analysis of somatic variation was used (57). This iteration of the somatic calling model enhanced the original Strelka approach for liquid and late-stage tumour analysis by taking into consideration the possibility of tumour cell contamination in the normal sample (57). It runs together with the Manta structural variant and indel caller to provide indel candidates up to a given maximum indel size (49 by default) (57). For the detection of the somatic CNVs, Control-FREEC was employed. This software profiled the copy number and beta allele frequencies by automatically computing, normalizing and segmenting them and then calling the copy number alterations and loss-of-heterozygosity (58).

Annotation was performed using ANNOVAR. This tool drew a comparison of the identified variants against a reference genome (in this instance hg38) and then annotated it by taking information from databases such as the single nucleotide polymorphism database (dbSNP) and the 1000 Genomes Project to identify the genes that were affected by the variants (59). It then predicted the potential consequences of the variants by identifying whether they caused a missense mutation, frameshift mutation, splice site alteration or any other form of genetic change (59). It also provided information on the potential impact of the variants on protein structure and disease association (59).

The resultant annotated somatic variants were then assessed for potential status as a driver gene mutation by: (i) whether their effect on the gene function would likely be deleterious; (ii) whether the mutated gene(s) had a known or plausible biological connection to cancer development; and (iii) whether the gene(s) were mutated in more than one patient. The assessment of these factors was done using *R*, a free and open-source software programming language used for statistical computing and graphics. *RStudio Desktop* (version 2023.03.0+386), an integrated development of R was used to explore and model the data. An *R-script* (Appendix 5) was generated to dissect and analyse the somatic variants that occurred within the dataset that was received from Novogene.

The annotated variant data was released as a VCF file which was then formatted into a MAF file so that it could be read into R, which made use of the Maftools package (60) to further summarise, analyse, annotate and visualize the mutation data against TCGA sources (Figure 2.3).

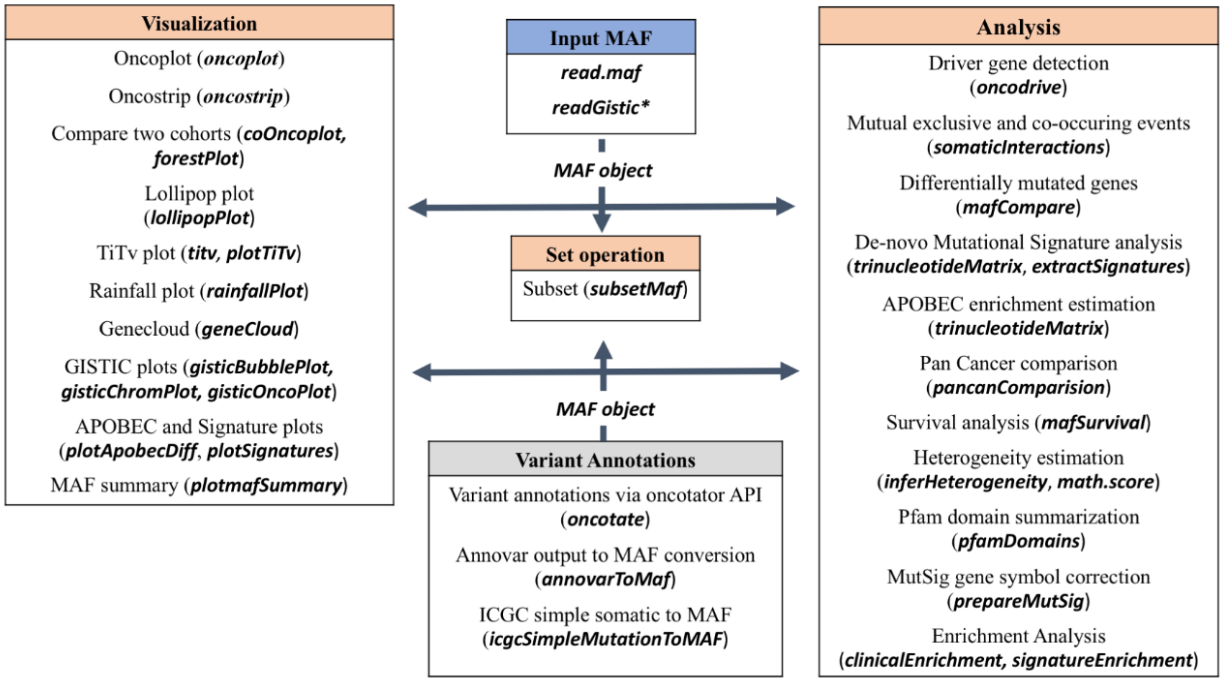

Figure 2.3: Summary indicating Maftools (60). The functions are categorised mainly into visualization and analysis modules.

To generate copy number variants, R segmented the genomic data into regions that possessed similar copy number profiles. After segmentation, the statistical algorithm GISTIC was applied to calculate the variants. It executed this calculation (G-score) by assessing the amplitude of the alterations as well the frequency at which they occurred for each sample in the data set as well as grading them individually as either gains or losses (61). In addition to this calculation, a FDR multiple testing correction was also incorporated to calculate a Q-bound significance score (61,62). Regions that met both the G-score and Q-bound threshold cut-offs were then visualised.

For the analysis of mutation signatures, a mutational matrix – the matrix representing the occurrence of the different types of mutations in the sample set – was generated. The mathematical algorithm, non-negative factorization matrix (NMF) was incorporated in order to decompose the mutational matrix into two lower-dimensional matrices which would illustrate the mutational processes and their activities (63). A comparison against the catalogue of somatic mutations in cancer (COSMIC) database was done to compare and identify known mutational signatures present in the dataset. An apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) enrichment was also performed to assess cancer development and prognosis.

Regarding molecular pathway analysis, the mutational data was analysed as well as its impact on cancer-related pathways. From the Maftools package, “Oncogenic pathways” was incorporated in order to detect the enrichment of the known oncogenic signalling pathways and illustrated it in graphical format (60).

3 Results

This chapter focuses on presenting the findings that were derived from the techniques and analyses that were described in the previous chapter. The results are presented as two discrete tumour sequencing projects. Project 1, where the candidate's role was to extract DNA from tumour tissue and carry out quality control so that samples with sufficient and good-quality DNA could be sent to the Sanger for WES. Project 2, where bioinformatic analysis of the samples sent to Novogene (UK) for WES was conducted under the supervision of Dr Carl Chen, to identify somatic mutations in the tumours and to look for possible associations of the molecular findings with clinical features.

3.1 Clinical and demographic data and potential association with molecular data

Regarding the clinical and demographic data of the Project 2 (21 OSCC participants), characteristics such as ethnicity, age, BMI, Tumour histology and HIV status were obtained, but the data was incomplete for some patients. The available data is shown in Table 3.1.

Table 3.1: Demographic and clinical data from OSCC patients

Sample code	**Ancestry**	**Age**	**BMI**	**Tumour histology**[1]	**HIV status**
s48	Black African	58	Unknown	3	Unknown
s41	Black African	46	Unknown	2	Unknown
s38	Black African	64	Unknown	2	Unknown
s180	Black African	64	Unknown	3	Unknown
s182	Black African	63	Unknown	3	Unknown
s186	Black African	64	Unknown	3	Unknown
LF141	Mixed ancestry	63	14.5	1	Negative
LF175	Black African	65	20.9	1	Positive
LF280	Black African	63	13.0	1	Unknown
LF289	Black African	65	30.2	3	Unknown
LF293	Black African	61	18.8	2	Negative
LF294	Black African	71	15.8	1	Negative
LF298	Black African	89	33.8	3	Negative

[1] Tumour histology is labelled numerically; 1 – Infiltrating keratinizing squamous cell carcinoma, 2 – Moderately differentiated squamous cell carcinoma, 3 – Poorly differentiated squamous cell carcinoma

LF306	Black African	84	Unknown	1	Unknown
LF312	Black African	36	24.7	1	Positive
LF317	Black African	60	Unknown	2	Negative
LF320	Black African	61	16.0	1	Unknown
LF357	Black African	55	27.6	1	Positive
LF367	Black African	69	19.1	1	Negative
LF374	Black African	87	23.9	1	Unknown
LF415	Black African	52	33.2	3	Negative

The table illustrates that the study participants were mostly of African ancestry with only one participant of mixed ancestry. The mean age at diagnosis was 63.8 years. BMI values could be obtained from only 13 patients, with a mean BMI of 22.4, and several patients were underweight as a result of oesophageal obstruction caused by the cancer. Tumours were graded by histological examination according to the type of differentiation seen with a value of either 1 – an infiltrating keratinizing squamous cell carcinoma, 2 – a moderately differentiated squamous cell carcinoma or 3 – a poorly differentiated squamous cell carcinoma, assigned to each sample. Approximately 48% of the samples had a grading of 1, meaning that they could be classified as having a high-grade, aggressive tumour, followed by 19% of the samples having a grading of 2, which suggests that 50 – 70% of the tumour cells show keratinization and 33% a grading of 3, meaning that <50% of the tumour cells show keratinization (64,65). HIV status was available on a subset of patients, with 14% of participants HIV-positive, 34% HIV-negative and 52% of the patients, their status was unknown.

3.2 Extraction of DNA from OSCC tumour tissue for sequencing by the Sanger Institute

In Project 1, 36 OSCC tumour tissue samples from patients recruited at Chris Hani Baragwanath and Charlotte Maxeke Johannesburg Academic Hospitals, Gauteng and from Grey's Hospital, Kwa-Zulu Natal were assessed by histopathological analysis for tumour content and for biopsy size. DNA was extracted at the Somatic Cell Genetics laboratory under the supervision of Mr Mishalan Moodley. As mentioned in the Methods component of this thesis, the samples were extracted by following the protocol in the QIAGen AllPrep® Universal kit. The Nanodrop was then used to establish the absorbance ratios of the resultant DNA. In addition to the absorbance ratios, the concentrations were determined using Qubit.

After filtering samples for tumour content, size and DNA yield, a total of 36 samples were deemed suitable for DNA sequencing. The QC results of these samples are shown in Table 3.2.

Table 3.2: Tumour DNA concentration and absorbance measured with Nanodrop

	DNA		
Sample ID	**Concentration (ng/μl)**	**Total yield (μg)**	**A260/A280**
LF153	25	1.5	1.63
LF154	35	2.1	1.90
LF159	55	3.3	1.47
LF162	50	3.0	1.80
LF163	29	1.8	1.61
LF172	37	2.2	1.81
LF185	13	0.79	1.92
LF187	21	1.3	2.02
LF190	7	0.43	1.58
LF191	10	0.62	1.94
LF192	-	-	-
LF198	8	0.49	1.52
LF204	5	0.3	1.69
LF212	35	2.1	1.84
LF221	5	0.29	1.83
LF224	8.7	0.52	0.97
LF225	50	3.0	1.48
LF227	35	2.1	1.83
LF234	126	7.5	1.65
LF235	23	1.4	1.83
LF238	27	1.7	1.53
LF239	-	-	-
LF240	10	0.61	1.91
LF243	28	1.7	1.79
LF246	-	-	-
LF248	24	1.4	1.66
LF249	41	2.5	1.79
LF252	21	1.3	1.88
LF257	31	1.9	1.80

LF260	-	-	-
LF261	52	3.1	1.70
LF266	-	-	-
LF284	108	6.5	1.26
LF325	12	0.71	1.59
LF342	26	1.6	1.70
LF385	14	0.86	1.81

[2]

In order to be able to determine the accuracy of the DNA concentration of the samples that would be sent for sequencing to the Sanger Institute, the Qubit spectrophotometer was then employed.

DNA concentrations of the samples measured using the Qubit fluorescence spectrophotometer are shown in Table 3.3.

Table 3.3: Tumour DNA yields measured with the Qubit

SAMPLE ID	**Concentration (ng/μl)**	**Volume (μl)**	**Final Total Yield (μg)**
LF153	15.7	60	0.942
LF154	34.1	60	2.046
LF159	36.1	60	2.166
LF162	53	60	3.18
LF163	30.6	60	1.836
LF172	38.8	60	2.328
LF185	11.6	60	0.696
LF187	24	60	1.44
LF190	3.81	60	0.229
LF191	9.72	60	0.583
LF192	23	60	1.38
LF198	13	60	0.78
LF204	3.4	60	0.204
LF212	53	60	3.18
LF221	1.56	60	0.936
LF224	50	60	3

[2] There are samples for which readings could not be taken when using the Nanodrop, however they could be measured when using the Qubit.

LF225	27.9	60	1.674
LF227	44.9	60	2.964
LF234	60	60	3.6
LF235	24.7	60	1.482
LF238	21.1	60	1.266
LF239	19.7	60	1.182
LF240	9.85	60	0.591
LF243	31.9	60	1.914
LF246	30.1	60	1.806
LF248	33.4	60	2.004
LF249	55	60	3.3
LF252	8.52	60	0.5112
LF257 [3]	>0.5	60	**(too low)**
LF260	9.59	60	0.5754
LF261	60	60	3.6
LF266	6.23	60	0.3738
LF284	29.2	60	1.752
LF325	6.42	60	0.3852
LF342	19.8	60	1.188
LF385	11.1	60	0.666

Prior to the Sanger Institute carrying out WES of the samples, they performed their own quality control measures by measuring the DNA yields in order to ascertain that the samples were adequate for conducting WES and also by genotyping DNA from all matched blood/tumour pairs on a genome-wide SNP array to check that the DNA samples originated from the same individual. They then sent a report to our team which showed that all 72 samples (36 matched pairs) had sufficient DNA for sequencing and that all pairs were genetically matched (Table 3.4).

Table 3.4: DNA yields and genotype match from the Sanger Institute

External ID	SANGER SAMPLE ID	TOTAL VOLUME (µl)	CONCENTRATION (ng/µl)	TOTAL AMOUNT (µg)	Genotype match
LF153_T_Tumour	6814STDY12716804	38	13,9	0,5282	y

[3] Had a low yield prior to being shipped to the Sanger, however when they performed their own QC, there was a greater yield.

LF153_N_Normal	6814STDY12716842	38	21,38	0,81244	y
LF154_T_Tumour	6814STDY12716805	38	12,67	0,48146	y
LF154_N_Normal	6814STDY12716843	38	11,63	0,44194	y
LF159_T_Tumour	6814STDY12716806	38	19,45	0,7391	y
LF159_N_Normal	6814STDY12716844	38	15,21	0,57798	y
LF162_T_Tumour	6814STDY12716807	38	25,26	0,95988	y
LF162_N_Normal	6814STDY12716845	38	15,29	0,58102	y
LF163_T_Tumour	6814STDY12716808	28	14,63	0,40964	y
LF163_N_Normal	6814STDY12716846	38	8,71	0,33098	y
LF172_T_Tumour	6814STDY12716809	38	22,92	0,87096	y
LF172_N_Normal	6814STDY12716847	38	16,37	0,62206	y
LF185_T_Tumour	6814STDY12716810	54	13,23	0,71442	y
LF185_N_Normal	6814STDY12716848	38	14,17	0,53846	y
LF187_T_Tumour	6814STDY12716811	28	12,33	0,34524	y
LF187_N_Normal	6814STDY12716849	38	14,61	0,55518	y
LF190_T_Tumour	6814STDY12716812	28	3,41	0,09548	y
LF190_N_Normal	6814STDY12716850	44,7	13,5	0,60345	y
LF191_T_Tumour	6814STDY12716813	28	11,72	0,32816	y
LF191_N_Normal	6814STDY12716851	38	14,83	0,56354	y
LF192_T_Tumour	6814STDY12716814	58	25,37	1,47146	y
LF192_N_Normal	6814STDY12716852	43,4	12,06	0,523404	y
LF198_T_Tumour	6814STDY12716815	38	10,57	0,40166	y
LF198_N_Normal	6814STDY12716853	38	11,51	0,43738	y
LF204_T_Tumour	6814STDY12716816	61,9	4,86	0,300834	y
LF204_N_Normal	6814STDY12716854	38	15,19	0,57722	y
LF212_T_Tumour	6814STDY12716817	38	10,11	0,38418	y
LF212_N_Normal	6814STDY12716855	38	15,17	0,57646	y
LF221_T_Tumour	6814STDY12716818	28	2,46	0,06888	y
LF221_N_Normal	6814STDY12716856	38	13,09	0,49742	y

LF224_T_Tumour	6814STDY12716819	38	10,58	0,40204	y
LF224_N_Normal	6814STDY12716857	38	13,48	0,51224	y
LF225_T_Tumour	6814STDY12716820	38	18,93	0,71934	y
LF225_N_Normal	6814STDY12716858	38	16,03	0,60914	y
LF227_T_Tumour	6814STDY12716821	28	7,22	0,20216	y
LF227_N_Normal	6814STDY12716859	46,2	13,3	0,61446	y
LF234_T_Tumour	6814STDY12716822	28	20,2	0,5656	y
LF234_N_Normal	6814STDY12716860	38	14,2	0,5396	y
LF235_T_Tumour	6814STDY12716823	31,7	27,36	0,867312	y
LF235_N_Normal	6814STDY12716861	38	13,83	0,52554	y
LF238_T_Tumour	6814STDY12716824	28	15	0,42	y
LF238_N_Normal	6814STDY12716862	38	12,18	0,46284	y
LF239_T_Tumour	6814STDY12716825	38	14,99	0,56962	y
LF239_N_Normal	6814STDY12716863	38	13,42	0,50996	y
LF240_T_Tumour	6814STDY12716826	38	12,9	0,4902	y
LF240_N_Normal	6814STDY12716864	38	19,54	0,74252	y
LF243_T_Tumour	6814STDY12716827	38	7,76	0,29488	y
LF243_N_Normal	6814STDY12716865	38	15,41	0,58558	y
LF246_T_Tumour	6814STDY12716828	38	8,62	0,32756	y
LF246_N_Normal	6814STDY12716866	38	13,4	0,5092	y
LF248_T_Tumour	6814STDY12716829	28	7,14	0,19992	y
LF248_N_Normal	6814STDY12716867	38	17,15	0,6517	y
LF249_T_Tumour	6814STDY12716830	58	14,83	0,86014	y
LF249_N_Normal	6814STDY12716868	38	12,64	0,48032	y
LF252_T_Tumour	6814STDY12716831	28	8,35	0,2338	y
LF252_N_Normal	6814STDY12716869	37,3	11,12	0,414776	y
LF257_T_Tumour	6814STDY12716832	38	30,93	1,17534	y
LF257_N_Normal	6814STDY12716870	38	15,98	0,60724	y
LF260_T_Tumour	6814STDY12716833	50	11,18	0,559	y

LF260_N_Normal	6814STDY12716871	38	14,47	0,54986	y
LF261_T_Tumour	6814STDY12716834	38	18,62	0,70756	y
LF261_N_Normal	6814STDY12716872	38	11,02	0,41876	y
LF266_T_Tumour	6814STDY12716835	58	7,53	0,43674	y
LF266_N_Normal	6814STDY12716873	38	17,47	0,66386	y
LF284_T_Tumour	6814STDY12716836	38	5,53	0,21014	y
LF284_N_Normal	6814STDY12716874	44,1	16,49	0,727209	y
LF325_T_Tumour	6814STDY12716837	28	7,43	0,20804	y
LF325_N_Normal	6814STDY12716875	45,4	9,22	0,418588	y
LF342_T_Tumour	6814STDY12716838	25,8	5,33	0,137514	y
LF342_N_Normal	6814STDY12716877	28	12,71	0,35588	y
LF385_T_Tumour	6814STDY12716839	50,7	13,88	0,703716	y
LF385_N_Normal	6814STDY12716878	28	12,97	0,36316	y

Whole exome sequencing on all 72 samples from the 36 matched pairs was then carried out at the Sanger Institute, followed by bioinformatic analysis to identify somatic mutations in the tumours. Their interim report on the results of this analysis will be discussed in Chapter 4.

3.3 Whole exome sequencing and bioinformatic analysis of 21 OSCC/blood tumour pairs

In Project 2, DNA extracted from 21 matched blood/tumour pairs in the Somatic Cell Genetics Unit at Wits University was sent to Novogene for whole exome sequencing as described in Chapter 2. The raw sequence data, aligned sequence data and preliminary annotated results were returned to our research team, together with a report on the quality of the sequencing they had carried out, as well as a detailed description of the workflow. The data was then analysed to determine the nature and extent of somatic mutations present in the tumour samples and to assess whether there was any association with clinical features.

3.3.1 DNA sequence quality in tumour samples

A Phred quality score of 30 (Q30), which equates to a probability of an incorrect base call of 1 in 1000 and a base call accuracy of 99.9%, was set as the benchmark by Novogene, and required

that this should be above 80% for paired-end sequence data. This Q30 criterion was met in all of the blood and tumour DNA samples sequenced in this study, with a range of 85.96 – 94.09% of bases across the samples. Data on the percentage of the coverage for each target region as well as the average sequencing depth for each tumour and its blood counterpart are shown in Table 3.5 below. This shows that the coverage of the target regions was >99% in the majority of samples, and that the percentage of bases with a sequencing depth of >10X in the target region was achieved in the majority of samples. An exception was tumour sample LF141T, in which the depth of coverage was low compared to other samples.

Table 3.5: Sequencing depth and coverage of target regions for the 21 samples

Sample name	***Coverage of target region***	***Average sequencing depth on target***	***Fraction of target covered with at least 50x***	***Fraction of target covered with at least 20x***	***Fraction of target covered with at least 10x***
s38T	99.5%	188.46	83.0%	94.7%	97.6%
s38B	97.5%	137.62	60.3%	79.7%	88.5%
s41T	99.7%	195.36	83.9%	95.0%	97.9%
s41B	99.3%	89.01	61.3%	86.6%	94.2%
s48T	99.5%	164.69	74.1%	90.9%	96.0%
s48B	99.5%	122.56	70.8%	90.2%	95.7%
s180T	99.4%	163.04	81.5%	94.3%	97.4%
s180B	98.7%	86.44	56.9%	83.3%	92.3%
s182T	99.7%	195.93	86.1%	95.9%	98.3%
s182B	99.2%	127.11	69.1%	88.8%	94.8%
s186T	99.7%	188.07	85.8%	95.8%	98.2%
s186B	99.1%	99.83	61.7%	85.8%	93.6%
LF141T	92.3%	54.51	30.5%	55.4%	69.9%
LF141B	99.5%	81.25	67.1%	91.8%	96.9%
LF175T	99.4%	80.26	54.8%	85.7%	94.7%
LF175B	99.4%	84.20	68.0%	91.9%	96.9%
LF280T	99.3%	69.69	46.7%	81.5%	93.1%
LF280B	99.5%	94.99	75.2%	93.5%	97.3%
LF289T	99.6%	154.66	86.3%	96.7%	98.6%
LF289B	99.4%	76.11	65.3%	91.1%	96.5%
LF293T	99.6%	154.61	87.3%	96.7%	98.5%

LF293B	99.4%	70.46	61.9%	90.4%	96.3%
LF294T	99.3%	62.01	41.8%	78.4%	91.9%
LF294B	99.4%	70.23	61.4%	90.0%	96.1%
LF298T	99.5%	137.16	85.2%	96.2%	98.3%
LF298B	99.4%	78.28	66.4%	91.4%	96.7%
LF306T	99.3%	68.28	46.6%	81.9%	93.2%
LF306B	99.5%	99.30	76.3%	93.7%	97.4%
LF312T	96.9%	60.12	38.2%	71.9%	85.9%
LF312B	99.5%	93.53	74.6%	93.3%	97.2%
LF317T	99.6%	150.52	87.5%	96.8%	98.6%
LF317B	99.4%	83.12	72.0%	92.7%	97.0%
LF320T	98.2%	160.69	71.5%	88.4%	93.9%
LF320B	99.4%	83.12	69.6%	92.1%	96.8%
LF357T	98.4%	102.61	62.2%	86.0%	93.4%
LF357B	99.4%	83.19	69.7%	92.1%	96.8%
LF367T	96.9%	78.33	47.2%	76.6%	88.0%
LF367B	99.4%	73.69	63.7%	90.8%	96.5%
LF374T	98.5%	200.29	76.4%	90.7%	95.1%
LF374B	99.5%	84.12	70.3%	92.4%	96.9%
LF415T	99.5%	148.55	87.2%	96.6%	98.5%
LF415B	99.4%	84.11	70.5%	92.4%	96.9%

Plots of the number of reference bases versus the depth of coverage of the six retrospective and fifteen recent patients are displayed in the coverage histograms below (Figures 3.1 and 3.2). The figures indicate a relatively high sequencing depth and base coverage across each chromosome for most samples.

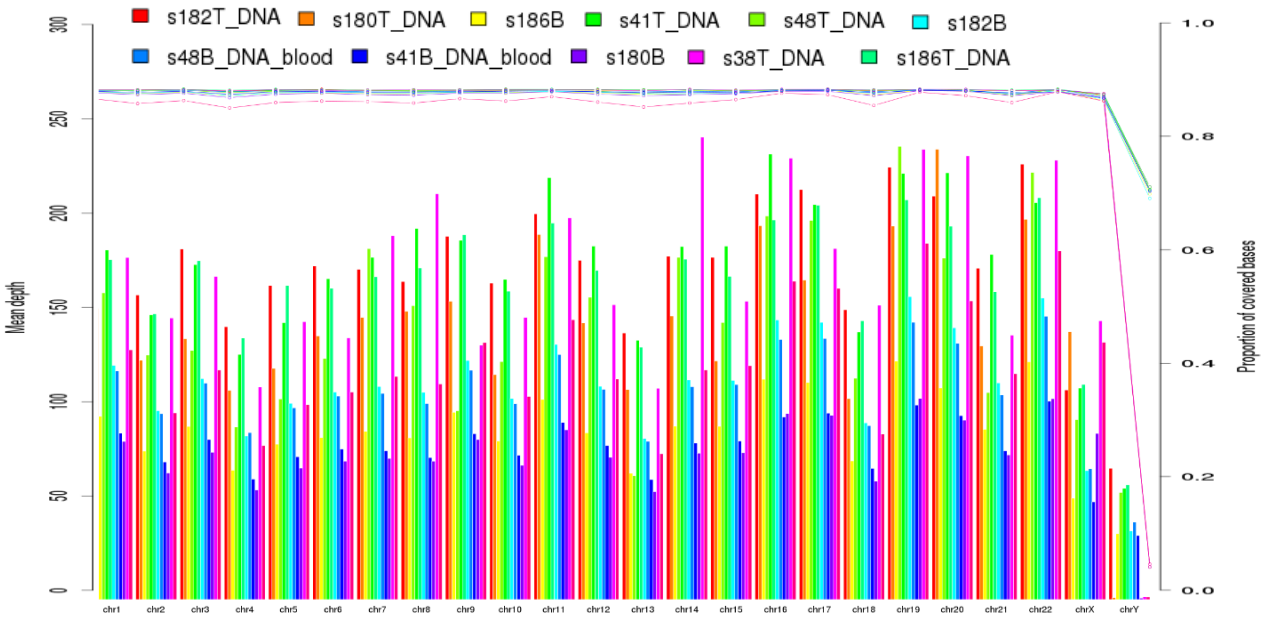

Figure 3.1: The average sequencing depth (bar plot) and the coverage (dot-line plot) present in each chromosome for the 6 retrospectively collected samples as received in the sequencing report from Novogene. The x-axis represents the chromosome; the left y-axis represents mean depth; the right y-axis represents the coverage.

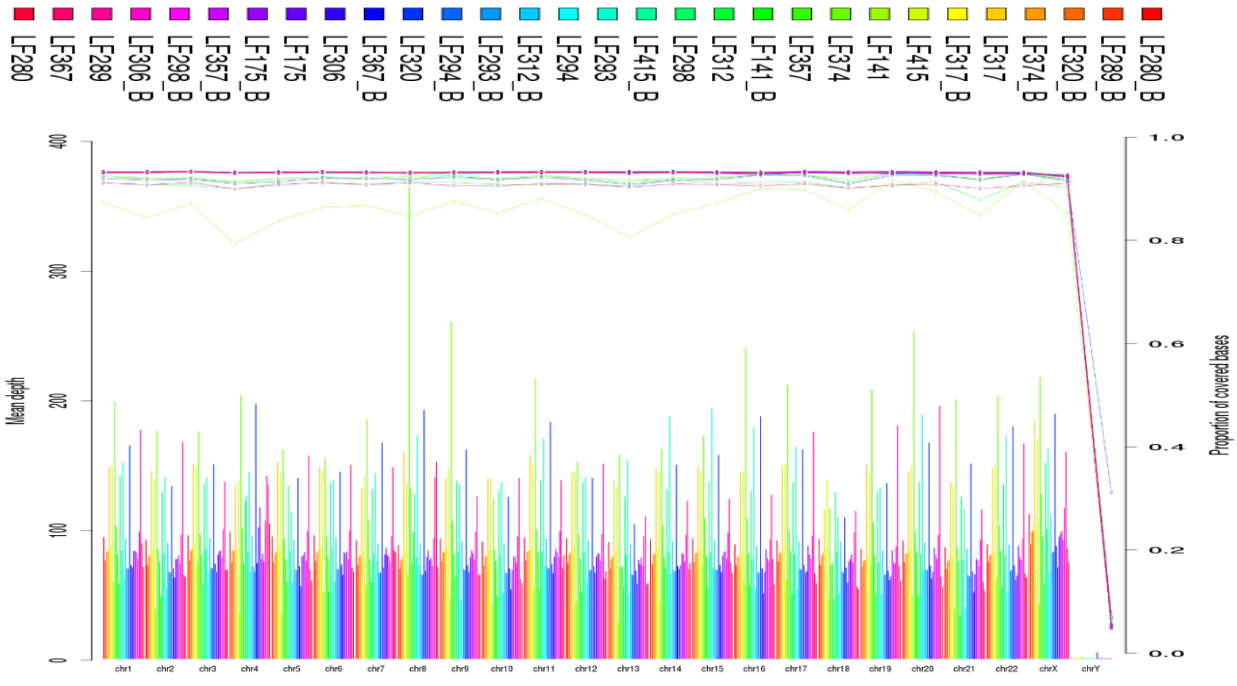

Figure 3.2: The average sequencing depth (bar plot) and the coverage (dot-line plot) present in each chromosome for the 15 more recently collected samples as received in the sequencing report from Novogene. The x-axis represents the chromosome; the left y-axis represents the average depth; the right y-axis represents the coverage

3.3.2 Bioinformatic analysis of somatic mutations in OSCC tumour tissue:

Obtaining, analysing, and visualising the mutational variables that were associated with OSCC in this study was done using the *R* package, Maftools (60). This analysis demonstrated that SNVs were much more prevalent in comparison to somatic indels (see Table 3.6), and the total number varied widely per tumour.

Table 3.6: The total number of somatic SNVs and indels for the 21 OSCC tumour samples derived from the Novogene sequencing report.

Sample name	***Total number of SNVs***	***Total number of indels***
s38T	507	288
s41T	951	154
s48T	438	200
s180T	814	252
s182T	286	260
s186T	267	212
LF141T	1640	125
LF175T	1782	37
LF280T	2009	49
LF289T	516	96
LF293T	465	78
LF294T	1824	33
LF298T	298	70
LF306T	1557	44
LF312T	27598	192
LF317T	404	52
LF320T	3498	50
LF357T	2886	66
LF367T	24379	233
LF374T	4026	77
LF415T	192	57

The total number of mutations, otherwise known as the median per mega-base tumour mutation burden discovered in the tumour sample was 1.94 (~2 mutations per Mb) with 36.18 Mb of the

coding region captured. The tumour mutation burden (TMB) in our set of 21 OSCC samples was then compared with TMBs in a wide variety of cancers in The Cancer Genome Atlas (TCGA – Figure 3.3). This showed that our limited sample had a lower TMB than oesophageal cancer (ESCA) in the TCGA, although the ESCA data is based on an amalgamation of both major types of oesophageal cancer, namely squamous cell carcinoma and adenocarcinoma. Our WES-OSCC dataset from this study had a fairly average mutation load, similar to uterine endometrial carcinoma and ovarian cancer.

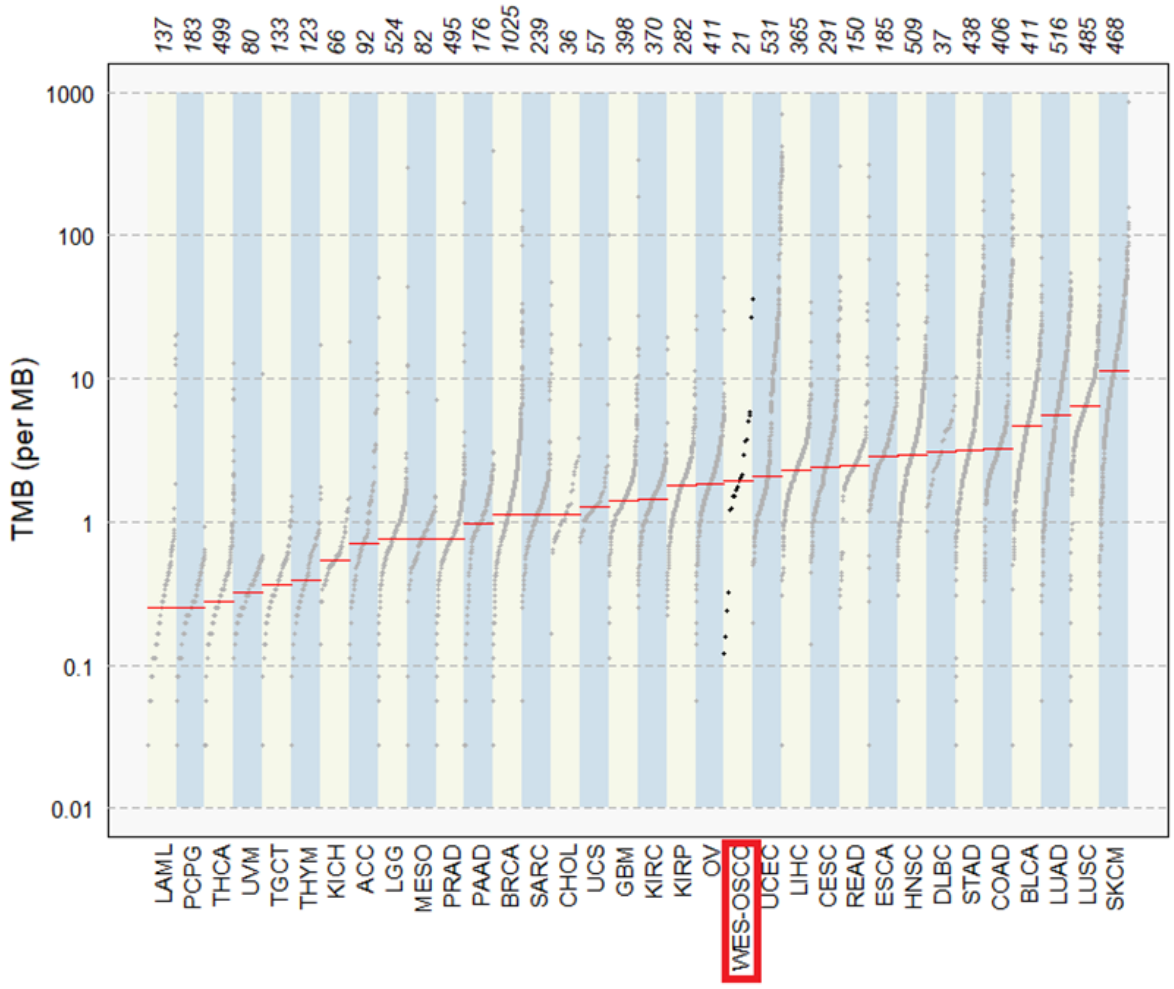

Figure 3.3: A comparison of the TMB in TCGA datasets.

An overview of the types of somatic variants detected in the tumours showed that single nucleotide variants (shown as "SNP" in the Maftools output) were far more common than indel variants, and that C>T substitutions were the most common type of SNV. Missense variants were much more prevalent than nonsense and splice site variants. The median number of non-

synonymous variants that occurred per sample was 87, whilst the highest number of variants in a single tumour was 1,815.

A chart of the genes which were most commonly mutated in the OSCC tumours is shown in Figure 3.4.

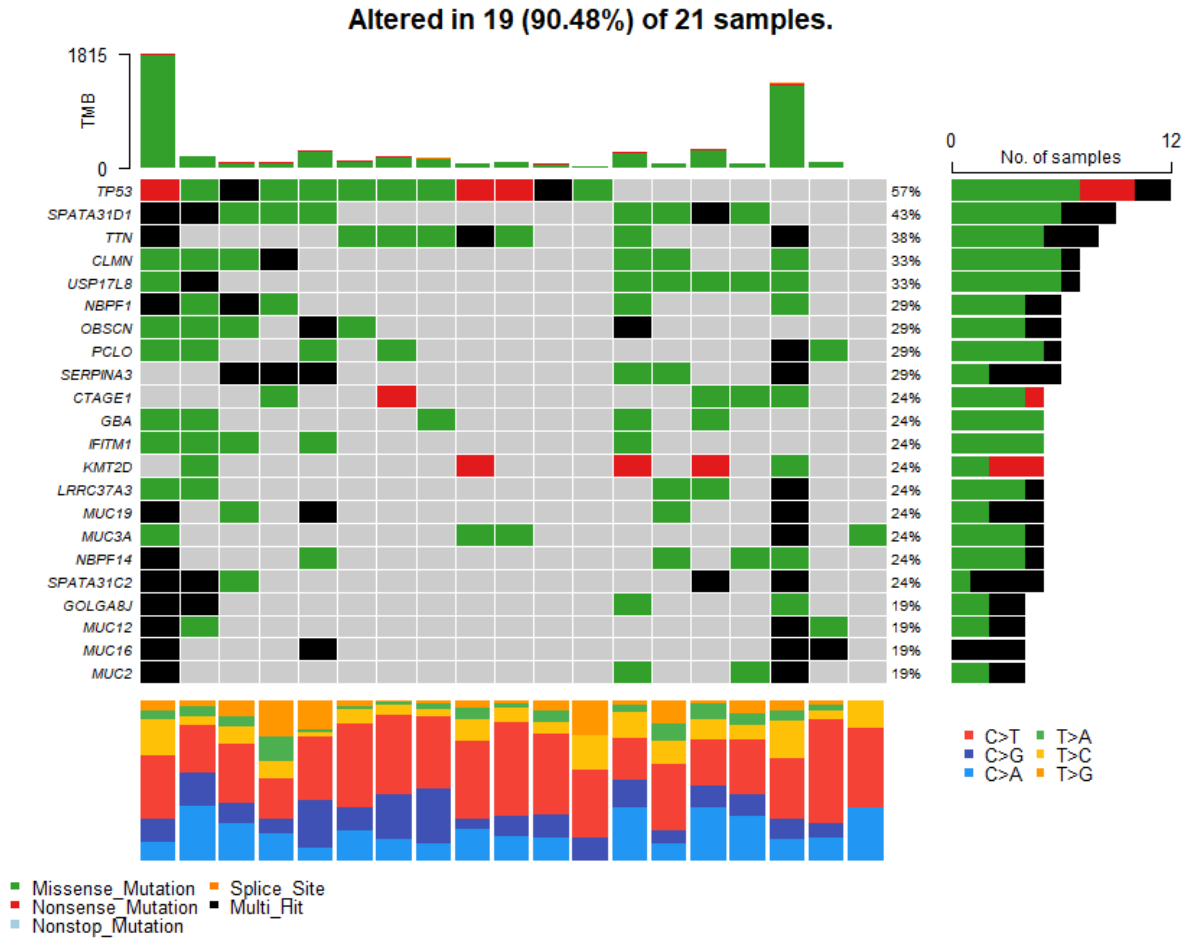

Figure 3.4: The 22 most frequently mutated genes in the OSCC tumours generated using R-Studio. Tumour samples are arranged from left to right and the type of mutation for each sample is annotated by different colours with the mutation frequency illustrated on the right side of the panel and the number of mutations per sample presented at the top of the panel. The proportions of the different types of base changes for each sample are shown at the bottom. Only 19 of the 21 samples had mutations in the 22 genes listed.

As can be seen in Figure 3.4, the most commonly mutated gene was *TP53*, which occurred in 57% of tumours. Other genes such as *SPATA31D1* and *TTN* were also frequently mutated; the possible biological significance of these findings will be discussed in Chapter 4. C>T substitutions were the most common mutations in the majority of tumour samples. A transition

and transversion plot was generated which summarised the manner in which SNVs were distributed among the six different categories.

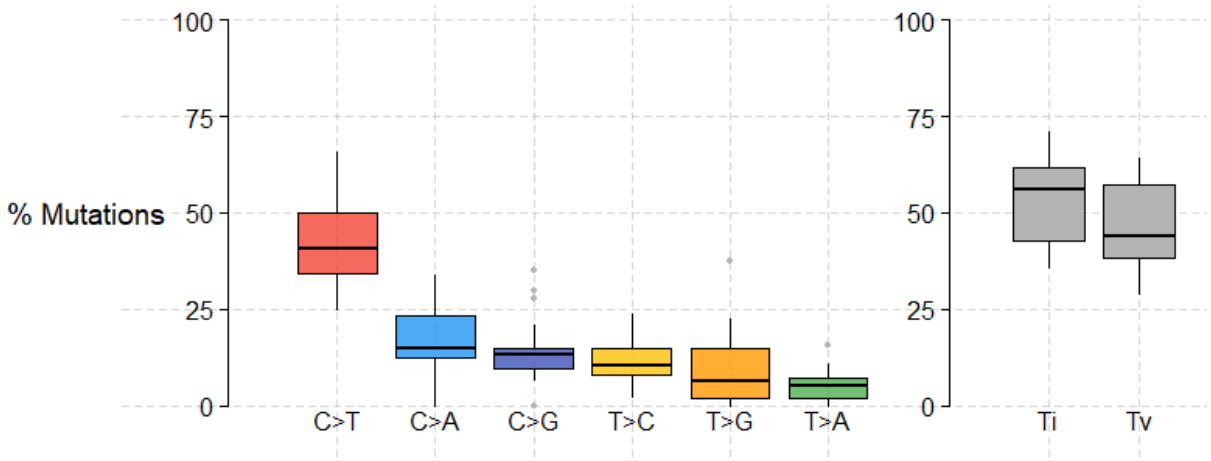

Figure 3.5: The distribution of the SNVs in OSCC is displayed in the transition and transversion plot generated using R-Studio.

In view of the importance of *TP53* mutations in OSCC, a lollipop plot of the distribution of the somatic mutations in the gene's protein domains and hotspots for mutations is shown in Figure 3.6. All but one of the mutations observed are in the DNA binding domain of the protein.

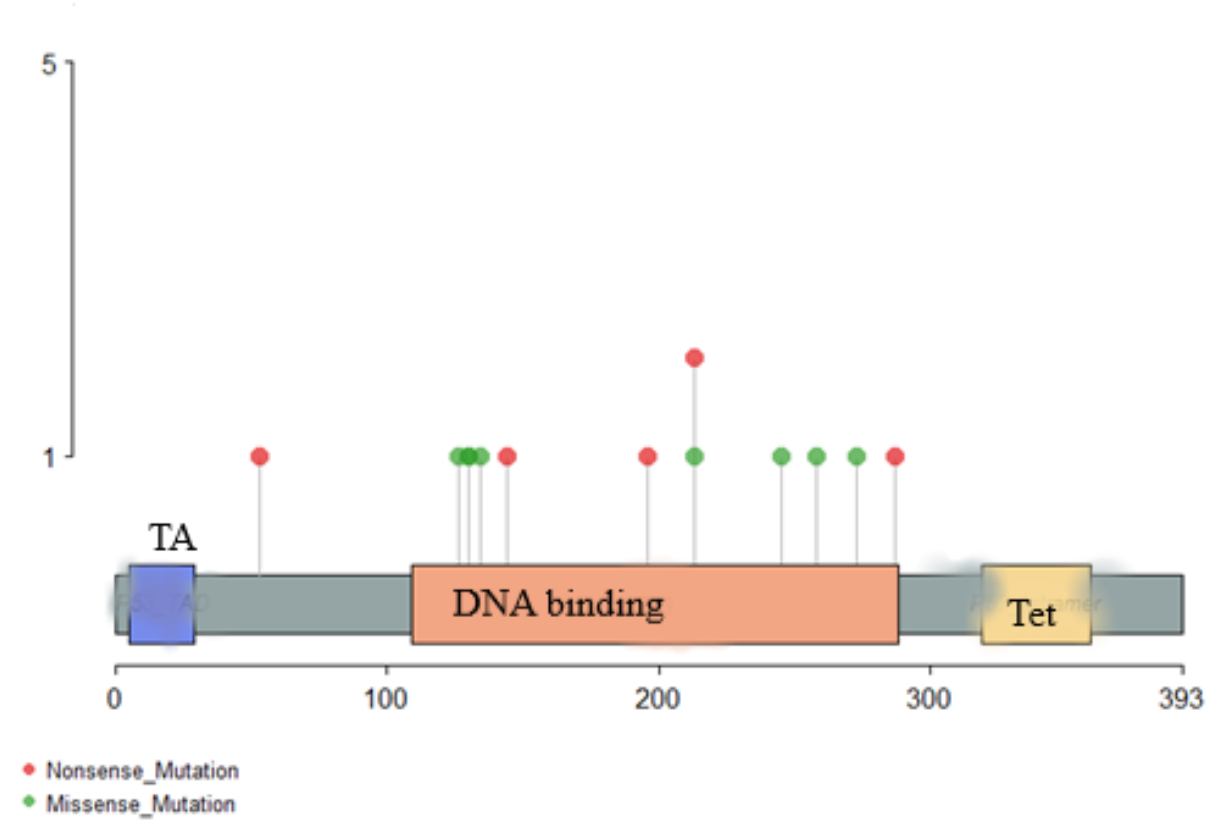

Figure 3.6: Distribution of TP53 mutations detected in this study in relation to p53 protein domains (TA = transactivation domain, DNA binding = DNA binding domain, Tet = tetramerisation domain) generated in R-Studio.

The variant allele frequencies (VAF) of the somatic variants detected in the tumours was then used to determine the clonal status of the top mutated genes in the sample set. This analysis (Figure 3.7) showed that the genes with the highest VAF were *TP53, TTN* and *MUC19*. This suggests that mutation of these genes were relatively early events in the development of the tumours.

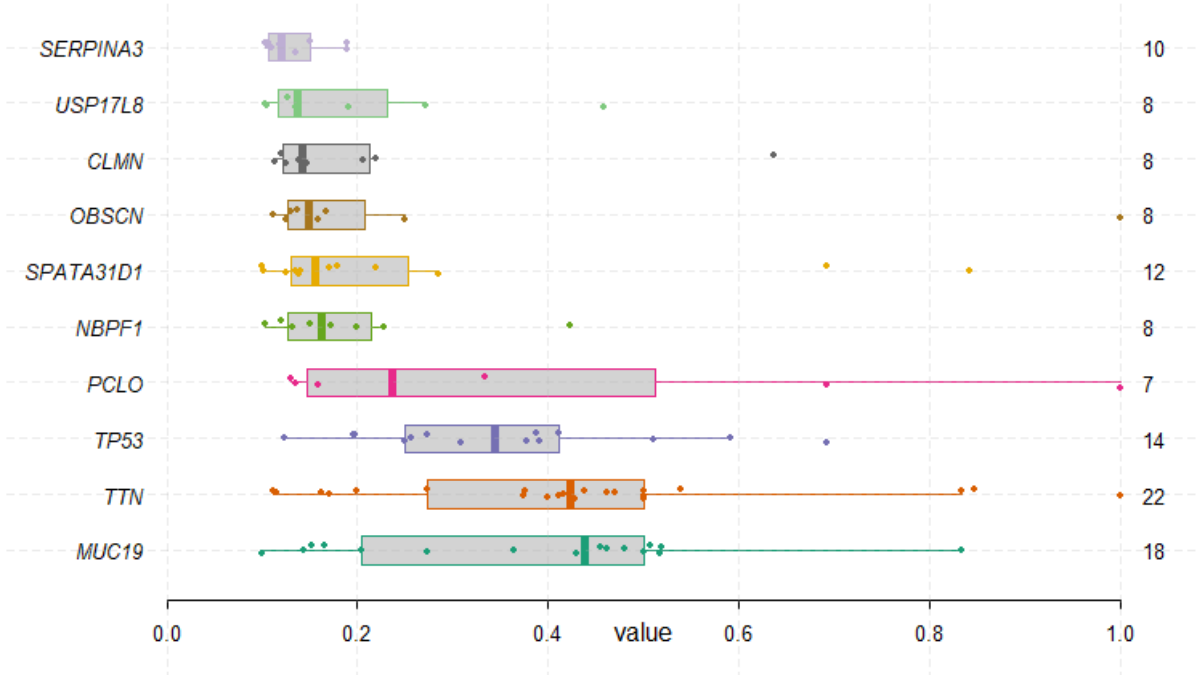

Figure 3.7: Variant allele frequency indicating the clonal status of the top mutated genes. The x-axis represents the median value of VAF distribution; y-axis represents the clonal status.

The analysis of the copy number alterations in the sample set was done using GISTIC 2.0. This revealed multiple regions showing gain or loss of chromosomal material, indicative of a high degree of complexity in the tumour genome. A plot of these abnormalities is shown in Figure 3.8.

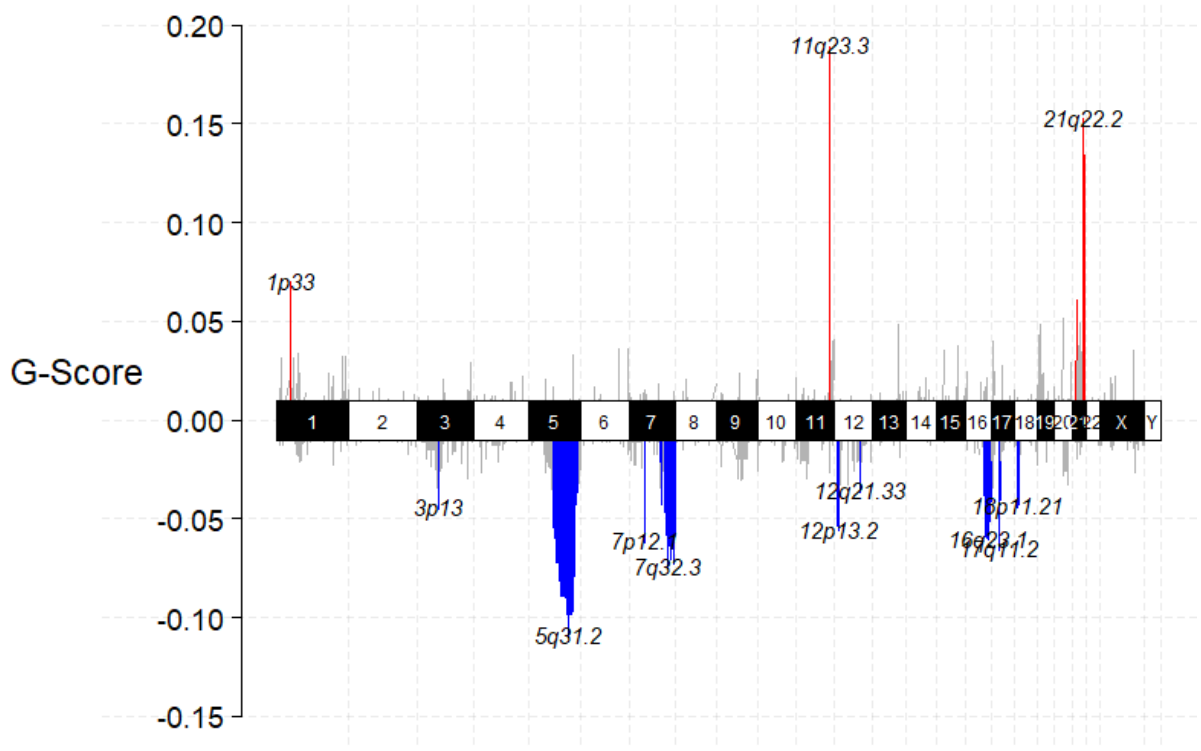

Figure 3.8: GISTIC Plot of the 21 tumours identifying gains (red peaks) and losses (blue peaks) regions derived through using R-Studio.

The G-score takes into account both the magnitude and frequency of aberrations, with a positive score indicating frequent and high-amplitude amplification and a negative score suggesting a frequent high-amplitude deletion. Three focal gains can be observed in Figure 3.8 involving chromosomes *1p33, 11q23.3* and *21q22.2*. Nine focal deletions were detected, covering chromosomal regions *3p13, 5q31.2, 7p12.1, 7q32.3, , 12p13.2, 12q21.33, 16q23.1, 17q11.2* and *18p11.21*. The genes that exist at these chromosomes or in regions close by could not be determined precisely as this was out of the scope of this study. Fine-mapping of these regions will provide important information on the genes involved.

An analysis of mutation signatures was performed with the intention of understanding the mutational patterns and underlying causes of disease. The NMF technique was used to estimate the signatures and an APOBEC-based enrichment analysis to estimate the strength of APOBEC-related mutagenic scores.

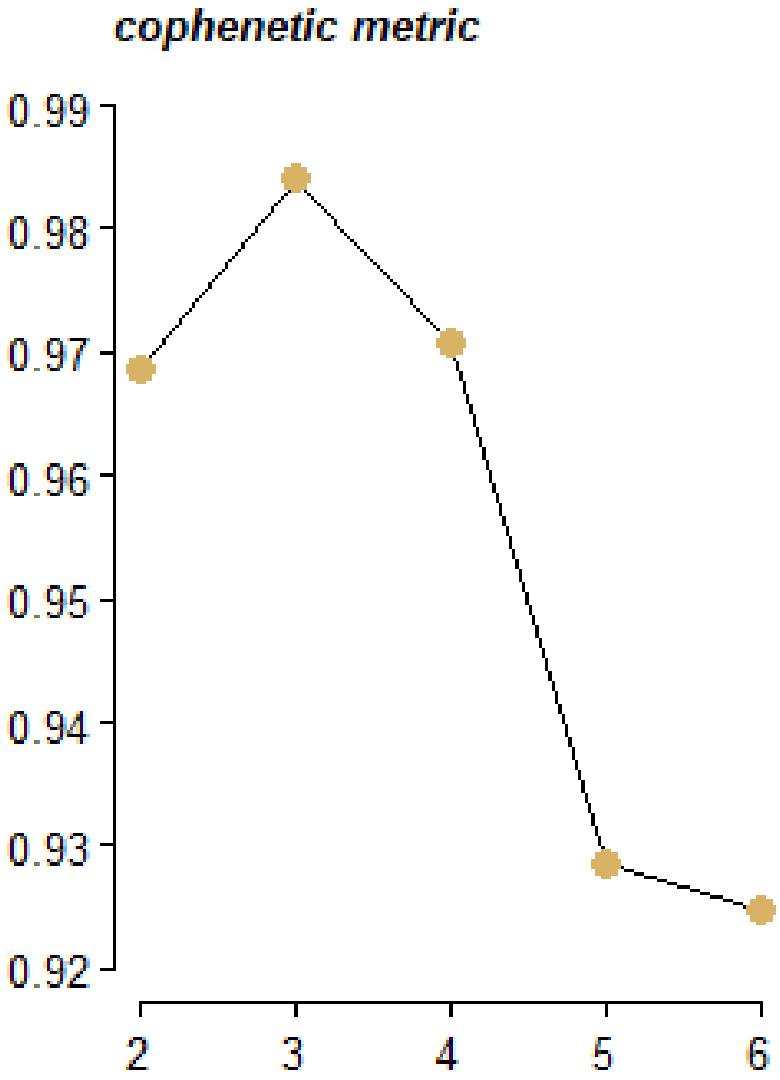

Figure 3.9: NMF Rank for the OSCC sample set (generated using Maftools on R-Studio).

The NMF algorithm identified the optimal number of mutational signatures by leveraging the cophenetic correlation coefficient (CPCC). On the elbow plot above (Figure 3.9), the point where the cophenetic correlation dropped significantly, coincided with the value, three (k = 3). This suggested that the number of mutational signatures for this sample set was three. The single-base substitution phenomenon was used to establish the signature profile of this dataset and the cosine similarity, an analysis used to measure the similarity between mutational signatures, used to test the accuracy of the signature analysis.

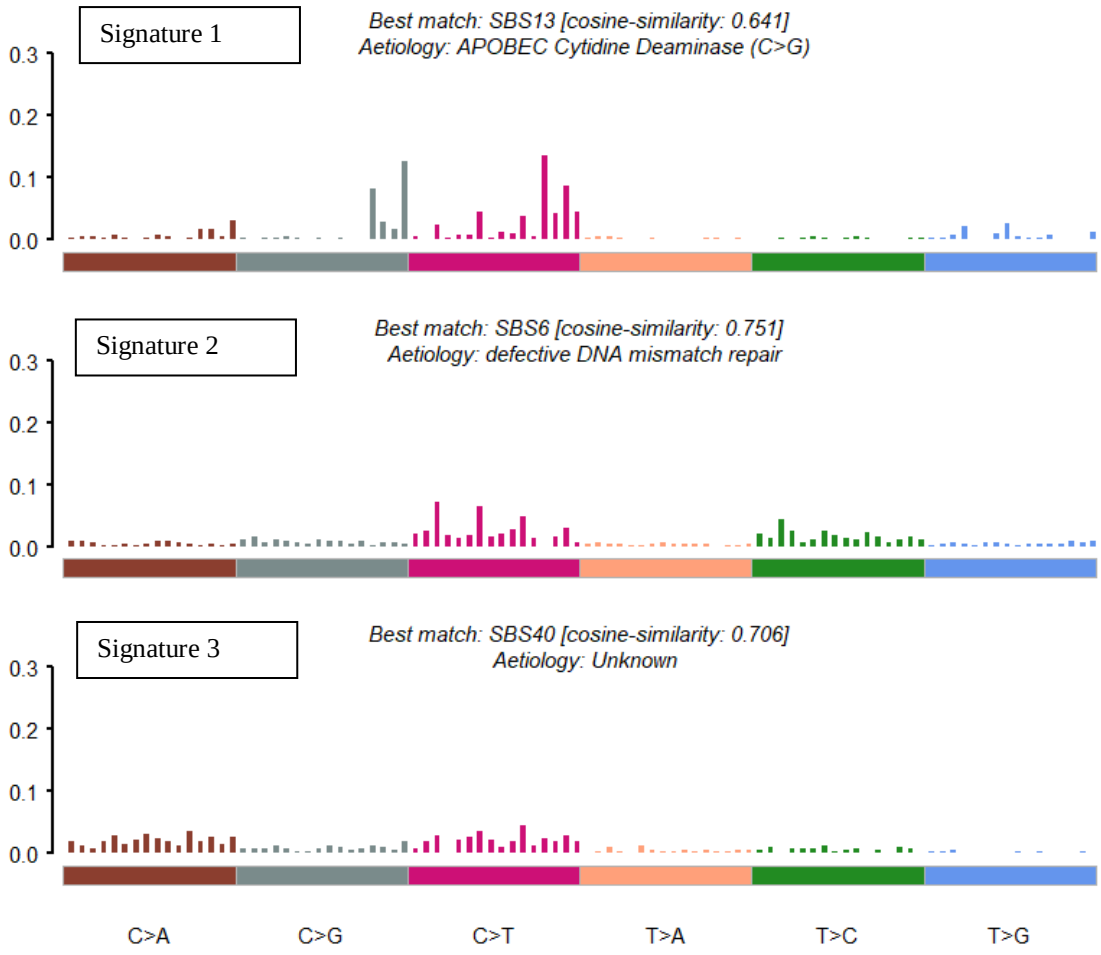

Figure 3.10: Signature profile of the three established mutational signatures. The probability bars for the six types of substitutions are coloured differently with the vertical axes showing the percentages of mutations attributed to a particular type of mutation. Signatures are based on the observed trinucleotide frequency of the human genome. This was achieved using the signature analysis tool of Maftools on R-Studio.

Signature 1 was characterised by the prominence of C>G and C>T substitutions with a cosine similarity of 0.641 and was best matched to signature SBS-13. The substitutions attributable to signatures 2 and 3 were predominantly C>T changes. The cosine similarity indexes were 0.751 and 0.706 respectively, thereby matching signature 2 to SBS-6 and signature 3 was linked to SBS-40. The mutational signature analysis assisted in establishing the mutational mechanisms responsible for the somatic mutations seen in OSCC.

The APOBEC-enrichment analysis was done to provide insights into the mutational processes influencing OSCC in this study and the results are illustrated below.

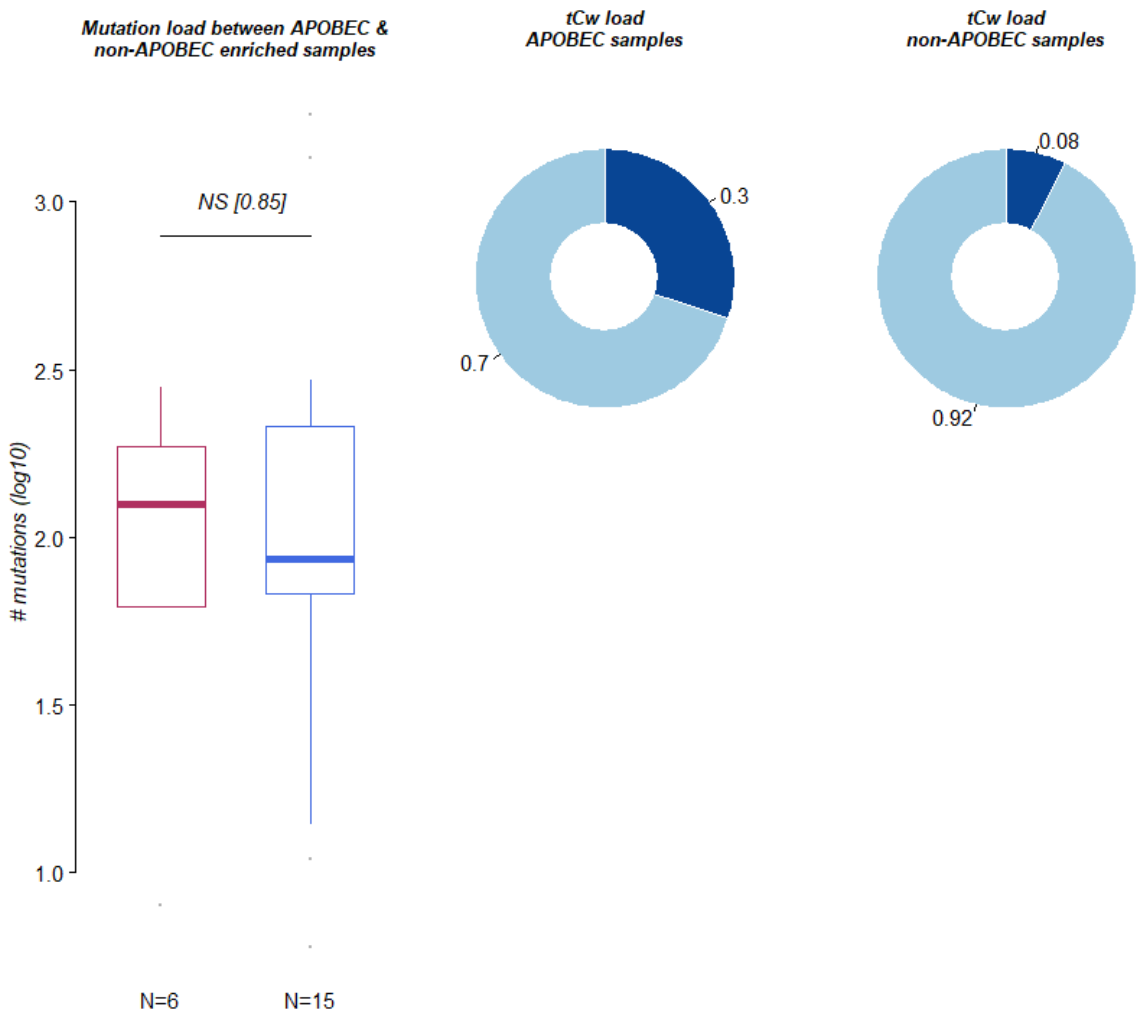

Figure 3.11: APOBEC-enrichment analysis. Out of the 21 samples, 6 were enriched (N=6) and 15 were not enriched (N=15).

A one-way Fisher's test was done for the APOBEC enrichment analysis and that it illustrated that approximately 29% of the samples had an APOBEC-related mutation and that the enrichment score was >2. This confirmed the presence of APOBEC-mediated mutational signatures in the tumour genome.

A molecular pathway analysis was carried out as described in Chapter 2 based on the somatic mutation data for the 21 OSCC tumours. The analysis revealed the number of genes that were affected in each pathway, as well as the number of samples from the dataset that were affected in each pathway (Figure 3.12).

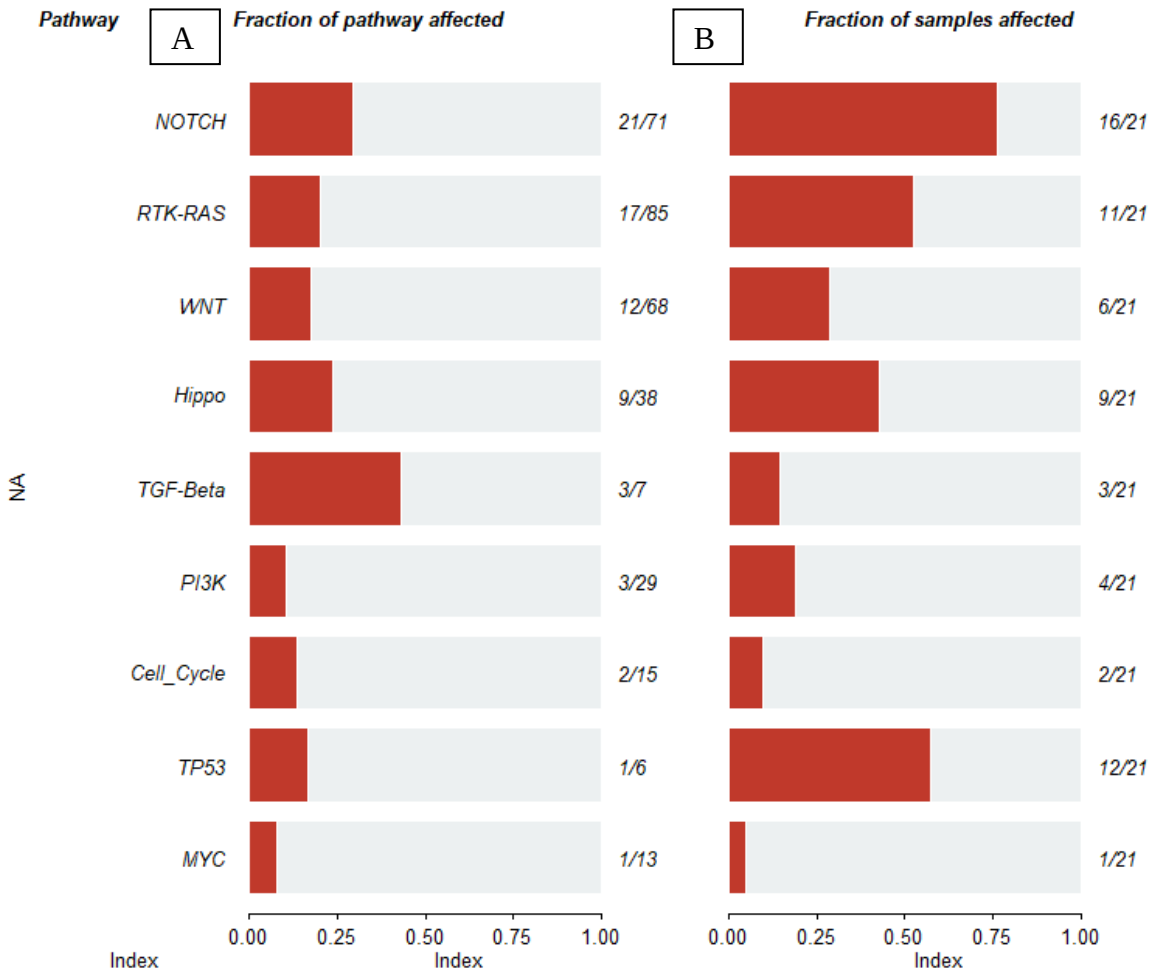

Figure 3.12: Molecular pathways (taken from the KEGG database) implicated in OSCC summarised as bar plots. The altered pathways are distributed according to the number of mutated genes in each pathway (left) and also according to the number of samples that were affected (right).

The bar plot on the left (labelled ‘A’) indicates how many genes were affected in at least one sample, whereas the bar plot on the right (labelled ‘B’) shows the total number of samples with mutations observed in a gene in that particular pathway. The mutation frequency of all the genes in the pathway is also illustrated. Plot A illustrates how many genes are affected in each pathway, for example, 3 out of 7 genes in the *TGF-Beta* pathway were affected, whereas Plot B, illustrates how many out of the 21 samples had some sort of mutation in a gene in a specific pathway (only 1 sample was affected by a gene in the *MYC* pathway). Placing emphasis on plot B (it serves the interests of this study more accurately), the pathways with the most number of samples mutated was found to be *NOTCH* (76%), *TP53* (57%) and *RTK-RAS* (52%). The

other pathways had relatively fewer samples affected [*Hippo* (43%); *WNT* (29%); *PI3K* (19%); *TGF-Beta* (14%); *Cell_cycle* (10%), and *MYC* (5%)] (Percentages are derived from fractions). Considering that *NOTCH, TP53* and *RTK-RAS* had the highest proportion of affected samples, a list of genes that could be mutated in these various pathways is shown below (Table 3.7).

Table 3.7: Genes mutated in the top 3 molecular pathways implicated in OSCC

Pathway		
NOTCH	***RTK-RAS***	***TP53***
CTBP2	NF1	TP53
NCOR1	RAPGEF1	
NOTCH1	ARAF	
SPEN	EGFR	
CIR1	ERRFI1	
DLK1	FGFR2	
DNER	INSR	
DTX2	KIT	
EP300	KSR1	
FBXW7	NTRK1	
JAG1	RAPGEF2	
MAML2	RASA3	
MAML3	RASGRP4	
MFAP2	ROS1	
NCOR2	SCRIB	
NOTCH2	SHC3	
NOTCH3	SOS1	
NOTCH4		
NUMBL		
RBPJL		
THBS2		
Total		
21	17	1

4 Discussion

This study aimed to identify the somatic mutation profiles of African patients with OSCC which might help to establish clinically useful molecular markers for the disease. This was accomplished by using the procedures described in the Methods section (Chapter 2), which led to the outcome displayed in the Results section (Chapter 3). This chapter will cover the project's overall outcomes and results by discussing the quality of the tumour tissue collected, and the significance of the clinical characteristics that were established. A discussion of the bioinformatic analysis outputs and the molecular characteristics of the tumours will be further elaborated upon, as well as addressing the project's constraints, potential future directions and overall conclusions to complete the chapter.

4.1 Purpose of study:

Oesophageal cancer is the sixth most common cause of cancer-related fatalities in South Africa, with a high incidence rate among men of African descent (13). The majority of patients present with a locally advanced or disseminated illness (66), thus necessitating the use of palliative rather than curative care for most patients. Multi-modal therapies including total endoscopic and/or surgical resection, chemotherapy and radiotherapy are the only current potentially curative courses of treatment (66). It is quite challenging to apply the treatment algorithms used in the western world to the South African setting since the genomic landscape of the disease is not generally well understood, and together treatment algorithms and results are not thoroughly documented (66). This gave rise to the focus of this study, which sought to understand the mutational patterns displayed in South African patients of African ancestry with OSCC which could lead to potentially useful molecular markers as well as the identification of treatment types that could be effective in African populations.

4.2 Clinical characteristics:

Regarding the demographic and clinical characteristics of the study participants, it can be noted that all except one of the participants in the study were of African ancestry, with a single participant of mixed ancestry. The recruitment of African ancestry patients was prioritised in order to build our limited knowledge of the underlying molecular basis of this disease in African populations. In the prospective study of fifteen patients from KwaZulu-Natal we prioritised female patients who had not used alcohol or tobacco products in order to assess

whether the molecular findings might be different to patients exposed to these common risk factors for OSCC.

Regarding age, the studies conducted in Malawi and Tanzania established the mean age at diagnosis to be 59 and 49 years old, respectively (16,40). Although, the mean age at diagnosis in our cohort was 63.8 years (~64), it is fairly consistent with that of the Malawian cohort as well as the pre-existing knowledge that OSCC is usually diagnosed at ages between 50-70 years.

In addition to ethnicity and age at diagnosis, the BMI, HIV status and tumour histology were also noted. A study done on the metabolic risk factors for OSCC and oesophageal adenocarcinoma (OAC) in patients from Austria, Norway and Sweden revealed that with regards to BMI in oesophageal cancer, a high BMI was associated with an increased risk of OAC and a decreased risk of OSCC. This correlates well with the different aetiology of these two different types of oesophageal cancer, with OAC being associated with reflux and obesity whereas many OSCC patients are very underweight at diagnosis owing to difficulty in swallowing their food. The mean BMI for our cohort was 22.4, but it ranged from 13.0-33.8, with several patients being severely underweight. The three participants that had BMI values that were considered to be obese (30.2, 33.8 and 33.2) presented with poorly differentiated squamous cell carcinoma. Of the patients of known HIV status, 14% were HIV-positive which is relatively consistent with the data from the Malawian study where 17% of that particular study cohort were HIV-positive (16).

With regards to the tumour histology, a significant portion (48%) of the 21 participants had an infiltrating keratinizing squamous cell carcinoma (IKSCC) which suggested that there was a higher risk of lymph node metastasis, an increased risk of local recurrence, poor prognosis and lower chance of survival in comparison to the proportion of the participants that exhibited a moderate differentiation (19%) and a poor differentiation (33%) (67). According the WHO classification for OSCC, well differentiated squamous cell carcinoma indicates that >75% of the tumour cells have some resemblance with the normal squamous cells, whilst moderately differentiated indicates that approximately 25-75% of the tumour cells resemble the normal squamous cells, and poorly differentiated squamous cells indicates that <25% of the tumour cells resemble the normal squamous cells (42). On the basis of the WHO classification, a significant proportion of our study participants had a very poor prognosis, suggesting that they

had an aggressive tumour which would mostly likely not respond well to any form of treatment interventions other than palliative care.

4.3 Tumour tissue quality control:

The integrity of genomic DNA (gDNA) is an important factor when defining DNA quality and carries the potential to have an impact on subsequent molecular applications (68). In order to ensure that the integrity of the resultant gDNA was not compromised, the tumour tissue samples were placed immediately into RNALater and stored in a -20°C freezer prior to DNA extraction. The other important factor was to ensure that the biopsies used for DNA extraction had at least 50% tumour cell content, which was assessed by histopathological examination of part of the biopsies. Research by our cancer team and Pathologist colleagues have shown that storage in RNALater at -20°C not only preserves DNA quality but also tissue morphology, which allows tumour content of the biopsies to be assessed (69).

During the course of the extraction of the gDNA, the samples were kept at 4°C to reduce the possibility of degradation. The Nanodrop was used to assess the purity of the DNA by using the ratio of absorbance at 260 and 280nm (70). However, given that the absorbance ratio could not discriminate between DNA and RNA, the possibility of the presence of RNA could lead to the ratio increasing and thus distorting the actual proportion of DNA present. DNA samples were therefore also analysed using the Qubit fluorometer and an assay that is highly selective for double-stranded DNA (dsDNA) over RNA. The sensitivity and specificity of the Qubit assay is deemed superior in comparison to the Nanodrop as it is more specific to their intended targets, thereby not allowing contaminants to interfere with the concentration of DNA – giving more accurate results (71).

When assessing the Nanodrop and Qubit results, they showed a relatively high consistency with one another. The average concentration values obtained from both of the samples were mostly within 10 – 20% of each other, indicative of a good agreement. The optimal A260/280 ratio for sequencing is 1.8 – 2.0, however ratios ranging from 1.7 – 2.2 tend to be accepted. The results showed that a few of our samples, namely, *LF159, LF224, LF225, LF284,* and *LF325* had absorbance ratios that were quite low (1,47; 0,97; 1,48; 1,26 and 1,59) suggesting severe to moderate contamination but not to the point where the samples were considered a fail as Novogene did conduct their own quality control and all the samples passed. The consistency in the results confirmed that the quality and integrity of the DNA samples was good, suggesting

a solid foundation for downstream applications. The final outcome of the qualitative and quantitative analysis carried by me and the Sanger Institute showed that all the DNA samples from the 36 matched pairs were suitable for submission to the Sanger whole exome DNA sequencing pipeline.

A further important aspect of quality control when sequencing matched blood/tumour pairs is to ensure that each of the two samples come from the same individual since potential mix-up of samples can occur in the clinic or in the laboratory. In the 36 matched pairs from which I extracted tumour DNA, the paired samples were genotyped by the Sanger Institute with a genome-wide SNP array, which showed that each pair did indeed come from the same individual.

We have received an interim report from the Sanger Institute on the results of WES of the 36 blood/tumour pairs. The top 10 mutated genes in these samples were: *TP53, TTN, MUC16, NOTCH1, PCLO, KMT2D, CSMD3, DNAH5, NFE2L2* and *SYNE1*, five of which (*TP53, TTN, PCLO, KMT2D* and *MUC16*) overlap with the top 22 genes that were mutated in the 21 blood/tumour pairs (Figure 3.4).

4.4 Quality and coverage of whole exome sequencing of the 21 blood/tumour pairs:

Key factors in the assessment of the DNA sequencing results from the 21 blood/tumour pairs were the accuracy of base calls, coverage of the target region and sequencing depth. Regarding calling accuracy, the Phred score of Q30 which equates to a base call accuracy of 99.9% was met in 86-94% of bases across all the samples, thus indicating high accuracy in the majority of the target regions sequenced. Coverage of the target region was >98% in 38 of the 42 samples, indicating that the majority of the target region had been sequenced. The sequencing depth is an indication of the number of times a given position in the genomic region (in this instance, exome) is sequenced. In this particular sample set, the fraction of the target covered with at least 10x was >90% in all but four of the samples, suggesting that the variants called at specific location within the exomes had a high probability of accuracy. The coverage histograms pointed towards the high range and uniformity in the depth and coverage across the genome. In summary, the data suggests that with a few exceptions good coverage of the target was achieved across the project.

4.5 Detection of somatic mutations in the 21 OSCC blood/tumour pairs:

The analysis of mutations in oesophageal squamous cell carcinoma tumour tissue showed that single nucleotide variants were much more common than indel variants, the majority of which were missense mutations, and that there was a very large variation in the number of mutations per tumour. The median number of non-synonymous variants that occurred per sample was 87, which is similar to the median number of 68 detected by Van Loon *et al.* (2023) in their recent report of somatic mutations in Tanzania (40).

Prior studies in other populations (72,73) have also demonstrated that missense mutations account for the bulk of somatic mutations in OSCC. However, a limitation of whole exome sequencing is that it will not detect variants in regulatory elements which can impact gene expression by means of distant control.

The comparison of the TMB of the OSCC tumours in our study with the TCGA dataset revealed that our samples had a TMB of 1.94 non-synonymous mutations per Mb, which was in the mid-range of TMBs compared to other cancer types and similar to the levels of TMBs seen in Uterine Corpus Endometrial Cancer and Ovarian Carcinoma. However, the level of TMB per Mb varied widely in our OSCC samples. TMB is emerging as a potentially important biomarker for cancer therapy. An elevated TMB corresponds to increased production of antigens by tumour cells, which can increase recognition of tumour cells by certain forms of immunotherapy and increased survival in a wide range of cancer types (74). Thus if immunotherapy were to become available to OSCC patients in Africa then analysis of TMB might become a clinically relevant biomarker.

The oncoplot in Figure 3.4 provided a representation of the top 22 mutated genes across the 21 samples. It established that other than the well-known OSCC gene *TP53*, several other genes were commonly mutated, including: *SPATA31D1, TTN, CLMN, USP17L8, NBPF1, OBSCN, PCLO, SERPINA3* and *KMT2D*. However, in comparison with other studies, only *TP53* and *KMT2D* are consistently mutated at similar levels (40,69,75–77). When compared to Western countries, Japan and China, *TP53* mutations in OSCC are much more common with frequencies of >90% in Asian OSCC patients. It has been suggested that variations in mutation frequencies are a consequence of ethnicity, although differences in exposure to environmental agents may also be involved (40). Some differences are seen in African studies done in Malawi and Tanzania with *TP53* mutated in 78% and 62% of patients, respectively, whereas in our sample set it was 57% (16,40). However, it should be noted that the sample sizes in the Malawian study

was 59, the Tanzanian study 61, whereas our cohort size was 21, which could account for the differences amongst the populations. *TP53* gene mutations in OSCC are associated with disruption of the p53 network's normal activities by suppressing immunological function, developing resistance to radiation and chemotherapy, as well as tumour formation and metastasis, thereby worsening a patient's prognosis (78). In addition to *TP53*, *KMT2D,* accounted for a significant portion of the nonsense mutations seen in this dataset and was mutated in 24% of our patients (Figure 3.4). *KMT2D* was mutated in 17% of Malawian OSCC patients (Liu *et al,* 2016), but was not reported to be significantly mutated in the Tanzanian study (40). This chromatin-modifying-enzyme may therefore be relevant to tumour development in a subset of African OSCC patients.

In order to assess the genetic heterogeneity of the tumours observed in this OSCC cohort, a variant allele frequency (VAF) assessment was performed. This metric measures the fraction of the sequencing reads that are inside a genomic location that support a particular variant allele as well as assist in differentiating between passenger and driver mutations (79). Driver mutations are responsible for providing cancer cells with a selective growth advantage and contribute directly to the onset and spread of the cancer whilst passenger mutations do not contribute directly as the genetic changes occur haphazardly (79).

The assessment of the VAF found that the highest values were in *TTN, MUC19* and *TP53,* which suggests that the mutations in these genes are more likely to be driver mutations. *SPATA31D1* and *SERPINA3* had intermediate VAF values, whereas *USP17L8, CLMN, OBSCN, NBPF1* and *PCLO* showed relatively low VAF values, suggesting that they are more likely to be passenger mutations. The biological relevance of the relatively high mutation frequency and VAF in the *TTN* gene is unclear. It is one of the longest coding genes known, thus offering a large target for non-consequential mutations, and is mutated in variety of cancers (80). Chen *et al.* (2020) found that a missense mutation in *TTN* contributed to the expression of lncRNA-TTN-AS1 which then stimulated the growth and metastasis of the OSCC cells (80). It encodes a muscle protein and is mutated in cardiomyopathies and is highly expressed in the heart, but is also expressed at relatively high levels in the oesophagus. Further research is required to investigate the role of this gene in the development of OSCC. *MUC19* is a gland-specific gel-forming mucin gene that is mostly expressed in mucous cells of several glands, but not in the oesophagus (80). Mutations in this gene were found in melanoma and are reportedly associated with poor prognosis in breast cancer, but the manner in which *MUC19* mutations influence OSCC development is still unclear and requires further exploration (80).

SPATA31D1 also displayed a relatively high clonal status, but it is only expressed in testes tissue and is predicted to be involved in cell differentiation and spermatogenesis, so seems unlikely to contribute significantly to the pathogenesis of OSCC (81). Like *TTN* it has a very long coding sequence of 4,731 nucleotides, thus presenting a large target for random mutations. Jung *et al.* (2022) proposed that it may be a possible driver gene in colorectal cancer, but did not explain the mechanism through which it influences disease development (82). *SERPINA3,* the last of the genes with a relatively high clonal status, is a serine protease inhibitor with multiple targets and given that it acts an acute-phase protein, it is released into the plasma by liver cells thereby allowing it to play a role in the anti-inflammatory and antiviral responses of the body (83). A review by Soman *et al.* (2022) revealed that it could serve as either a tumour promoter or suppressor in a variety of malignancies, however, the localization, probable effectors and mode of action are seemingly unknown (84).

In summary, with the exception of *TP53,* the biological and functional evidence that several of the genes frequently mutated in our samples contribute to the development of OSCC is not compelling, but warrants further investigation in a larger cohort of South African patients.

The analysis of the copy number alterations in our OSCC tumours established that three focal gains and nine focal deletions were present. The most significant of the three gains was amplification of chromosomes *11q23.3* and *21q22.2*. A gain in chromosome *11q23.3* is reportedly associated with *KMT2A* which is affiliated with an adverse risk for Acute Myeloid Leukemia, however the mechanisms that regulate its expression are unclear (85). A study by He *et al.* (2020) suggested that *KMT2A* is responsible for missense mutations observed in OSCC, however that was not observed in our study cohort (86). *ERG,* a proto-oncogene has been linked to the gain-of-function alteration in chromosome *21q22.2* (87). In prostate cancer, it is proposed to work in conjunction with p53 by driving β-catenin activation and pyrimidine synthesis, yet no mode of action has been identified for OSCC (87). Of the focal deletions present, the most significant of the nine deletions were those on chromosomes *5q31.2, 7q32.3* and *16q23.1*. They are proposed to harbour the genes, *LRRMT2, FOXP2* and *WWOX*, respectively (88–90). The first two are proposed to be involved in the development of intellectual disabilities whilst the latter (*WWOX*) is presumed to be involved in childhood acute lymphoblastic leukeumia (90). In the literature, there is currently no mention of these genes relating to OSCC and thus the mechanisms that drive their expressions are unclear. Further studies are warranted to understand how they may possibly exert their functions in OSCC.

The non-negative matrix factorization was employed in order to understand the mutational patterns exhibited in this cohort. This method estimated signatures by factorising mutational catalogues of observed counts of mutation types in the sample set and produced them in the form of a matrix (91). This algorithm used the cophenetic correlation coefficient (CPCC) to measure the similarity between the whole-exome data's hierarchical clustering structure and the clustering structure obtained by NMF (41). The calculation of the CPCC values across the various NMF ranks (k), determined that at the 'elbow point' (region where CPCC values is at its highest and begins to decrease) was the optimal number of mutational signatures (41). Given that single-base substitutions are the most investigated mutation types, a signature profile (SBS96) was then made on the basis of the NMF ranking with the accuracy of the analysis tested using the cosine similarity. This similarity index enabled the comparison and identification of similar signatures across datasets which helped facilitate the discovery of the underlying mutational processes (41). The 96 substitution classification visually illustrates each signature.

In this study, the first signature was attributed to C>G substitutions with a cosine similarity of 0.641 and was best matched to SBS-13. These mutations are likely to be produced by the response of the DNA repair machinery to cytidine deaminases of the APOBEC family, which convert cytidine to uracil. Error prone polymerases then replicate DNA across basic sites which have been generated by base excision repair removal of uracil. Signature 2 was as a result of C>T changes with a cosine similarity index of 0.706 which then matched it to SBS-6. This signature is associated with a pattern of indels (microsatellite instabilities) known to result from defective DNA mismatch repair. The substitutions attributable to signature 3 were also predominantly C>T changes. The cosine similarity index was 0.751, thereby matching it to SBS-40 which currently, according to the COSMIC database has no known aetiology. However a determination has been made that suggests that the mutations mostly correlated with the ages of the patients for some types of human cancers. Signatures 1 and 2 are consistent with the findings of the study conducted in Malawi as well as other studies done in China in that C>G and C>T substitutions and the cytidine deaminases of APOBEC invoke some form of mutagenic influence.

Regarding the APOBEC enrichment analysis, the sample set had an enrichment score of >2 with 29% of the samples having a mutation directly attributable to APOBEC activity. Although, the genes associated with enrichment could not be identified, the results of this analysis illustrate how APOBEC mutagenesis may be involved in OSCC progression (high

mutational burden) and tumour tissue transformation (advanced tumour stage, poor differentiation or lymph node metastasis) (92).

The pathway analysis revealed that the *NOTCH*, *RTK-RAS*, and *TP53* pathways were the significantly altered pathways involved in OSCC in this particular study. In the oesophagus, the NOTCH signalling pathway is involved in cellular functions relating to differentiation, proliferation and apoptosis. It encompasses the *NOTCH1/2/3* genes and many others which are proposed to be involved in the development of OSCC. A review into this pathway found that in addition to its role as a tumour suppressor, it also exerts an oncogenic function during carcinogenesis (93). Approximately 76% of the samples were mutated in at least one of the genes in this pathway, suggesting that it has an important role in OSCC.

The RTK-RAS pathway is involved in functions related to cell differentiation, cell cycle progression, cytoskeletal alteration and inhibition of cellular senescence as well as apoptosis. It encodes for a variety of genes however in this study, 52% of the samples were linked to this pathway.

It comes as no surprise that the samples are affected by an alteration in the *TP53* pathway. Seeing that *TP53* is one of the most commonly mutated genes in OSCC and is known to be responsible for tumour formation and metastasis, the pathway analysis confirmed the consistency of the involvement of *TP53* in pathogenesis.

4.6 Project limitations

There were significant limitations faced during the undertaking of this study. The start of the patient recruitment process was initially disrupted due to the fire that affected CMJAH which had resulted in the closing of various departments in the hospital, inclusive of the Oncology department, and as a result potential participants were then referred to other healthcare institutions and thus keeping track of these individuals was difficult. This was circumvented by adding CHBAH as a second recruitment location, however various permissions needed to be obtained which added a significant delay towards starting the collection process. After having being able to able to secure permissions for collections of samples from the second recruitment site, South Africa was still in the midst of COVID lockdown restrictions which meant that all work that the national government deemed non-essential in the healthcare field was put on hold which then delayed the start process even further.

In addition to the difficulties associated with patient recruitment, the consequent sample size was also a limiting factor, as well as funding for whole exome sequencing. Studies done with a small sample size tend to lack statistical power to identify robust associations. The consequence of the relatively small sample size of this study is that the results may not necessarily depict an accurate reflection of the molecular basis of OSCC in the local population.

The African continent is regarded as a low-resource setting due to the limited availability of resources and information. In South Africa, whole-exome sequencing is a developing sequencing technology, and therefore it is costly to perform and the existence of service providers who can perform this form of sequencing is limited, which then meant that the services of overseas companies, Novogene have to be utilized. The result of using their services was that the samples had to be shipped to the United Kingdom, with accompanying logistical issues such as required permits and having to wait for a significant amount of time before the data could be received. This added further time delays to the way in which the bioinformatic analysis took place.

In addition, we did not have the time or funding to carry out RNA sequencing, which would have provided additional insights into the pathways involved in tumour development. In recent times, studies that have incorporated this analysis have yielded effective results in that their ability to detect different types of cancers and rare diseases has shed a light on the ways in which adequate treatment can be refined. In other cancer studies, they have also proven to be useful in predicting diagnostic disease biomarkers which is desperately needed in OSCC of patients of African ancestry.

4.7 Conclusion and Future directions:

This study aimed to identify the somatic mutation profiles of African patients with OSCC and establish clinically useful molecular markers for this disease and possibly identify effective treatment types. The findings provided some insight into the somatic mutation profiles, however there is still a need for more information.

The limited amount of demographic and clinical information hindered the ability to yield significant results with regards to the association of the clinical characteristics with the most commonly mutated genes and CNVs. Therefore a reanalysis with a significantly larger sample size into the association between these parameters may be able to yield more statistically

significant results, thus, illustrating how clinical characteristics influence the genomic landscape.

The study identified *TP53* as the top commonly mutated gene, which is consistent with studies done in other African countries as well as Asia. Furthermore, the study introduced other commonly mutated genes (*SPATA31D1, TTN, MUC19, SERPINA3, CLMN, USP17L8, NBPF1, OBSCN* and *PCLO)* not frequently seen in other populations. Although the VAF gave rise to the possible driver genes involved in OSCC, apart from *TP53*, a well-known commonly mutated gene in OSCC, the rest of the genes mentioned have not been validated well enough to suggest that, they too, are definite driver genes. An analysis into the molecular mechanisms of each of the genes in the context of OSCC needs to be performed to ascertain the degree to which they influence the occurrence of the somatic mutations observed.

Together with the commonly mutated genes, the copy number alterations involved some amplifications and deletions not previously seen in other populations, thus contributing to the greater scheme of the molecular landscape of OSCC. Although, some genes of interest could be attributed to these changes, a detailed investigation of the genes within these regions should be established and then followed by an analysis into the manner in which they serve their function in OSCC. The fact that five of the top ten genes found to be mutated in the Sanger Institute study of our other 36 OSCC patients were also on our list of top twenty mutated genes is encouraging. A joint study combining all the data from patients from Johannesburg, Kwa-Zulu Natal and Professor Parker's study in the Western Cape is being planned which will provide much improved power to determine the major drivers of the development of OSCC in South African patients of African ancestry.

The mutational signature analysis identified three signatures which, as in previous studies, implicated the role of APOBEC enzyme activity in generating mutations on OSCC. A much larger study of mutation signatures in OSCC is warranted, together with a detailed analysis of potential association of these signatures with environmental exposures in patients.

In order to be able to increase the statistical power of the findings of this study, further investigation with an expanded sample size is warranted in order to circumvent possible biases. In being able to reproduce the mutational processes with a larger sample size, the findings of this study would be validated in that they would could be considered as accurately representing the genomic landscape of South African patients of African ancestry who are diagnosed with OSCC.

In conclusion, the study succeeded in establishing the somatic mutation profiles of African patients with oesophageal squamous cell carcinoma. In future, the identification of relevant molecular markers in local patients may serve as a guideline in discovering effective personalised treatment for this common and fatal disease.

5 References

1. Caspa Gokulan R, Garcia-Buitrago MT, Zaika AI. From genetics to signaling pathways: molecular pathogenesis of esophageal adenocarcinoma. Biochim Biophys Acta Rev Cancer. 2019 Aug;1872(1):37–48.

2. Yang T, Hui R, Nouws J, Sauler M, Zeng T, Wu Q. Untargeted metabolomics analysis of esophageal squamous cell cancer progression. J Transl Med. 2022 Dec;20(1):127.

3. Thun MJ, Linet MS, Cerhan JR, Haiman C, Schottenfeld D, editors. Schottenfeld and Fraumeni cancer epidemiology and prevention. Fourth edition. New York, NY: Oxford University Press; 2018. 1308 p.

4. Ness-Jensen E. Epidemiology and prevention of oesophageal adenocarcinoma. Scand J Gastroenterol. 2022 Aug 3;57(8):891–5.

5. Mazidimoradi A, Ghavidel F, Momenimovahed Z, Allahqoli L, Salehiniya H. Global incidence, mortality, and burden of esophageal cancer, and its correlation with SDI, metabolic risks, fasting plasma glucose, LDL cholesterol, and body mass index: An ecological study. Health Sci Rep. 2023 Jun;6(6):e1342.

6. Zhou B, Bie F, Zang R, Zhang M, Song P, Liu L, et al. Global burden and temporal trends in incidence and mortality of oesophageal cancer. J Adv Res. 2023 Aug;50:135–44.

7. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory. 2020 [cited 2023 Jul 25]. Global Cancer Observatory: Cancer Today. Lyon, France. International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today/

8. He H, Chen N, Hou Y, Wang Z, Zhang Y, Zhang G, et al. Trends in the incidence and survival of patients with esophageal cancer: A SEER database analysis. Thorac Cancer. 2020 May;11(5):1121–8.

9. Van Loon K, Mwachiro MM, Abnet CC, Akoko L, Assefa M, Burgert SL, et al. The African Esophageal Cancer Consortium: A Call to Action. J Glob Oncol. 2018 Dec;(4):1–9.

10. Mangalaparthi KK, Patel K, Khan AA, Manoharan M, Karunakaran C, Murugan S, et al. Mutational Landscape of Esophageal Squamous Cell Carcinoma in an Indian Cohort. Front Oncol. 2020 Aug 20;10:1457.

11. Niu C, Liu Y, Wang J, Liu Y, Zhang S, Zhang Y, et al. Risk factors for esophageal squamous cell carcinoma and its histological precursor lesions in China: a multicenter cross-sectional study. BMC Cancer. 2021 Dec;21(1):1034.

12. Simba H, Kuivaniemi H, Lutje V, Tromp G, Sewram V. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. Front Genet. 2019 Aug 2;10:642.

13. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Meryl L, et al. Global Cancer Observatory. 2024 [cited 2024 Sep 14]. Global Cancer Observatory: Cancer Today. Lyon,

France. International Agency for Research on Cancer. Available from: https://gco.iarc.who.int/today

14. Kachala R. Systematic review: epidemiology of Oesophageal Cancer in SubSaharan Africa. Malawi Med J [Internet]. 2010 Sep 18 [cited 2024 Sep 14];22(3). Available from: https://www.ajol.info/index.php/mmj/article/view/62190

15. Ferndale L, Aldous C, Hift R, Thomson S. Gender Differences in Oesophageal Squamous Cell Carcinoma in a South African Tertiary Hospital. Int J Environ Res Public Health. 2020 Sep 28;17(19):7086.

16. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight [Internet]. 2016 Oct 6 [cited 2021 Jun 1];1(16). Available from: https://insight.jci.org/articles/view/88755

17. Chen WC, Bye H, Matejcic M, Amar A, Govender D, Khew YW, et al. Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. Carcinogenesis. 2019 Jun 10;40(4):513–20.

18. Brown J, Stepien AJ, Willem P. Landscape of copy number aberrations in esophageal squamous cell carcinoma from a high endemic region of South Africa. BMC Cancer. 2020 Dec;20(1):281.

19. Abnet CC, Arnold M, Wei WQ. Epidemiology of Esophageal Squamous Cell Carcinoma. Gastroenterology. 2018 Jan;154(2):360–73.

20. Simba H, Kuivaniemi H, Abnet CC, Tromp G, Sewram V. Environmental and life-style risk factors for esophageal squamous cell carcinoma in Africa: a systematic review and meta-analysis. BMC Public Health. 2023 Sep 14;23(1):1782.

21. Alaouna M, Hull R, Penny C, Dlamini Z. Esophageal cancer genetics in South Africa. Clin Exp Gastroenterol. 2019 Apr;Volume 12:157–77.

22. Come J, Pereira JB, Pinto R, Carrilho C, Pereira L, Lara Santos L. The Upper Digestive Tract Microbiome and Oesophageal Squamous Cell Carcinoma: Epidemiology, Pathogenesis, and Clinical Implications in Africa. Pathobiology. 2021;88(2):141–55.

23. Yang CS, Chen XL. Research on esophageal cancer: With personal perspectives from studies in China and Kenya. Int J Cancer. 2021 Jul 15;149(2):264–76.

24. Chen HY, Chen IC, Chen YH, Chen CC, Chuang CY, Lin CH. The Influence of Socioeconomic Status on Esophageal Cancer in Taiwan: A Population-Based Study. J Pers Med. 2022 Apr 7;12(4):595.

25. Mwachiro MM, Pritchett N, Calafat AM, Parker RK, Lando JO, Murphy G, et al. Indoor wood combustion, carcinogenic exposure and esophageal cancer in southwest Kenya. Environ Int. 2021 Jul;152:106485.

26. Akbari MR, Malekzadeh R, Lepage P, Roquis D, Sadjadi AR, Aghcheli K, et al. Mutations in Fanconi anemia genes and the risk of esophageal cancer. Hum Genet. 2011 May;129(5):573–82.

27. Bhandari J, Thada PK, Puckett Y. Fanconi Anemia. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 [cited 2023 Dec 4]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK559133/

28. Rosenberg PS, Greene MH, Alter BP. Cancer incidence in persons with Fanconi anemia. Blood. 2003 Feb 1;101(3):822–6.

29. Zeng B, Huang P, Du P, Sun X, Huang X, Fang X, et al. Comprehensive Study of Germline Mutations and Double-Hit Events in Esophageal Squamous Cell Cancer. Front Oncol. 2021 Apr 6;11:637431.

30. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019 Aug;20(8):467–84.

31. Wu C, Wang Z, Song X, Feng XS, Abnet CC, He J, et al. Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nat Genet. 2014 Sep;46(9):1001–6.

32. Chen WC, Brandenburg JT, Choudhury A, Hayat M, Sengupta D, Swiel Y, et al. Genome-wide association study of esophageal squamous cell cancer identifies shared and distinct risk variants in African and Chinese populations. Am J Hum Genet. 2023 Oct 5;110(10):1690–703.

33. Ramarao-Milne P, Kondrashova O, Patch AM, Nones K, Koufariotis LT, Newell F, et al. Comparison of actionable events detected in cancer genomes by whole-genome sequencing, in silico whole-exome and mutation panels. ESMO Open. 2022 Aug;7(4):100540.

34. Goh G, Choi M. Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases. Genomics Inform. 2012;10(4):214.

35. Burdick KJ, Cogan JD, Rives LC, Robertson AK, Koziura ME, Brokamp E, et al. Limitations of exome sequencing in detecting rare and undiagnosed diseases. Am J Med Genet A. 2020 Jun;182(6):1400–6.

36. Mazzarotto F, Olivotto I, Walsh R. Advantages and Perils of Clinical Whole-Exome and Whole-Genome Sequencing in Cardiomyopathy. Cardiovasc Drugs Ther. 2020 Apr;34(2):241–53.

37. Ohashi S, Miyamoto S, Kikuchi O, Goto T, Amanuma Y, Muto M. Recent Advances From Basic and Clinical Studies of Esophageal Squamous Cell Carcinoma. Gastroenterology. 2015 Dec;149(7):1700–15.

38. Bethune G, Bethune D, Ridgway N, Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. J Thorac Dis. 2010 Mar;2(1):48–51.

39. Zhang X, Wang Y, Meng L. Comparative genomic analysis of esophageal squamous cell carcinoma and adenocarcinoma: New opportunities towards molecularly targeted therapy. Acta Pharm Sin B. 2022 Mar;12(3):1054–67.

40. Van Loon K, Mmbaga EJ, Mushi BP, Selekwa M, Mwanga A, Akoko LO, et al. A Genomic Analysis of Esophageal Squamous Cell Carcinoma in Eastern Africa. Cancer Epidemiol Biomarkers Prev. 2023 Oct 2;32(10):1411–20.

41. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb 6;578(7793):94–101.

42. WHO Classification of Tumours Editorial. Digestive System Tumours [Internet]. [cited 2024 Sep 21]. Available from: https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/Digestive-System-Tumours-2019

43. Yang Y, Zhu L, Cheng Y, Liu Z, Cai X, Shao J, et al. Three-arm phase II trial comparing camrelizumab plus chemotherapy versus camrelizumab plus chemoradiation versus chemoradiation as preoperative treatment for locally advanced esophageal squamous cell carcinoma (NICE-2 Study). BMC Cancer. 2022 May 6;22(1):506.

44. Abelson S, Anderson N, Bloomfield C, Brown B, Chen F, Chiappinelli K, et al. Cancer Genomics 2023: Advances in Cancer Research, Precision Oncology and Clinical Care [Internet]. Winchester, Hampshire, United Kingdom; 2023 Nov [cited 2023 Nov 29]. Available from: https://frontlinegenomics.com/cancer-genomics-2023-advances-in-cancer-research-precision-oncology-and-clinical-care/

45. Guo XF, Li SS, Zhu XF, Dou QH, Liu D. Lapatinib in combination with paclitaxel plays synergistic antitumor effects on esophageal squamous cancer. Cancer Chemother Pharmacol. 2018 Sep;82(3):383–94.

46. Lu M, Wang X, Shen L, Jia J, Gong J, Li J, et al. Nimotuzumab plus paclitaxel and cisplatin as the first line treatment for advanced esophageal squamous cell cancer: A single centre prospective phase II trial. Cancer Sci. 2016;107(4):486–90.

47. Xin Z, Song X, Jiang B, Gongsun X, Song L, Qin Q, et al. Blocking FGFR4 exerts distinct anti-tumorigenic effects in esophageal squamous cell carcinoma. Thorac Cancer. 2018;9(12):1687–98.

48. Testa U, Castelli G, Pelosi E. The Molecular Characterization of Genetic Abnormalities in Esophageal Squamous Cell Carcinoma May Foster the Development of Targeted Therapies. Curr Oncol. 2023 Jan;30(1):610–40.

49. Liu J, Pan S, Hsieh MH, Ng N, Sun F, Wang T, et al. Targeting Wnt-driven cancer through the inhibition of Porcupine by LGK974. Proc Natl Acad Sci. 2013 Dec 10;110(50):20224–9.

50. Chen WC, Kerr R, May A, Ndlovu B, Sobalisa A, Duze ST, et al. The Integrity and Yield of Genomic DNA Isolated from Whole Blood Following Long-Term Storage at −30°C. Biopreservation Biobanking. 2018 Apr;16(2):106–13.

51. Ferndale L, Moodley M, Chen WC, Wadee R, Wright CA, Parker MI, et al. Processing and Analysis of Tissue Samples from Esophageal Cancer Patients in an African Setting. Biopreservation Biobanking. 2022 Apr 1;20(2):185–94.

52. Abdel-Latif A, Osman G. Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize. Plant Methods. 2017 Dec;13(1):1.

53. Genohub. Genohub. 2018 [cited 2023 May 1]. Sample Storage and Shipping Conditions for Next Generation Sequencing. Available from: https://genohub.com/dna-rna-shipping-for-ngs/

54. Broad Institute. Genome Analysis Toolkit. 2024 [cited 2024 Sep 5]. GATK. Available from: https://gatk.broadinstitute.org/hc/en-us

55. Broad Institute. GATK. 2024 [cited 2024 Sep 5]. Mutect2. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

56. Cibulskis K, Lawrence MS, Carter S, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogenous cancer samples. Nat Biotechnol. 2013 Feb 10;31(3):213–9.

57. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018 Jul 16;(15):591–4.

58. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012 Feb 3;28(3):423–5.

59. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Jul 3;38(16):e164: 1-7.

60. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018 Oct 19;(28):1747–56.

61. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proc Natl Acad Sci. 2007 Dec 11;104(50):20007–12.

62. Bionano Genomics. Identifying Statistically Significant Frequent Copy Number Alterations: STAC and GISTIC [Internet]. Identifying Statistically Significant Frequent Copy Number Alterations: STAC and GISTIC. 2016 [cited 2024 Feb 24]. Available from: https://bionano.com/blogs/identifying-statistically-significant-frequent-copy-number-alterations/

63. Yu N, Gao YL, Liu JX, Wang J, Shang J. Robust Hypergraph Regularized Non-Negative Matrix Factorization for Sample Clustering and Feature Selection. IEEEE Trans Cybern. 2019 Oct 22;49(4):1155–116.

64. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, et al., editors. AJCC Cancer Staging Manual [Internet]. 8th ed. American College of Surgeons: Springer Cham; 2016 [cited 2024 Sep 5]. XVII, 1032. Available from: https://link.springer.com/book/9783319406176

65. NCCN Guidelines for Patients: Esophageal Cancer. Esophageal Cancer [Internet]. 2024; Available from: https://www.nccn.org/patients/guidelines/content/PDF/esophageal-patient.pdf

66. Loots E, Sartorius B, Madiba TE, Mulder CJJ, Clarke DL. Is Clinical Research in Oesophageal Cancer in South Africa in Crisis? A Systematic Review. World J Surg. 2017 Mar;41(3):810–6.

67. Lam AK. Histopathological Assessment for Esophageal Squamous Cell Carcinoma. In: Lam AK, editor. Esophageal Squamous Cell Carcinoma: Methods and Protocols [Internet]. New York, NY: Springer US; 2020 [cited 2024 Sep 18]. p. 7–18. Available from: https://doi.org/10.1007/978-1-0716-0377-2_2

68. Malentacchi F, Ciniselli CM, Pazzagli M, Verderio P, Barraud L, Hartmann CC, et al. Influence of pre-analytical procedures on genomic DNA integrity in blood samples: The SPIDIA experience. Clin Chim Acta. 2015 Feb;440:205–10.

69. Lin DC, Hao JJ, Nagata Y, Xu L, Shang L, Meng X, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. Nat Genet. 2014 May;46(5):467–73.

70. Lucena-Aguilar G, Sánchez-López AM, Barberán-Aceituno C, Carrillo-Ávila JA, López-Guerrero JA, Aguilar-Quesada R. DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis. Biopreservation Biobanking. 2016 Aug;14(4):264–70.

71. Bhat S, Curach N, Mostyn T, Singh Bains G, Griffiths KR, Emslie KR. Comparison of methods for accurate quantification of DNA mass concentration with traceability to the international system of units. Anal Chem. 2010 Aug 6;82(17):7185–92.

72. Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, et al. Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. Nat Commun. 2017 Nov 16;8(1):1533.

73. He W, Leng X, Yang Y, Peng L, Shao Y, Li X, et al. Genetic Heterogeneity of Esophageal Squamous Cell Carcinoma with Inherited Family History. OncoTargets Ther. 2020 Sep;Volume 13:8795–802.

74. Cao D, Xu H, Xu X, Guo T, Ge W. High tumor mutation burden predicts better efficacy of immunotherapy: a pooled analysis of 103078 cancer patients. OncoImmunology. 2019 Sep 2;8(9):e1629258.

75. Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, Sun ZM, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet. 2014 Oct;46(10):1097–102.

76. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, et al. Identification of genomic alterations in oesophageal squamous cell cancer. Nature. 2014 May 1;509(7498):91–5.

77. Sawada G, Niida A, Uchi R, Hirata H, Shimamura T, Suzuki Y, et al. Genomic Landscape of Esophageal Squamous Cell Carcinoma in a Japanese Population. Gastroenterology. 2016 May;150(5):1171–82.

78. Zhong L, Li H, Chang W, Ao Y, Wen Z, Chen Y. TP53 Mutations in Esophageal Squamous Cell Carcinoma. Front Biosci Landmark Ed. 2023 Sep 24;28(9):219.

79. Boscolo Bielo L, Trapani D, Repetto M, Crimini E, Valenza C, Belli C, et al. Variant allele frequency: a decision-making tool in precision oncology? Trends Cancer. 2023 Dec;9(12):1058–68.

80. Chen Z, Yao N, Zhang S, Song Y, Shao Q, Gu H, et al. Identification of critical radioresistance genes in esophageal squamous cell carcinoma by whole-exome sequencing. Ann Transl Med. 2020 Aug;8(16):998–998.

81. The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017 Jan;541(7636):169–75.

82. Jung S, Lee JL, Kim TW, Lee J, Yoon YS, Lee KY, et al. Molecular characterization of dysplasia-initiated colorectal cancer with assessing matched tumor and dysplasia samples. Ann Coloproctology. 2022 Feb 28;38(1):72–81.

83. De Mezer M, Rogaliński J, Przewoźny S, Chojnicki M, Niepolski L, Sobieska M, et al. SERPINA3: Stimulator or Inhibitor of Pathological Changes. Biomedicines. 2023 Jan 7;11(1):156.

84. Soman A, Asha Nair S. Unfolding the cascade of SERPINA3: Inflammation to cancer. Biochim Biophys Acta BBA - Rev Cancer. 2022 Sep;1877(5):188760.

85. Tsai HK, Gibson CJ, Murdock HM, Davineni P, Harris MH, Wang ES, et al. Allelic complexity of *KMT2A* partial tandem duplications in acute myeloid leukemia and myelodysplastic syndromes. Blood Adv. 2022 Jul 26;6(14):4236–40.

86. He W, Leng X, Wang K, Mao T, Peng L, Fang Q, et al. Genomic characteristics in neoadjuvant chemoradiotherapy for locally advanced esophageal squamous cell carcinoma. J Gastrointest Oncol. 2020 Dec;11(6):1105–12.

87. Ding D, Blee AM, Zhang J, Pan Y, Becker NA, Maher LJ, et al. Gain-of-function mutant p53 together with ERG proto-oncogene drive prostate cancer by beta-catenin activation and pyrimidine synthesis. Nat Commun. 2023 Aug 3;14(1):4671.

88. Kleffmann W, Zink AM, Lee JA, Senderek J, Mangold E, Moog U, et al. 5q31 Microdeletions: Definition of a Critical Region and Analysis of LRRTM2, a Candidate Gene for Intellectual Disability. Mol Syndromol. 2012 Aug;3(2):68–75.

89. Jiménez-Romero MS, Barcos-Martínez M, Espejo-Portero I, Benítez-Burraco A. Language Impairment Resulting from a de novo Deletion of 7q32.1q33. Mol Syndromol. 2016 Oct;7(5):292–8.

90. Shi Y, Du M, Fang Y, Tong N, Zhai X, Sheng X, et al. Identification of a novel susceptibility locus at 16q23.1 associated with childhood acute lymphoblastic leukemia in Han Chinese. Hum Mol Genet. 2016 Apr 19;ddw112.

91. Pelizzola M, Laursen R, Hobolth A. Model selection and robust inference of mutational signatures using Negative Binomial non-negative matrix factorization. BMC Bioinformatics. 2023 May 8;24(1):187.

92. Zhang SQ, Zhang J, Yu Y, Yu MM, Wei J, Tang YH. APOBEC3B expression has prognostic significance in cervical cancer. Int J Clin Exp Pathol. 2023 Mar 15;16(3):48–56.

93. Li Y, Li Y, Chen X. NOTCH and Esophageal Squamous Cell Carcinoma. In: Reichrath J, Reichrath S, editors. Notch Signaling in Embryology and Cancer [Internet]. Cham: Springer International Publishing; 2021 [cited 2024 Mar 17]. p. 59–68. (Advances in Experimental Medicine and Biology; vol. 1287). Available from: https://link.springer.com/10.1007/978-3-030-55031-8_5

Appendices

Appendix 1 – Patient information and consent form

UNIVERSITY OF THE WITWATERSRAND & NATIONAL CANCER REGISTRY, NHLS

OESOPHAGEAL CANCER RESEARCH STUDY

PARTICIPANT INFORMATION FORM (CASE)

VERSION 2.3, Dated 11 December 2020
SPONSORS: South African Medical Research Council and CANSA
PRINCIPAL INVESTIGATORS: C Mathew, WC Chen, M Ramsay, E Singh

---

INTRODUCTION

Good day. We are members of the oesophageal cancer research team from the University of the Witwatersrand and the National Cancer Registry, National Health Laboratory Service (NHLS) in Johannesburg.

We are working on a project aimed at identifying the genetic factors which are important in the development cancer of the food pipe, the oesophagus. Our aim is to discover genetic markers which increase the risk of this disease, and to develop sensitive diagnostic tests to improve the early diagnosis of oesophageal cancer.

We are leading an international research project to achieve these aims, and we would like to invite you to participate in this important research study.

WHAT DO I HAVE TO DO IF I AM IN THIS STUDY?

If you agree to participate in this study we will ask you to provide a small sample of saliva (spit) and about 10 ml (2 table-spoons) of blood which will be taken by a nurse or doctor. We

also ask you provide a small biopsy sample of a potential tumour that will be extracted (by the doctor) during the routine scoping procedure. We will ask you some routine questions about your home language, education, income, how you cook your food, HIV status and your general health. We will also ask about possible exposure to substances in the environment and your habits such as smoking and drinking hot or alcoholic liquids which may play a role in oesophageal disease. This should take no longer than 30 minutes at most.

WHAT DOES THE STUDY INVOLVE?

We will extract the genetic material known as nucleic acid (DNA) from your samples, which will be labelled with a study number (not your name). We will analyse your DNA using various techniques which may include whole genome sequencing to look for variants in your genetic code which may increase the risk of oesophageal cancer. Your anonymous sample may be stored by the project for an indefinite period as DNA analysis technology develops, and part of your anonymous DNA sample may be sent to reputable laboratories elsewhere in South Africa or overseas for analysis. We will also ask for confirmation of your diagnosis from the Division of Anatomical Pathology, who examine under a microscope the biopsy taken from you as part of your routine examination. At the end of the project, the anonymised results will be stored indefinitely in an electronic archive which may be accessible to qualified researchers worldwide to compare your data with those from other patients.

WHAT ARE THE BENEFITS OF BEING IN THIS STUDY?

There is no direct immediate benefit to you. Your participation will contribute to developing greater knowledge of oesophageal cancer and hopefully to the development of new diagnostic tests and improved treatments.

WHAT ARE THE RISKS OF BEING IN THIS STUDY?

You may feel discomfort or slight pain during the blood collection. You may have bruising at the site where the needle is inserted. Some people may feel light-headed during blood collection. The tumour sample collection will be done at the same time as your routine gastroscopy and biopsy procedure so you will not be subjected to additional discomforts.

WHAT ARE MY RIGHTS AS A PARTICIPANT IN THE STUDY?

Your participation is entirely voluntary, and you can decline to participate or stop at any time without stating a reason. There will be no consequences if you decide not to take part in the study and you will still receive the best treatment available.

You also have the right to withdraw from the study at any time by calling the numbers below and your samples would then be destroyed.

COST AND PAYMENT

There is neither cost nor payment involved in participating in this study.

WHAT ABOUT CONFIDENTIALITY?

Confidentiality will be ensured by assigning a unique study number to your blood and saliva samples and questionnaire, which will not be labelled with your name. The link between your study number and your personal details will be restricted to qualified staff in the NHLS National Cancer Registry and held in a secure database at the registry.

QUESTIONS

You are free to ask any questions about this study and discuss any concerns you may have with the research staff.
Our contacts are as follow:

Research Study team: 011 885 5305

Email: WenlongC@nicd.ac.za
Human Research Ethics Committee: 011 717 1234 or 011 717 1252

ETHICS CLEARANCE

This study has been approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, Johannesburg ("Committee"). A principal function of this Committee is to safeguard the rights and dignity of all human subjects who agree to participate in a research project and the integrity of the research.

If you have any concern over the way the study is being conducted, please contact the Chairperson of this Committee who is Dr Clement Penny, who may be contacted on telephone number 011 717 2301, or by email on Clement.Penny@wits.ac.za. The telephone numbers for the Committee secretariat are 011 717 2700/1234 and the e-mail addresses are Zanele.Ndlovu@wits.ac.za and Rhulani.Mukansi@wits.ac.za

Thank you for reading this Study Information Sheet.

December 2020

V2.3. 11.12.2020

STATEMENT OF CONSENT AND SIGNATURE

VOLUNTARY INFORMED CONSENT TO PARTICIPATE IN THE WITS-NHLS OESOPHAGEAL CANCER RESEARCH STUDY

1. I have been given a Participant Information Sheet which explains the nature and processes involved in this study, which is attached hereto;
2. I was given time to read it, or had it read to me, in the language I best understand;
3. I was given time to ask any questions I wanted to and found any answers given to me to be reasonable and satisfactory;
4. I believe I fully understand why the study is being conducted and what the intended outcomes will be;
5. I understand that there will be no immediate benefit to me, should I agree to participate, nor will I receive any payment; conversely, participation will not cost me anything but my time;
6. I understand that, even if I initially consent to take part in the study, I may subsequently withdraw at any time and would not be required to give any reasons; if that happened, any data collected about me for the purposes of the study would immediately be destroyed, unless I give consent for it to be retained
7. I have been given a range of contact details, listed below. If I require further information or become concerned about any aspect of this study, I am free to speak to any of these contacts.

DECLARATION

I have received, read and understood the above information regarding this research study. I have had the opportunity to ask questions and received adequate responses.

I have fully understood this consent form.

I have no further questions and am prepared to participate in the study

Contact details:

Carl Chen, Research coordinator, telephone no. 0118855305, or by e-mail at wenlongc@nicd.ac.za,

Professor CB Penny, Chairperson of the Human Research Ethics Committee (Medical) at the University of

Witwatersrand, on telephone no. 011 717 2301, or by e-mail at Clement.Penny@wits.ac.za.

Ms. Z Ndlovu or Mr Rhulani Mkansi, Committee Secretariat, telephone nos.: 011 717 2700 or 1234, or by e-mail at: Zanele.Ndlovu@wits.ac.za or Rhulani.Mkansi@wits.ac.za

_____ _____

Participant Name and Surname (print) Participant signature or mark and date

_____ _____

Witness Name and Surname (print) Witness signature or mark and date

V2, v11122020

Retain original informed consent form on file. Offer participant a copy.

Appendix 2 – Questionnaire

R: SAMRC CRCEC Genetics Questionnaire

Event Name: Research

Database ID

Record ID

OESOPHAGEAL CANCER STUDY QUESTIONNAIRE FOR PATIENTS (WITS-SAMRC)

Disease Diagnosis:

Type of Cancer:	

Participant number:		Hospital number:	
Date of interview:	Y-M-D	Interviewer:	
Pathology/Episode number:		Stage:	
Clinical Diagnosis:		Histology result:	
Upload Pathology report (1):	Upload le	Upload Pathology report (2):	Upload le

Demographics:

Date of birth:	Y-M-D	Age in years:	
Gender:		Race group:	

Place of Birth & Residence:

In which province were you born?		Is the area of birth Urban or Rural or Peri-urban?	
Place of birth - other specify:			

How long have you lived in each province:

	Gauteng	Limpopo	Mpumalanga	North West	Free State	Eastern Cape	Western Cape	Northern Cape	Kwazulu-Natal	Other
In years:										

Is the area you live/ lived in Urban or Rural or Peri-urban?	

Education:

Yourself:

What is the highest grade you passed at school/university?	

Your Spouse/ Partner:

Not married?	(1) Yes
What is the highest grade your pouse/ partner passed at school/university?	

Income:

What is your monthly income?	

Home Language:

What language do you speak the most at home?		Other - specify:	
What language do you speak second-most at home?		Other - specify:	
What language does/did your father speak the most at home?		Other - specify:	
What language does/did your father speak the second-most at home?		Other - specify:	
What language does/did your mother speak the most at home?		Other - specify:	
What language does/did your mother speak the second most at home?		Other - specify:	

Fuel Usage:

For current cooking & warmth:

Where do you normally cook food?			
What type of fuel do you mostly use in your home for cooking?		Other fuel cook - specify:	
What type of fuel do you mostly use in your home for keeping warm?		Other fuel warmth specify:	

For 20 years ago cooking & warmth:

Where did you (or your family) normally cook food 20 years ago?			
What type of fuel did you mostly use in your home for cooking 20 years ago?		Other fuel cook - specify:	
What type of fuel did you mostly use in your home for keeping warm 20 years ago?		Other fuel warmth - specify:	

Risk Factors:

Have you ever smoked cigarettes or a pipe regularly?	

If 'Yes' (or 'Yes, in the past'), how many cigarettes/pipe/hand-rolled cigarettes would you usually smoke in a day?

Number of cigarettes:	Number of pipes:	Number of hand-rolled cigarettes:

If 'Yes' (or 'Yes, in the past'), how old were you when you first started smoking cigarettes/pipe/hand-rolled cigarettes regularly?	
If 'Yes' (or in the past'), for how long did you smoke cigarettes/pipe/hand-rolled cigarettes regularly?	
If 'Yes' (or in the past'), at what age did you stop smoking cigarettes/pipe/hand-rolled cigarettes?	

Have you ever used snuff regularly?	
If yes, in the past 5 to 10 years, how often would you use snuff each day?	

Have you ever used cannabis regularly?		**If yes, how often would you use cannabis (in a week)?**	
If yes, how would you use cannabis?		**Other - specify:**	

The following questions are about your alcohol drinking habits before you became ill:

Do you or have you ever drunk alcohol regularly?	○ Yes ○ No reset

Maize Homebrew "Mqomboti"		Sorghum Homebrew "Mbila"		Commercial Beer		Spirits (glasses)		Wine/sherry (glasses)	
Days/week:	**Quantity:**	**Days/week:**	**Quantity:**	**Days/week:**	**Quantity:**	**Days/week:**	**Quantity:**	**Days/week:**	**Quantity:**
Age started:	**Age stopped:**	**Age started:**	**Age stopped:**	**Age started:**	**Age stopped:**	**Age started:**	**Age stopped:**	**Age started:**	**Age stopped:**

How long since you last drank alcohol regularly?		**Years ago**

Previous Cancer History & Co-morbidities:

Before you became ill, did you ever have your blood pressure measured?	○ Yes ○ No reset	**Were you treated for high blood pressure?**	○ Yes ○ No reset
Do you have diabetes?	○ Yes ○ No reset	**Do you use tablets or insulin to manage your diabetes?**	
Have you been diagnosed with cancer before?	○ Yes ○ No reset		
If yes, when?		**Type of cancer:**	
Do you have blood relatives* with cancer?			
If yes, then who:		**Type of cancer:**	
If yes, then who:		**Type of cancer:**	
If yes, then who:		**Type of cancer:**	

***Blood relatives can be your children, parents, siblings, children of siblings, anyone related to you by blood.**

HIV Status:

Self-reported HIV status:	
Diagnostically confirmed HIV status:	
Are you on ARTs (Anti retroviral treatment)?	
If yes, when did you start? (YEAR)	
How many tablets are you taking a day?	

Occupation:

Describe your occupation (e.g. Driver, machine operator)		**Describe your usual workplace (e.g. A chemical factory, a gold mine)**	
How long did you work there (Year)?		**Occupation of your spouse/ partner**	
Have you ever worked in any of the following places?	☐ (1) Gold mine ☐ (2) Other mines ☐ (3) Glass/ Brick/ Tile factory ☐ (4) Asbestos factory ☐ (5) Car repair or manufacturing shop ☐ (6) Foundry ☐ (7) Chemical factory ☐ (8) Tyre or Rubber factory ☐ (9) Building site ☐ (10) A farm ☐ (11) Transport industry ☐ (12) A dusty factory ☐ (13) Other - non-industrial		

	Gold mine:	**Other mines:**	**Glass / brick / Tile factory:**	**Asbestos factory:**	**Car repair or manufacturing shop:**	**Foundry:**	**Chemical factory:**	**Tyre or Rubber factory:**	**Building site:**	**A farm:**	**Transport industry:**	**A dusty factory:**
The number of years worked:												

History of Other Ailments:

Do you have a history of any ailments e.g. Skin disorders, heartburn, stomach ulcers etc.	○ Yes ○ No reset	**If Yes, state the ailment**	
		And for how long (years) ?	**years**
Have you ever used medication to make you vomit?		**If yes, how often**	
Do you use/take regular medication?		**For how long did you use this medication?**	
What medication did you use		**For which diseases did you use this medication**	
Did you use traditional or complementary medicine for the current ailment?	○ Yes ○ No reset		

Bio-sample:

Bio-samples collected:	☐ (1) Blood sample - purple tube ☐ (2) Not done due to poor veins/circulation ☐ (3) Saliva sample tube ☐ (4) Yellow serum tube ☐ (5) Tissue sample - Biopsy ☐ (6) Tissue sample - Resection ☐ (7) Stool sample

Appendix 3 – Sub-study ethics clearance

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

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

TO: Ms NL Ngundu and Dr CB Penny
School of Pathology
Department of Human Genetics
Medical School
University

E-mail: 2461584@students.wits.ac.za

CC: Supervisor: Professor CG Mathew
<Chris.G.Mathew@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: 2021/12/01

REF: R14/49

PROTOCOL NO: **M2111143** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *The clinical genomics of African oesophageal cancer*

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.

MSWorks2005\Iain\0007\Clearscan.wps

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Ms NL Ngundu and Dr CB Penny

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2111143

NAME: (Principal Investigator)	Ms NL Ngundu and Dr CB Penny
DEPARTMENT:	School of Pathology Department of Human Genetics Medical School University
PROJECT TITLE:	*The clinical genomics of African oesophageal cancer*
DATE CONSIDERED:	Ad hoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study under M180306
NOTE:	If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>
SUPERVISOR:	Professor CG Mathew
APPROVED BY:	Dr N Naran, Co-Chairperson, HREC (Medical)
DATE OF APPROVAL:	2021/12/01

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 those conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. **I agree to submit a yearly progress report.** When a funder requires annual re certification, the application date will be one year after the date when the study was initially reviewed. In this case, the study was initially reviewed in November and therefore reports and re-certification will be due in the month of **November** each year. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator _____ Date _____

Appendix 4 – Larger study ethics clearance

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Professor C Mathew, et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M2310111

NAME: Professor C Mathew, et al **DEGREE:** N/A
(Principal and Co-Investigators)

DEPARTMENT: Faculty of Health Sciences
Sydney Brenner Institute for Molecular Bioscience
Medical School
University
University

PROJECT TITLE: *Investigation of genetic biomarkers for African oesophageal cancer*

DATE CONSIDERED: Ad hoc

DECISION: Approved unconditionally

CONDITIONS: Renewal of M18/03/06 - expired on 2023/10/08

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>

SUPERVISOR: Not applicable

APPROVED BY: _____
Professor P Ruff, Chairperson, HREC (Medical)

DATE OF APPROVAL: 23 November 2023 **EXPIRY DATE:** 22 November 2028

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

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

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

TO: Professor C Mathew et al
School of Pathology
Department of Human Genetics
Sydney Brenner Institute for Molecular Bioscience
Medical School

E-mail: Chris.G.Mathew@wits.ac.za

CC: Supervisor: Not applicable <>
and <HREC-Medical.ResearchOffice@wits.ac.za>

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

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

DATE: 09/10/2018

REF: *R14/49*

PROTOCOL NO: **M180306** *(This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)*

PROJECT TITLE: *Investigation of genetic biomarkers for African oesophageal cancer*

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

MS/Works2008/Iain0007/Clearacan.wps

R14/49 Professor C Mathew et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M180306

NAME: (Principal Investigator)	Professor C Mathew et al
DEPARTMENT:	School of Pathology Department of Human Genetics Sydney Brenner Institute for Molecular Bioscience Medical School University
PROJECT TITLE:	Investigation of genetic biomarkers for African oesophageal cancer
DATE CONSIDERED:	06/04/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Not applicable
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	09/10/2018

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

DECLARATION OF INVESTIGATORS

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

I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report.** When a funder requires annual re-certification, the application date will be one year after the date of the meeting when the study was initially reviewed. In this case, the study was initially reviewed in **March** and will therefore reports and re-certification will be due early in the month of **March** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature _____ Date _____

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 5 – *R* script

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

BiocManager::install("maftools")

install.packages("R.utils")

BiocManager::install("BSgenome.Hsapiens.UCSC.hg38")

BiocManager::install("NMF")

BiocManager::install("pheatmap")

BiocManager::install("barplot3d")

# https://bioconductor.statistik.tu-dortmund.de/packages/3.5/bioc/vignettes/maftools/inst/doc/maftools.html#oncoplots-aka-waterfall-plots


library("maftools")

library("R.utils")

library("BSgenome.Hsapiens.UCSC.hg38", quietly = TRUE)

library('NMF')

library("barplot3d")


setwd("C:\\Users\\2461584\\Documents\\Masters_WITS\\Data Analysis")


# this is the functional mutation dataset

combined = read.maf(maf = "COHORT.strelka_muTect.somatic_filtered_v2.maf")


# mutational load needs to be on the full dataset

laml.mutload = tcgaCompare(maf = combined, cohortName = 'WES-OSCC', logscale = TRUE, capture_size = 50)


# pathways
```

```
oncoplot(maf = combined, draw_titv = TRUE, top = 22, gene_mar=20,legendFontSize = 1.5,annotationFontSize = 5, anno_height = 2, pathways = 'auto')


# mutational signatures

laml.tnm = trinucleotideMatrix(maf = combined, prefix = '', add = TRUE, ref_genome = "BSgenome.Hsapiens.UCSC.hg38")

plotApobecDiff(tnm=laml.tnm, maf = combined, pVal = 0.2)

laml.sign = estimateSignatures(mat = laml.tnm, nTry = 6)

plotCophenetic(res = laml.sign)

laml.sig = extractSignatures(mat = laml.tnm, n = 3)

laml.v3.cosm = compareSignatures(nmfRes = laml.sig, sig_db = "SBS")

maftools::plotSignatures(nmfRes = laml.sig, title_size = 1.2, sig_db = "SBS")


PlotOncogenicPathways(maf = combined, pathways = "TP53")

OncogenicPathways(maf = combined)

laml.sig = oncodrive(maf = combined, AACol = 'HGVSp', minMut = 5, pvalMethod = 'zscore')

head(laml.sig)

plotOncodrive(res = laml.sig, fdrCutOff = 0.1, useFraction = TRUE, labelSize = 0.5)


getSampleSummary(combined)

getGeneSummary(combined)


combined.titv = titv(maf = combined, plot = FALSE, useSyn = TRUE)

plotTiTv(res = combined.titv, sampleOrder = sample)

data.titv = titv(maf = combined, plot = FALSE, useSyn = TRUE)

plotTiTv(res = data.titv)


lollipopPlot( maf = combined, gene = 'TP53', AACol = 'HGVSp',  showMutationRate = TRUE)
```

```
# vaf plot

vafPlot = plotVaf(maf = combined, vafCol = 'vaf', flip = TRUE)
```