Analysis of Driver Gene Mutations in Oesophageal Squamous Cell Carcinoma.

Hendrina Nelao Mwiiwete Shipanga

SHPHEN003

Thesis is presented for the degree of

Doctor of Philosophy

Division of Medical Biochemistry and Structural Biology

Department of Integrative Biomedical Sciences

Faculty of Health Sciences

University of Cape Town

November 2024

Supervisor: Professor: M.I. Parker
Co-Supervisor: Associate Professor D. Hendricks

Declaration

I Hendrina Nelao Mwiiwete, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature of candidate: Date: 12/11/2024

Acknowledgements

I would like to express my sincere thanks to my supervisor, Prof. M.I. Parker, for his invaluable advice, unwavering support, and continuous guidance throughout my PhD journey. His constant efforts to ensure I had everything needed to complete my research, coupled with his immense knowledge, have greatly empowered me to excel in this project.

To my co-supervisor, Associate Professor D. Hendricks, thank you for your invaluable ideas and advice throughout this project. Your encouragement and insightful suggestions have been instrumental in completing my PhD research. Your endless support throughout my journey is hard to forget throughout my life.

I extend my sincere gratitude to Dr. H. Bendou for his crucial role in the bioinformatics analysis of whole genome sequencing (WGS) data for this thesis. His expertise and support added immense value to my research. I am deeply appreciative of his willingness to share his time and knowledge.

Special thanks to Dr J.D Woodward for his exceptional guidance with the *in silico* structural modelling and analysis in my PhD thesis. His technical skills and thoughtful advice were crucial to overcoming the challenges faced during this stage of my research.

Many thanks to the Division of Medical Biochemistry at UCT and my fellow lab mates especially Dr. Victoria Patten. Your support, advice, and kindness have been invaluable, and your readiness to help has made for a truly pleasant working environment. I am deeply grateful for all you have done.

Thank you to my family—especially my uncle Kauluma Shipanga, my mom, and my grandparents—whose prayers and unwavering support have been a source of strength and motivation throughout this journey. I truly believe I would not be where I am today without them. A heartfelt thank you to Laudika John for your constant encouragement and motivation; your support has been invaluable. I also want to extend my thanks to my friends in Cape Town, who have shared both the joys and challenges of this long journey with me.

I would like to acknowledge the continuous financial support I have received over the years, as well as the contributions of collaborators involved in this project. My sincere thanks go to the University of Cape Town, the L'Oréal–UNESCO For Women in Science Programme, the Newton Fund, UK Medical Research Council and the South African Medical Research Council for their generous funding.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
AfrECC	African Esophageal Cancer Consortium
AID/APOBECs	Activation-induced cytidine deaminase/Apolipoprotein B mRNA editing enzyme catalytic subunit
ANK	Ankyrin repeat
ARF	Alternative reading frame
ATCC	American Type Culture Collection
bp	Base pair
BPE	Bovine Pituitary Extract
BWA	Burrows-Wheeler Aligner
CaVEMan	Cancer Variants through Expectation Maximization
cDNA	Complementary DNA
CDK	Cyclin-dependent kinase
ChEBI	Chemical Entities of Biological Interest
CNAs	Copy number alterations
COSMIC	Catalogue of somatic mutations in cancer
CRCS1	Colorectal cancer
CRISPR/Cas9	Clustered regularly interspaced palindromic repeats/CRISPR-associated protein 9
CSR	Class switch recombination
CT	Threshold cycle
DDR	DNA damage repair
DEPC	Diethyl pyrocarbonate
DEGs	Differentially expressed genes
DMEM	Dulbecco's modified Eagle's medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
dN/dS	Ratio of non-synonymous to synonymous mutations
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacetic acid
EGF	Epidermal Growth Factor
EthBr	Ethidium Bromide
FBS	Foetal bovine serum
FDR	False discovery rate
GenVisR	Genomic Visualizations in R
GERD	Gastroesophageal reflux disease
GO	Gene Ontology
GRCh37	Genome Reference Consortium Human Build 37
GRCh38	Genome Reference Consortium Human Build 38
HNSCC	Head and neck squamous cell carcinoma
HDP	Hierarchical Dirichlet process
HFTC	Hyperphosphatemia familial tumoral calcinosis
HPV	Human papillomavirus
H3K4me3	Histone H3 Lysine 4 Trimethylation
Indels	Insertions/deletions
KEGG	Kyoto Encyclopedia of Genes and Genomes
KSFM	Keratinocyte Serum-Free Median
Mb	Megabase
MDR	Multi-drug resistance

miRNA	MicroRNA
MOPS-EDTA	3-(N-Morpholino)propane sulfonic acid-EDTA
mRNA	Messenger RNA
MSAs	Multiple sequence alignments
NCBI	National Center for Biotechnology Information
ND	Not determinable
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NSCLC	Non–small cell lung cancer
PAHs	Polycyclic aromatic hydrocarbons
PBS	phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PES	Polyethersulfone
PolyPhen-2	Polymorphism Phenotyping v2
OAC	Oesophageal adenocarcinoma
OC	Oesophageal cancer
ORFs	Open reading frames
OSCC	Oesophageal squamous cell carcinoma
qPCR	Quantitative PCR
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
RISCs	RNA-induced silencing complexes
RNA	Ribonucleic acid
RNases	Ribonucleases
RNA-Seq	RNA-Sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
rsID number	Reference SNP cluster ID
RT-qPCR	Reverse-transcription quantitative PCR
SBSs	Single-base substitutions
SD	Standard deviation
SDS	Sodium dodecyl sulphate
SIFT	Sorting Intolerant From Tolerant
siRNA	Small interfering RNA
SNVs	Single nucleotide variations
SNP	Single nucleotide polymorphisms
STRs	Short tandem repeats
SSA	Sub-Saharan Africa
TBE	Tris-Borate-EDTA
ti	Transition
TNPS	TN polyagglutination syndrome
tv	Transversion
UV	Ultraviolet
v/v	Volume/volume
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Abstract

Oesophageal cancer (OC) is the eleventh most diagnosed cancer and the seventh most common cause of cancer-related deaths worldwide. The two main subtypes are oesophageal adenocarcinoma (OAC) and oesophageal squamous cell carcinoma (OSCC). OAC is more common in North America and Europe, while OSCC predominantly occurs in Eastern Asia, Sub-Saharan Africa and Latin America. Over 80% of the OSCC cases and deaths worldwide occur in less developed regions, including Sub-Saharan Africa. The asymptomatic development of OSCC, results in late diagnosis of the disease with a poor prognosis, typically ranging from 5-10% at 5-year post-diagnosis in Africa.

This study investigated the genomic landscape of OSCC in the South African population by whole-genome sequencing (WGS) and whole-exome sequencing (WES). Normal and tumour DNA and RNA was isolated from OSCC patient biopsies prior to the commencement of any form of chemotherapy or radiotherapy. WGS was performed on 31 samples, while WES was conducted on 67 samples. The mRNA levels of selected genes in OSCC were quantitated by RT-qPCR. KYSE30 cells were used for siRNA-mediated knockdown experiments targeting *p14ARF* and *p16INK4a* in OSCC. In silico structural analysis of missense mutations in *p14ARF* and *p16INK4a* was conducted using UCSF Chimera tool.

WGS analysis identified 35 frequently mutated genes in OSCC, among these, *TP53*, *CDKN2A.p16INK4a*, *CDKN2A.p14ARF*, and *KMT2D* were identified as OSCC driver genes. Based on the mutation spectrum analysis, samples clustered into two distinct groups, cluster 1 and cluster 2b, characterized by *TP53* alterations and mutation rates per megabase (Mb). WES expanded findings across 67 samples, identifying *TP53*, *NFE2L2*, *CDKN2A.p16INK4a*, *ZNF750*, and *NOTCH1* as OSCC driver genes. Samples clustered into three groups: cluster 1, cluster 2a, and cluster 2b, expanding upon the two clusters identified in our WGS analysis. In both WGS and WES analyses, cluster 1 exhibited *TP53* mutations and relatively high somatic mutation rates per Mb, while cluster 2 lacked *TP53* mutations. Cluster 2 is further subdivided into clusters 2a and 2b in WES. Cluster 2a samples display a high mutation rate per Mb, while cluster 2b samples display fewer genomic alterations. By quantifying the contribution of the mutational signatures to the mutation spectrum, we found a relatively high contribution of mutation signature SBS1, SBS2, and SBS13, implicating aging and AID/APOBEC (activation-induced cytidine deaminase/apolipoprotein B mRNA editing enzyme catalytic subunit) activation in OSCC tumourigenesis. WGS analysis revealed three novel mutational signatures that had not been previously identified. Interestingly, these signatures were not observed in the samples analysed

by WES, even though the WES cohort included a larger sample size. The significance of these novel mutational signatures remains unclear.

Evaluation of selected differentially expressed genes in OSCC involved in cell cycle control, the KEAP1-NFE2L2 (NRF2) pathway, and DNA damage response pathways showed variable expression of these genes in OSCC suggests potential dysregulation of the genes in OSCC. Furthermore, *p16INK4a* and *p14ARF* mRNA levels were significantly lower in 61% and 48% of OSCC tumour samples, respectively, while elevated levels were observed in 16% and 25% of tumours, respectively. Knockdown of *p14ARF* and *p16INK4a* in KYSE30 cells resulted in dysregulation of key regulators involved in multiple cancer signalling pathways, including cell cycle, apoptosis, and KEAP1-NFE2L2 pathways. This dysregulation could promote cell survival, growth of apoptosis-resistant cells, and resistance to stress, which are critical events in tumorigenesis. *In-silico* mutation analysis revealed damaging mutations in p16INK4a, such as p.A68V, p.D84N, p.D108H, p.D108N, p.D108Y, and p.L130P. These mutations cause significant structural alterations that disrupt interactions crucial for p16INK4a stability and function, possibly affecting cell cycle regulation and potentially promoting tumorigenesis in OSCC.

Our findings highlights novel molecular features of OSCC and provides comprehensive insights into the genomic and molecular mechanisms driving OSCC within the South African population.

Chapter 1
Literature Review

1.1 Oesophageal squamous cell carcinoma

Histologically, there are two major types of OC: oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC) (Figure 1.1) [1, 2]. OSCC typically develops in the middle and upper parts of the oesophagus, originating from squamous cells in the oesophageal lining through the progression of premalignant precursor lesions. These lesions often arise due to chronic irritation and inflammation induced by various risk factors [3-5]. OSCC's aetiology is multi-factorial, with studies highlighting excessive tobacco smoking and alcohol consumption as major contributors to the disease [6-9]. Furthermore, dietary factors such as a diet low in fresh fruits and vegetables leading to low antioxidant levels and vitamin deficiencies also contribute to development of OSCC [10]. On the other hand, OAC arises from glandular cells in the lower third of the oesophagus [11]. Risk factors commonly associated with oesophageal adenocarcinoma include a high body mass index, Barrett's oesophagus, and gastroesophageal reflux disease (GERD) [12].

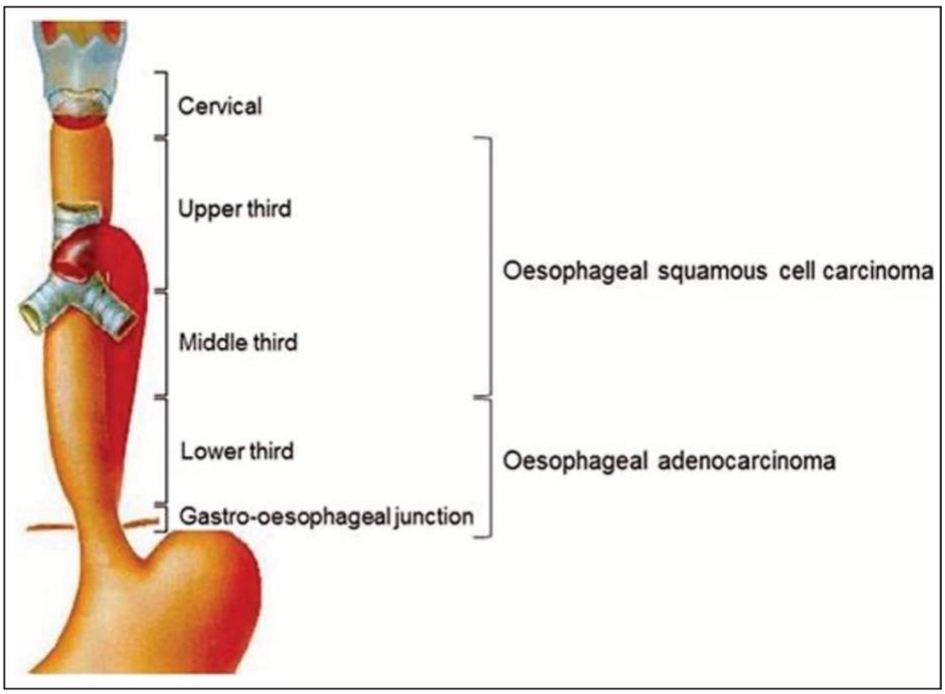


Figure 1.1 Location of the two major subtypes of oesophageal cancer.

Oesophageal squamous cell carcinoma is typically diagnosed in the upper and middle thirds of the oesophagus, while oesophageal adenocarcinoma predominantly occurs in the lower third of the oesophagus and at the gastro-oesophageal junction. Source: ICD–10, WHO [13].

OSCC (the focus of this study) can be an aggressive cancer and is often diagnosed at an advanced stage due to its asymptomatic development, lack of early diagnostic markers, and limited effective treatments [3, 14-17]. Survival rates for OSCC remain low, typically ranging from 5-10% at 5-year post-diagnosis, especially in developing countries such as Sub-Saharan Africa (SSA) countries [18, 19]. Common symptoms of OSCC include dysphagia (difficulty or pain while swallowing), weight loss and chronic indigestion, although these may all reflect fairly advanced stage symptoms [11].

1.2 Epidemiology of oesophageal cancer

Worldwide, OC is the eleventh most common cancer, and the seventh most common cause of cancer related deaths [20]. The highest rates are observed in East Asia, Eastern and Southern Africa, and some parts of Europe, with Malawi having the highest incidence rates worldwide in both men and women (Figure 1.2). Additionally, it is the leading cause of cancer death among men and women in Bangladesh, as well as among men in Malawi and Botswana [20, 21]. However, epidemiological data may be less accurate due to inadequate cancer reporting in many African countries. One distinctive epidemiological characteristic of OC is its uneven geographic distribution (Figure 1.3), which is attributed to differences in underlying risk factors for the two most common subtypes [1, 2]. Among the histological subtypes, OSCC is the most prevalent worldwide [5], with high incidence rates in specific regions of the world such as South America, and the two high-risk belts: the "Asian oesophageal cancer belt" spanning central Asia, from Northern Iran through the Central Asian republics to North-Central China, and the "African oesophageal cancer corridor" extending from eastern to southern regions of SSA [22-24]. In contrast, OAC is the dominant subtype in high-income countries including the United States, Australia, Finland, France, and the United Kingdom [1, 25].

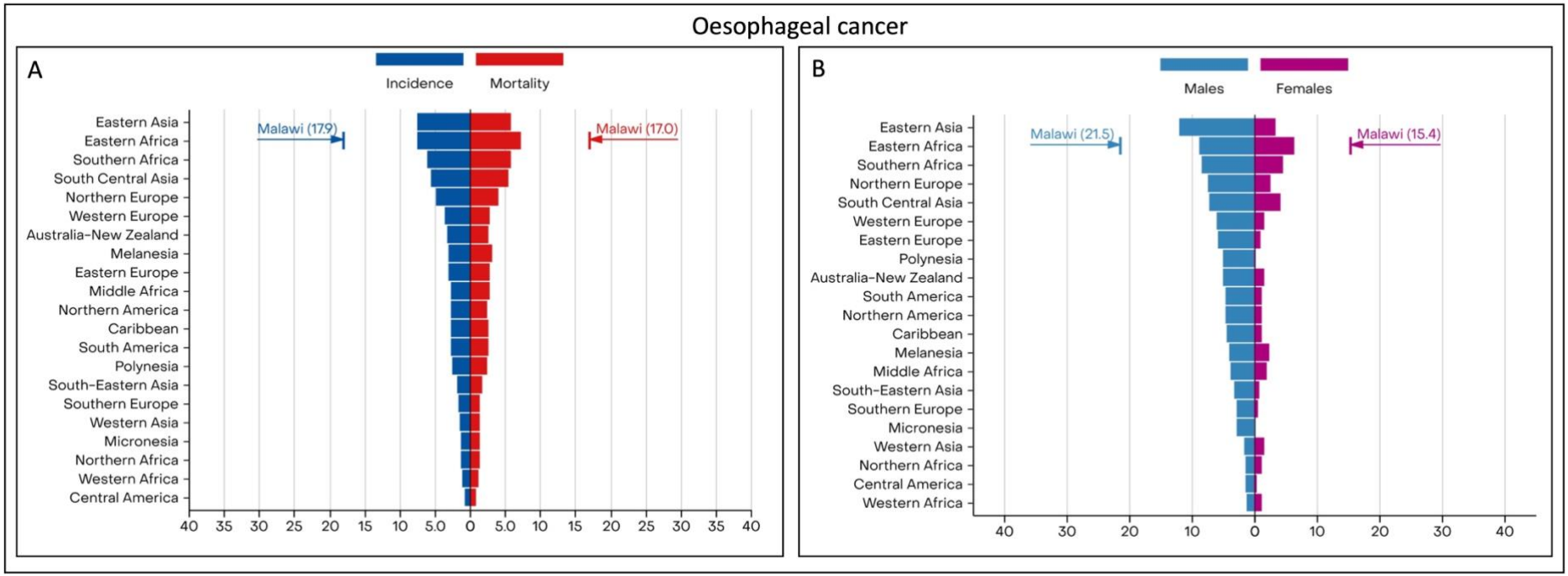


Figure 1.2 Age-standardized oesophageal cancer incidence rates by world and per sex in 2022.

(**A**) Oesophageal cancer incidence and mortality rates, both sexes, per regions; (**B**) Oesophageal cancer incidence rates, per sex, per region. Oesophageal cancer incidence rates vary internationally, with the highest rates observed in Eastern Asia and Eastern Africa, with Malawi having the highest incidence rates worldwide in both men and women. Bar chart depicts the region - specific incidence age - standardized rates of oesophageal cancer in 2022, categorized by sex. The rates are shown in descending order. Values for each world area correspond to incidence of oesophageal cancer per 100,000 individuals among males (left) and females (Right). Data source: GLOBACAN 2024, Cancer today| IARC, (http://gco.iarc.fr/today), accessed 06 June 2024 [21].

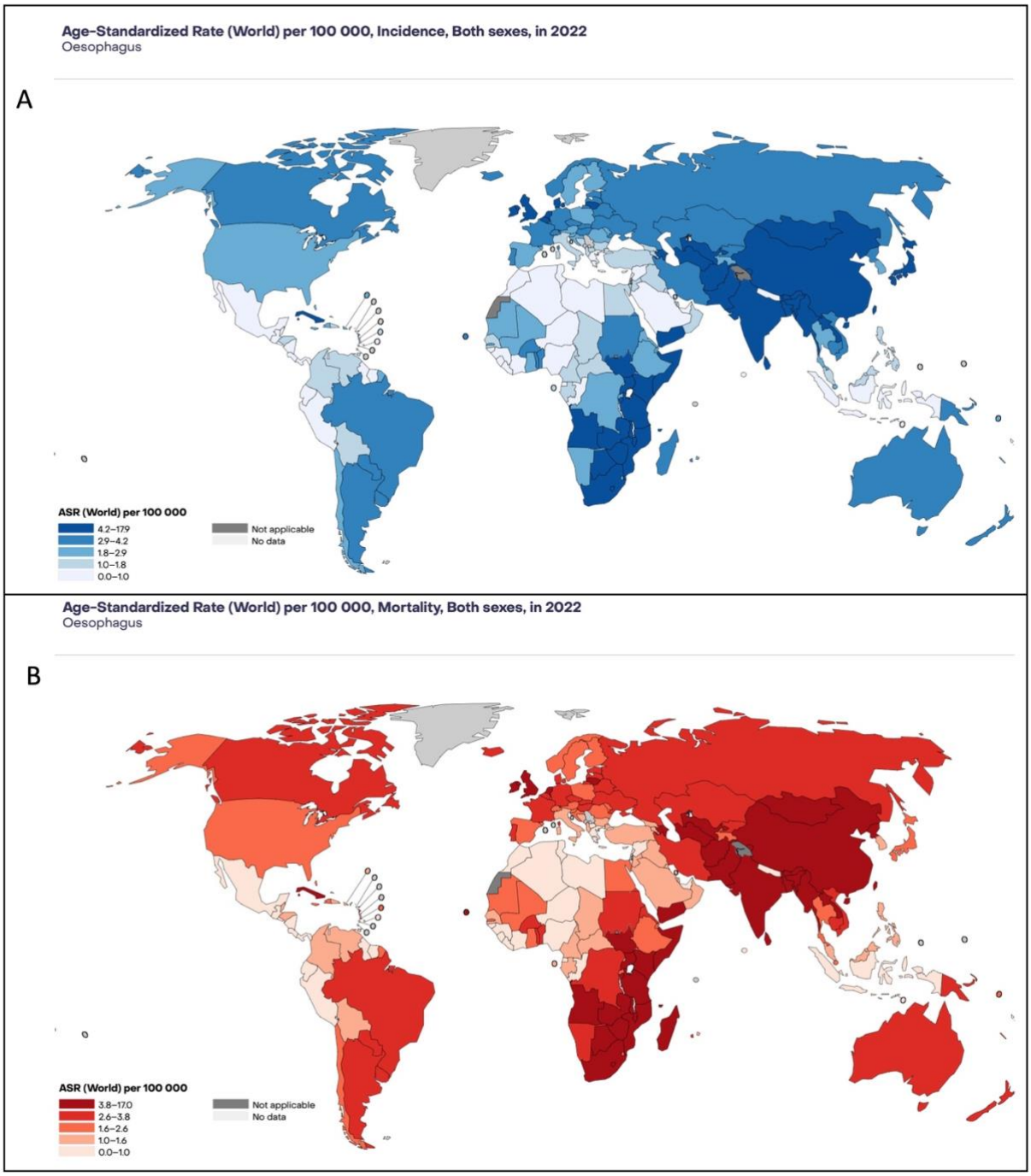


Figure 1.3 Worldwide age-standardized incidence and mortality rates (per 100,000 individuals) of oesophageal cancer (including both OSCC and OAC) in both sexes in 2022.

(**A**) Oesophageal cancer incidence; (**B**) Oesophageal cancer mortality. The highest incidence and mortality rates of OC, especially OSCC, are observed in developing countries such as those in Eastern to Southern regions of Sub-Saharan Africa, parts of Asia such as China, India, and South America. Additionally, high incidence rates of OAC are observed in certain European countries, such as the United Kingdom, France, and Portugal. Data source: GLOBACAN 2024, Cancer today| IARC, (http://gco.iarc.fr/today), accessed 01 May 2024 [21].

In the African continent, the incidence of OSCC is more than 20-fold higher than in developed countries [22]. Specifically, Malawi, Kenya, and Uganda report the highest OSCC incidence

rates in SSA, with Malawi leading at 17.9 cases per 100,000 individuals, followed by Uganda at 13.2 and Kenya at 13.0 per 100,000 individuals [21, 26, 27]. According to Ferlay et al., OSCC is the second cause of cancer death among men and third cause of cancer death among women of Eastern Africa, while in Southern Africa, it is the third most common cause of cancer-related death among men [28]. In South Africa, OSCC has an incidence rate of 6.4 per 100,000 individuals [21]. It ranks the seventh most common type of cancer and the sixth leading cause of cancer death nationwide, with the Eastern Cape province exhibiting one of the highest rates in the country [21, 29]. Despite its significant incidence across African countries, OSCC aetiology and molecular landscape remain understudied [30]. To address this research gap, the members of African Esophageal Cancer Consortium (AfrECC) has undertaken several studies in recent years [31].

1.3 Aetiology of OSCC

Several factors contribute to the risk of OSCC, including lifestyle choices, environmental risk factors and genetic factors [32-36]. Tobacco use and alcohol consumption are considered as major risk factors for OSCC [6, 11, 36-42]. Various risk factors are implicated in OSCC development, particularly in high-incidence areas. These include age, which stands out as the most significant risk factor of OSSC and many other cancer types [43-45]. Exposure to polycyclic aromatic hydrocarbons (PAHs) exposure from biomass combustion [46-49], dietary factors such as micronutrient deficiencies [37, 50-52], consumption of hot foods and beverages [53-56], poor oral hygiene [57] and viral infection [58, 59] also play roles in OSCC development. Furthermore, genomic alterations and epigenetic modifications play a role in OSCC development [60-64]. Both inherited and somatic genetic alterations contribute to OSCC development [65-68].

1.3.1 Environmental and lifestyle risk factors

1.3.1.1 Smoking and alcohol consumption

Tobacco smoke contains several carcinogens, including nitrosamines, polycyclic aromatic hydrocarbons, and acetaldehyde [69], while the alcohol carcinogenic effects arises from its metabolite, acetaldehyde, rather than alcohol itself [70]. Tobacco exposure includes pipe smoking, cigars, hookah, and chewing tobacco [71]. The intensity and duration of exposure have been reported as relevant factors in the risk of developing OSCC [72]. Numerous studies have identified tobacco smoking and alcohol consumption as significant risk factors for OSCC

[36, 37, 73-75]. There is evidence suggesting that tobacco and alcohol may interact synergistically, increasing the risk of OSCC beyond that predicted by either one alone [76].

In lower income countries, such as in parts of Asia and Sub-Saharan Africa, additional suspected risk factors for OSCC have been reported [25], consistent with the observation that over 60% of cases in many studies are non-smokers [39]. Furthermore, a subsequent study in the Taihang Mountains of China found no smoking mutation signatures in OSCC patients, and there was no significant difference in the frequency of smoking-related mutations between smokers and non-smokers in OSCC cases [77].

1.3.1.2 Age

While aging is typically considered a natural process rather than a pathology, and not all elderly individuals develop cancer, most cancer risks increase with age, making them age-related diseases [78]. There is a notable correlation between advanced age and increased cancer incidence [79, 80]. DePinho suggests that the cumulative effects of aging molecular and physiological processes, including mutation accumulation, epigenetic alterations, telomere dysfunction, and alterations in the stromal environment, contribute significantly to cancer development, particularly epithelial cancer [79]. In addition to environmental exposures and habits, age stands out as one of the strongest risk factors of OSSC and numerous other cancer types [45]. OSCC, for instance, predominantly affects older individuals. Studies in India and China indicate that the majority of oesophageal cancer cases occur in individuals aged over 50, and between 60–69 years, respectively [81-84]. However, despite age being a significant factor in the development of OSCC, it does not fully account for the striking geographic disparities in the OSCC incidence between high-risk and low risk-regions. This implies that age alone may not be an independent risk factor.

1.3.1.3 Diet

Several studies have reported a significant association between dietary habits and the risk of developing OSCC. Diets rich in anti-oxidant-containing vegetables and fruits may lower the risk of OSCC [37, 50, 51, 85], whereas diets high in processed foods, exposure to carcinogens from red meat, and excessive salt intake may increase the risk [37, 85, 86]. Ironically, the consumption of some wild vegetables such as umsobo (*Solanum nigrum*), lima beans, maize and pumpkin was associated with increased risk of OSCC in a study conducted in South

African population [87]. This association could be attributed to the presence of protease inhibitors in *Solanum nigrum*, beans, and pumpkin [87]. Epidermal growth factor and transforming growth factor-α are produced in significant amounts by the salivary glands and are present in the oesophageal lumen. Additionally, the oesophageal mucosa secretes its own growth factors, which are crucial for the repair and regeneration of the oesophageal lining. These growth factors are broken down by luminal proteases such as pepsin and trypsin, which are found in the gastric contents that periodically reflux into the oesophageal [87-89]. However, if protease activity is inhibited in the stomach and subsequently in the lower oesophagus, it can lead to an overexpression of growth factors in the oesophagus, creating an environment conducive to proliferation and oncogenesis [87].

Additionally, there is a strong association between consumption of hot foods and beverages and OSCC across various populations [7, 9, 48, 54-56]. The mechanism through which heat-induced lesions can contribute to OSCC development is not fully understood [54]. However, some studies suggest that sustained thermal injury increases the risk of OSCC by inducing inflammatory processes, which could directly impact DNA integrity or increase the exposure of the oesophageal mucosa to luminal carcinogens including N-nitroso compounds and polycyclic aromatic hydrocarbons [54, 90-92].

1.3.1.4 Indoor air pollution: polycyclic aromatic hydrocarbons (PAHs) exposure

In some communities of Africa, Asia, and South America, carbon-containing flammable materials such as wood, coal, gas, or biomass are the primary source of fuel for cooking and heating [93]. The incomplete combustion of organic materials represents a major source of indoor exposure of toxic, mutagenic and carcinogenic PAHs and nitro-PAHs [94]. Besides tobacco use, individuals can be exposed to PAHs through the combustion wood, coal or dung in open fires for cooking or heating, as well as from certain dietary sources [46, 95]. PAHs such as benzo[a]pyrene are known to exert carcinogenic and mutagenic effects [96]. Moreover, exposure to nitro-PAHs like 1,6-Dinitropyrene has been linked to various mutations, including G:C→A:T transitions and G:C→T:A transversion, in mice [97].

Studies in China, Brazil and Iran have highlighted low levels of tobacco and alcohol consumption alongside with high exposure to carcinogenic PAHs in their populations, possibly due to extensive use of coal and wood for cooking and heating. This exposure could contribute

to the high prevalence of OSCC in these regions [98-101]. Furthermore, elevated PAH-metabolite levels were associated with indoor cooking with wood on open, unvented stoves and the presence of advanced oesophageal dysplasia in Kenyan individuals [46], suggesting that PAH exposures increases the risk of OSCC.

1.3.2 Genetic predisposition

The significant geographic variation in OSCC suggest a notable influence of environmental factors on its development. However, genetic factors are also considered to contribute to OSCC [32, 33, 60-62, 68, 102-104]. OSCC is influenced by both inherited and somatic genetic alterations [67]. Inherited genetic alterations associated with familial syndromes play a significant but relatively small role in OSCC development, estimated to contribute about 5-10% of the risk [105]. Syndromes such as tylosis and Fanconi anaemia are associated with an increased risk of various malignancies, including OSCC [67, 106-108]. Tylosis with oesophageal cancer is associated with missense mutations in the *RHBDF2* gene located at 17q25 [109]. On the other hand, mutations in Fanconi anaemia predisposing genes; *FANCD2*, *FANCE* and *FANCL* have been associated with increased risk of OSCC in Iran [107]. Individuals with genetic susceptibility to oesophageal cancer may have an elevated risk of developing OSCC when exposed to environmental or lifestyle factors known to increase OSCC risk [110, 111]. Furthermore, familial aggregation of oesophageal cancer in Turkmen communities in northern Iran was reported to be influenced by environmental factors activating genetically predisposed individuals [112].

Certain single nucleotide polymorphisms (SNPs), notably within the aldehyde dehydrogenase 2 family (*ALDH2*) and alcohol dehydrogenase 1B (*ADH1B*), have been associated with OSCC [67, 113, 114]. These genes are involved in alcohol metabolism, where ethanol is oxidised to acetaldehyde by alcohol dehydrogenases, and subsequently acetaldehyde is oxidised to acetate by aldehyde dehydrogenase enzymes [115]. Acetaldehyde, a known carcinogen, accumulates at higher levels in individuals with genetic variations altering ADH and ALDH activity, thereby increasing susceptibility to OSCC among alcohol drinkers [67, 116]. Additionally, alcohol flushing, often referred to as the "Asian flush," is a genetic trait characterized by an inability to metabolize alcohol properly, primarily due to the *ALDH2* deficiency [117, 118]. This trait is prevalent in East Asian populations and has been extensively studied in relation to OSCC [72, 119, 120]. Studies have shown that individuals with the *ALDH2* deficiency who experience

flushing and consume alcohol moderately or heavily are at a higher risk of developing OSCC compared to those without this deficiency [72, 120, 121].

1.4 Somatic alterations in OSCC

In recent years, numerous molecular studies have characterized somatic mutations in OSCC cohorts from different geographic regions using whole-exome or whole-genome sequencing. These investigations have revealed frequently mutated genes, driver genes, disrupted pathways, copy number alterations (CNAs), and mutation signatures closely associated with the OSCC development [22, 77, 122-143]. The numerous somatic mutations and alterations reported likely represent the impact of several risk factors on the genomes of exposed individuals.

These studies demonstrate frequent mutations in several genes, including *TP53*, *NOTCH1*, *KMT2D*, *KMT2C*, *NFE2L2*, *ZNF750*, *RB1*, *FAT1*, *PIK3CA*, *EP300*, *CDKN2A*, *TNN*, *CREBBP*, *NOTCH3*, *TET2*, *FBXW7*, *TGFBR2*, *KDM6A* and *AJUBA*. Furthermore, copy-number changes were also observed in genes such as *CCND1*, *CDKN2A*, *TP63*/*SOX2*, *FGFR1*, *MYC*, *SHANK2*, *CTTN*, *PIK3CA*, and *RB1*. These genes play important roles in essential signalling pathways including the cell cycle, PI3K-AKT, histone modification, epigenetic regulation pathway, NFE2L2-KEAP pathway, Hippo, and NOTCH pathways [77, 122-134, 136-142]. These findings provide valuable insights into the underlying genetic landscape and pathogenic mechanisms of OSCC, suggesting potential diagnostic, prognostic, and therapeutic markers for this disease.

Mutation signatures are distinct patterns of mutations within an organism's DNA, particularly relevant in cancer research, reflecting potential exposure to DNA-damaging agents [144-146]. These mutation signatures arise from mutation processes, including exposure to environmental factors such as ultraviolet (UV) radiation, or tobacco smoke, as well as defects in DNA repair mechanisms. These processes generate unique combinations of mutation types, termed 'signatures' [144-146]. In addition to frequently mutated genes and disrupted pathways, various mutation signatures have been commonly reported in OSCC. These include the SBS1 mutation signature, associated with aging, and APOBEC-mediated mutation signatures, SBS2 and SBS13 [22, 77, 128, 131, 133, 135, 136, 138]. Beyond these, additional mutation signatures have been reported in various OSCC cohorts, some with unknown aetiologies [22, 77, 131,

135, 136, 138, 139, 142]. These signatures provide insights into the underlying causes and mechanisms of OSCC development. For instance, the presence of SBS1 suggests that age-related processes may contribute to the mutational landscape of OSCC, implying that older individuals might be more susceptible to certain types of mutations associated with increased OSCC risk. On the other hand, APOBEC-mediated mutation signatures indicate the involvement of APOBEC enzymes in the mutation processes driving OSCC. Different patients may exhibit distinct patterns of mutations, potentially reflecting diverse environmental exposures and genetic backgrounds.

1.4.1 Frequently altered genes in OSCC

TP53 is the most frequently mutated gene in OSCC [77, 122-134, 136-141, 143]. It functions as a crucial tumour-suppressor gene, and mutations in *TP53* alter the function of the p53 protein, thereby promoting tumorigenesis [147]. Several well-known tumour-associated genes such as *NOTCH1*, *CDKN2A*, *KMT2D* and *PIK3CA*, exhibit frequent alterations across different OSCC cohorts including Africa patients [22], East Asian populations such as Chinese, Korean and Japanese patients [77, 123-125, 127, 128, 131, 136, 138, 139, 142, 143], Brazilian patients [140], and Indian patients [135]. Furthermore, Song et al., [125] identified *ADAM29* and *FAM135B* as two novel significantly altered genes in their cohort. *FAM135B* promotes the malignancy of OSCC cells. In another study, Gao et al., [123] showed the tumour-suppressive role of *EP300* and found that mutations in *EP300* were associated with poor survival in OSCC. Zhang and others [77] identified *FBXW7*, *FAT1*, *AJUBA*, and *ZNF750* as significantly mutated genes. Their study revealed that *AJUBA* knockdown in KYSE140 and KYSE510 cells led to increased cell growth, colony formation, cell migration, and cell invasion potentially contributing to OSCC tumorigenesis [77]. A subsequent study by Du et al., [131] reported a significant correlation between *AJUBA* mutations and poorer survival among OSCC patients [131]. Lin et al., [124], through WES analysis of 139 OSCC samples, identified *ZNF750* and *FAT1* as tumour suppressors frequently mutated in OSCC. A recent study by Cui et al., [136] identified *NFE2L2* as a tumour suppressor in OSCC, and that *NFE2L2* mutations impair its tumour-suppressive function and confer oncogenic activities [136]. These findings highlight the complex landscape of genetic alterations in OSCC, and their profound implications for tumour development and patient prognosis.

Various studies highlighted the heterogeneous nature of OSCC among patients from the same and different OSCC populations [19]. A comprehensive analysis of molecular alterations in OSCC patients originating from various geographical regions including Western regions (North America, Eastern Europe or West Europe and Brazil) and Eastern regions (Vietnam) identified three main OSCC subtypes: OSCC1, OSCC2, and OSCC3 [132].

Subtype 1 (OSCC1), also known as the "classical" subtype, exhibits similar somatic mutations as the head and neck and lung squamous cell carcinomas. This subtype was characterized by mutations in the NFE2L2-KEAP pathway, involved in oxidative stress and detoxification, including alterations in *NFE2L2*, *KEAP1*, *CUL3* and *ATG7*. OSCC1 is predominantly observed in Asian patients [132]. Similarly, a comparative analysis between Asian and Caucasian patients revealed higher mutational frequencies of *TP53*, *EP300*, and *NFE2L2* in Asian patients with OSCC than in their Caucasian patients [130].

Subtype 2 (OSCC2) displays frequent alterations in *NOTCH1* or *ZNF750*, inactivating mutations of *KDM6A* and *KMT2D* and inactivation of *PTEN* or *PIK3R1*. This subtype primarily occurs primarily in Eastern European and South American patients.

Subtype 3 (OSCC3) exhibits no evidence of genetic deregulation of the cell cycle, with only one out of four samples displaying mutations of *TP53*. Moreover, all samples in subtype 3 had activating alterations of the PI3K pathway, alongside *KMT2D* and *SMARCA4* mutations. This subtype has been identified in African American patients [132]. However, due to the small sample size of just four, further research is needed to validate these findings. The identification of three molecular subtypes suggests genetic differences existing among Caucasian, Asian and African American OSCC patients [130, 132].

Similarly, a WES analysis of 59 Malawian OSCC samples revealed three subtypes 1a, 1b and 2 based on a combination of sequencing data and RNA expression analysis [22]. Subtype 1b exhibited fewer genomic alterations, with the fewest somatic mutations per Mb, fewer amplifications, particularly in *MYC*, *EGFR*, and *TP63* and fewer deletions of *CDKN2A/B* compared to the other subgroups. In contrast, subtype 1a, displayed an overexpression of genes associated with DNA replication, repair, and recombination, while subtype 2 demonstrated increased expression of genes associated with neural differentiation [22]. The identification of molecular subtypes in OSCC in the Malawian population, further provide insights of the

molecular heterogeneity of this disease. These subtypes exhibit distinct patterns of gene expression and genomic alterations, which may have implications for prognosis and treatment strategies in OSCC.

1.4.2 Driver genes in OSCC

Numerous genes display frequent mutations in OSCC, yet cancer is driven by a few key driver mutations. These mutations disrupt key cellular regulatory pathways, leading to abnormal proliferation and the development of cancer [148, 149]. Differentiating driver genes (genes carrying mutations directly responsible for cancer development and progression) from passenger genes (genes with mutations that do not directly drive cancer initiation and progression) is crucial for diagnostic and prognostic purposes [148, 150]. Several studies used methods such as MutSigCV method [151], oncodriveCLUST [152], OncodriveFML [153]and dNdScv [149] to identify driver genes. In total, 53 driver genes were identified across 14 studies [77, 124, 125, 127, 128, 131-133, 136, 142, 154-157]. Among the identified driver genes, a few appeared in seven or more studies. *TP53* was found in all 14 studies, while *CDKN2A* and *ZNF750* were present in 12 out of 14 studies. *NOTCH1* and *NFE2L2* was found in 11 out of 14 studies, and *KMT2D* and *PIK3CA* were each identified in 10 out of 14 studies. Additionally, *RB1* was detected in 9 out of 14 studies, *AJUBA*, *FBXW7* and *FAT1* were detected in 8 out of 14 studies, while *TGFBR2* and *EP300* were found in 7 out of 14 studies (Table 1.1). The identified driver genes suggest dysregulation in multiple signalling pathways, including cell cycle control, the NOTCH signalling pathway, differentiation, cell-cell adhesion, apoptosis, and signalling cascades, all of which play crucial roles in OSCC pathogenesis. Genes discussed below were identified as driver genes in our study.

1.4.2.1 TP53

Somatic mutations in *TP53* are one of the most prevalent alterations observed in human cancer [158]. Over 80% of patients with OSCC carry *TP53* mutations [77, 123, 125, 138-140, 142, 157]. Most of these mutations occur within the p53 DNA-binding domain (i.e. R175H, G245S, R248Q, R248W, R249S, R273H, R273S, and R282W) [140, 156]. These mutations abrogate the tumour-suppressive function of p53, leading to genomic instability (reviewed in [158]). Previous studies have highlighted the significance of mutants like G245C, p.R273C and R273H in the pathogenesis of various cancer types, including OSCC [159-161]. Subsequent studies

revealed an association between the *TP53* missense mutations and reduced overall survival rates [157].

Interestingly, an analysis by Martincorena et al., [45] of normal oesophageal epithelium from nine individuals (aged 20 to 75 years) found that 5 to 10% of the epithelium exhibited *TP53* mutations, with the mutation frequency seemingly increasing with age; with the oldest subjects displaying *TP53* mutations in 20–35% of their cells. Similarly, another study reported early *TP53* mutations in normal oesophageal epithelia [162]. These mutations may accumulate over time due to exposure to environmental factors such as tobacco smoke or alcohol or due to errors in DNA replication and repair mechanisms as cells divide and age. *TP53* mutations could potentially predispose these cells to accumulate additional mutations, potentially leading to the development of cancerous cells over time. The implications of *TP53* mutations in histologically normal tissue remain uncertain and require further investigation. In OSCC, *TP53* mutations disrupt p53-related pathways including those involved in cell cycle regulation, DNA repair, apoptosis or senescence pathways, thereby contributing to OSCC development. Additionally, these mutations may serve as potential prognostic biomarkers of OSCC [158].

Table 1.1 Driver genes in OSCC.

Author	D.-C. Lin et al., 2014	Song et al., 2014	Zhang et al., 2015	Qin et al., 2016	Sawada et al., 2016	Cancer Genome Atlas Research et al., 2017	Chang et al., 2017	Dai et al., 2017	Du et al., 2017	Lin et al., 2018	X. Li et al., 2018	Cui et al., 2020	Dutta et al., 2020	Li et al., 2022
	TP53	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*	*TP53*
	KMT2D	*CDKN2A*	*NOTCH1*	*CDKN2A*	*CDKN2A*	*ZNF750*	*NOTCH1*	*CDKN2A*	*AJUBA*	*CDKN2A*	*KMT2D*	*FAT1*	*ZNF750*	*NOTCH1*
	FAT1	*RB1*	*CDKN2A*	*NOTCH1*	*CCND1*	*KMT2D*	*KMT2D*	*ZNF750*	*CDKN2A*	*FBXW7*	*NOTCH1*	*NOTCH1*		*KMT2D*
	FAT2	*NFE2L2*	*ZNF750*	*NFE2L2*	*FBXW7*	*NFE2L2*	*FAT1*	*KMT2D*	*KMT2D*	*PIK3CA*	*FAT1*	*KMT2D*		*ZNF750*
	RB1	*ADAM29*	*PIK3CA*		*NOTCH1*	*TGFBR2*	*NFE2L2*		*ZNF750*	*NFE2L2*	*PIK3CA*	*CDKN2A*		*CDKN2A*
	NOTCH1	*FAM135B*	*RB1*		*NOTCH3*	*NOTCH1*	*PIK3CA*		*FAT1*	*ZNF750*	*NFE2L2*	*FBXW7*		*NFE2L2*
	EP300	*PIK3CA*	*AJUBA*		*KMT2D*		*EP300*		*NOTCH1*	*NOTCH1*	*ZNF750*	*ZNF750*		*EP300*
	ZNF750	*NOTCH1*	*FBXW7*		*EP300*		*ZNF750*		*NOTCH3*	*KMT2D*	*EP300*	*FAT2*		*PIK3CA*
	PIK3CA		*FAT1*		*CREBBP*		*CDKN2A*		*PIK3CA*	*FAT1*	*CDKN2A*	*PIK3CA*		*FBXW7*
	KDM6A				*TET2*		*CREBBP*		*NFE2L2*	*KDM6A*	*CREBBP*	*EP300*		*NOTCH3*
	NFE2L2				*FAT1*		*FBXW7*		*RB1*	*PTCH1*	*FBXW7*	*KRT5*		*CREBBP*
	PTEN				*YAP1*		*NOTCH3*		*KDM6A*	*PTEN*	*NOTCH3*	*NFE2L2*		*AJUBA*
	CDKN2A				*AJUBA*		*PTCH1*		*FBXW7*	*TGFBR2*	*RB1*	*CDH10*		*RB1*
					PIK3CA		*RB1*		*CREBBP*	*RB1*	*KDM6A*	*RB1*		*PPFIA2*
					EGFR		*KDM6A*		*TGFBR2*	*ATF5*	*TET2*	*AJUBA*		*KRT5*
					ERBB2		*AJUBA*		*CUL3*	*AJUBA*	*CUL3*	*KMT2C*		*KEAP1*
					NFE2L2		*TGFBR2*		*PTEN*	*ZFP36L2*	*PTEN*	*LILRB3*		*CASP8*
					ZNF750		*CUL3*		*DCDC1*	*EP300*	*AJUBA*	*KDM6A*		*CUL3*
					KLF5		*PTEN*			*CUL3*	*NAV3*	*CREBBP*		*TGFBR2*
					FOXA1		*RBPJ*				*TENM3*	*YEATS2*		*ZBBX*
					SOX2						*PTCH1*	*TGFBR2*		*ATP13A5*
					TERT						*TGFBR2*	*CASP8*		*IRF2BPL*
					LRP1B						*RIPK4*			
											PBRM1			
											USP8			
											BAP			
Software	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV	MutSigCV, MutSigCL, MutSigFN, '20/20þ' ratio-metric	MutSigCV algorithm22, oncodriveFML	dNdScv	MutSigCV, oncodriveCLUST,dNdScv
Journal	Nature Genetics	Nautre	The American Journal of Human Genetics	The American Journal of Human Genetics	Gastroenterology	Nautre	Nature Communications	Journal of Pathology	Scientific Reports	The BMJ	Annals of Oncology	Cell Research	PeerJ	Nature Communications

53 driver genes have been identified across 14 studies in OSCC. Various methods, including MutSigCV, MutSigCL, MutSigFN, 20/20þ' ratio-metric, oncodriveCLUST, OncodriveFML, and dNdScv, were employed for this purpose. Driver genes detected in seven or more studies were shown by highlighting them in yellow.

1.4.2.2 CDKN2A

CDKN2A is located at locus 9p21, encoding two different proteins, *p16INK4a* and *p14ARF*, which are produced through alternative mRNA splicing of the same gene [163]. These proteins are derived by utilisation of different first exons (exon 1α for p16INK4a; exon 1β for p14ARF) and a shared exon 2 but with different open reading frames (ORFs) [164, 165]. Both p16INK4a and p14ARF play crucial roles as cell cycle regulators. While p16INK4a binds and inhibits cyclin-dependent kinase (CDK)4/6, thereby preventing Rb phosphorylation and G1–S phase progression [166], p14ARF protein bind to and inhibits the functions of the murine double minute 2 protein (MDM2), consequently stabilizing p53 levels and activating p53-dependent pathways [167-169].

Alterations and inactivation of the *CDKN2A* locus impairs both the p16INK4a/Rb and p14ARF/p53 pathways, potentially leading to uncontrolled cell growth, a hallmark of cancer [170, 171]. *CDKN2A* mutations are reported in precancerous oesophageal lesions, indicating that these mutations may represent early events in OSCC carcinogenesis [172]. The frequency of *CDNK2A* mutations in OSCC varies, ranging from 4% to 21%, especially in patients from high-risk regions, including China, Japan, Brazil, and SSA [22, 123-125, 128, 138-140, 156, 173]. Truncating mutations in *CDNK2A* are common in OSCC, and a significant copy loss has been reported in the 9p21.3 region encoding *CDKN2A* [22, 123, 125, 127, 128, 136, 138, 140, 143]. Additionally, inactivation of the *p16INK4a* and *p14ARF* has been reported in OSCC [174].

While point mutations in exon 1β (encoding for *p14ARF*) are infrequent [175], genomic alterations in *p16INK4a* are more common in OSCC [176]. Notably, several of these mutations occur at the *p16INK4a* mutational hotspot (e.g., p.R58*, p.W110*, and p.D108X) in OSCC [124, 128, 174]. These alterations in *CDKN2A* (*p16INK4a* and *p14ARF*) likely result in a loss of protein function, disrupting cell cycle control mechanisms and promoting uncontrolled cell proliferation in OSCC [177]. In our study, both *p16INK4a* and *p14ARF* were identified as driver genes.

1.4.2.2.1 *p14ARF*

p14ARF plays a crucial role in regulating cell cycle arrest and apoptosis through both p53-dependent and independent pathways [178]. Decreased or absent *p14ARF* in several

malignancies often results from gene deletion or intragenic mutations, and epigenetic changes including DNA methylation [175, 179]. In humans, the significance of *p14ARF* inactivation in cancer is somewhat less understood compared to mice (*p19ARF* for the mouse), primarily due to its low mutation frequency in humans [169, 175]. However, several studies have reported a significant association between *p14ARF* expression and the risk of various cancers, such as breast cancer, lung cancer, liver cancer, ovarian cancer, and laryngeal cancer [180-184].

1.4.2.2.2 *p16INK4a*

The role of *p16INK4a* has been more extensively investigated and plays an important role in carcinogenesis in several cancer types including cervical cancer, familial melanoma, pancreatic cancer-melanoma syndrome and head and neck squamous cell carcinoma [185-188]. Numerous studies have reported multiple mechanisms of *p16INK4a* inactivation in different human cancers, including OSCC (45, 47, 63-71). Loss of heterozygosity, promoter hypermethylation, histone modification, point mutations and small deletions in *p16INK4a* are commonly observed in pancreatic adenocarcinomas, oesophageal squamous cell carcinoma, prostate cancer, gastric cancer, head and neck squamous cell carcinoma (HNSCC), non–small cell lung cancer (NSCLC), skin cancer and melanoma (reviewed in [189]).

Point mutations and deletion of *p16INK4a* detected in melanomas have been associated with an increased risk of metastasis and disease progression [185, 186]. In addition, alterations in *p16INK4a* in NSCLCs samples were significantly associated with lymph node metastasis [190, 191]. Several studies have demonstrated that loss of *p16INK4a* expression is attributed to promoter methylation of *p16INK4a* in numerous cancers including gastric lymphoma, skin cancer, and neck cancer (reviewed in [189]).

In OSCC, several studies have examined the expression of *p14ARF* and/or *p16INK4a*, revealing variability among samples [174, 175, 192]. For example, Xing et al., [175] observed lower levels of *p16INK4a* and *p14ARF* mRNAs in 12 out of 18 Chinese OSCC samples, with 4 of the 18 samples showing elevated level of *p16INK4a* and *p14ARF* mRNAs. Similarly, de Almeida Simao et al., [192] found significantly reduced mRNA levels of *p14ARF* in 58.8% and *p16INK4a* in 64.7% of the OSCC samples analysed. In a study from Germany, 36 of 53 samples did not express or showed decreased p16INK4a protein, as determined by immunohistochemistry [192]. Similarly, in French samples, 15 out of 33 showed reduced

p16INK4a protein [193]. In Japanese samples, the proportion was notably higher, with 38 out of 42 lacking or exhibiting decreased levels of p16INK4a protein [194]. This consistent pattern of *p14ARF* and *p16INK4a* alterations across different populations underscores the potential significance of *p14ARF* and *p16INK4a* dysregulation in OSCC. Furthermore, low expression of p14ARF and p16INK4*a* disrupts cell cycle regulation, increases susceptibility to oncogenic stimuli, thereby promoting tumour progression, and influencing therapeutic responses in various cancer types, including OSCC [168, 170, 186, 195-199].

1.4.2.3 ZNF750

Zinc finger protein 750 (*ZNF750*) plays a role in epidermal differentiation and is closely associated with cell differentiation in OSCC [200, 201]. In OSCC, *ZNF750* is frequently mutated, with a majority of mutations being truncating mutations [77, 124, 125, 128, 136, 142, 155, 157, 201]. The frequency of *ZNF750* mutations in OSCC varies, ranging from 3.9% to 17% [77, 124, 136, 155, 157], and 87.5% of these mutations in *ZNF750* result in decreased mRNA expression in OSCC [128]. Furthermore, *ZNF750* deletion occurred in 3.4% of OSCC tumours, and its mRNA expression was lower in oesophageal tumours compared with normal tissue [124, 201]. These findings suggest that *ZNF750* functions as a tumour suppressor in OSCC [124, 125]. Additionally, *ZNF750* knockdown significantly enhances proliferation, colony formation, migration and invasion in OSCC cells [202]. Consistently, several studies showed that low *ZNF750* expression correlates with lymph node metastasis [203] and poor prognosis in OSCC patients [201, 203, 204]. These findings indicate that *ZNF750* may confer selective advantages to OSCC cells and play an important role in the progression of OSCC [77].

1.4.2.4 *NOTCH1*

The NOTCH signalling pathway plays a crucial in cell fate determination and differentiation [205, 206]. Mutations in *NOTCH1* can disrupt these processes, leading to abnormal cell growth and differentiation in OSCC. According to data from cBio-Portal, *NOTCH1* mutation occurs in 14% OSCC [158]. These mutations typically result in loss-of-function, suggesting that the loss of NOTCH pathway activity is important for the growth of tumour cells exhibiting squamous differentiation characteristics [207]. Mutations in *NOTCH1* in OSCC tend to cluster in the EGF-like repeats and often lead to loss of function [208]. Loss of *NOTCH1* has been implicated in predisposing the oesophagus to precancerous and squamous cell carcinoma,

partially due to accelerated telomere erosion [209]. Mutations in *NOTCH1* have been associated with well-differentiated, early-stage malignancy and less metastasis to regional lymph nodes. Patients with *NOTCH1* mutations also tend to have shorter survival times compare to patients without *NOTCH1* mutations [210]. These findings suggest that *NOTCH1* may act as a tumour suppressor by regulating tumour growth rather than metastasis in OSCC [210].

1.4.2.5 KMT2D

KMT2D, also known as *MLL2*, is involved in histone methylation and the regulation of gene expression. Mutations in *KMT2D* can disrupt gene expression patterns, potentially promoting OSCC progression. Histone modifications are key players in the occurrence and development of various cancers, including OSCC [211]. *KMT2D* plays critical roles in regulation of development, differentiation, metabolism, and tumour suppression [212]. OSCC exhibits frequent mutations in genes involved in histone modification, with *KMT2D* mutations present in 5% to 19% of OSCC samples. Many of these mutations result in truncated KMT2D protein lacking the crucial methyltransferase domain [22, 123-125, 127, 128, 136, 141, 158, 213], suggesting a potential tumour suppressor role of *KMT2D* in OSCC. Notably, *KMT2D* mutations are more prevalent in metastatic OSCC (60%) compared to primary OSCC (15.3%), indicating a significant involvement of *KMT2D* in OSCC metastasis [157]. However, the clinical significance and prognostic value of *KMT2D* mutations in OSCC remain poorly understood and warrants further investigation [158].

1.4.2.6 ***NFE2L2***

NFE2L2 (NRF2) is a transcription factor that regulates and activates the expression of antioxidant response genes crucial for cell defence against oxidative stress [214]. Under normal conditions, *NFE2L2* binds to *KEAP1* through its DLG and ETGE domains, leading to its proteasomal degradation and maintaining low cellular levels [215, 216]. However, upon exposure to stresses, inactivation of *KEAP1* stabilizes *NFE2L2*, promoting the upregulation of its target genes, thereby enhancing stress resistance and cell proliferation [214, 217].

The *NFE2L2*/*KEAP1* pathway is frequently activated in human cancers through mutations in *NFE2L2* or its negative regulator *KEAP1*. These alterations in *NFE2L2* are associated with poor prognosis [218]. In OSCC, *NFE2L2* mutations are occur in 4%-18% samples [22, 77, 123,

124, 127, 128, 136, 140, 142, 155, 158], primarily, clustering in several *NFE2L2* mutation hotspots (p.W24, p.V32, p.R34, p.D77, p.E79 and p.E82), notably within *KEAP1* binding motifs (ETGE and DLG) [136, 156, 219]. Mutations such as p.R34Q and p.E79K hinder NFE2L2's tumour suppressive activity, exerting an oncogenic role in OSCC and significantly associated with poor prognosis [136].

A study analysing 1145 tumour samples detected *NFE2L2* mutations in 11.4% of samples, often accompanied with an increased NFE2L2 expression in the nuclei [220]. In contrast, Cui et al., [136] observed reduced NFE2L2 expression in tumours samples compared to matched normal samples. Silencing endogenous wild-type *NFE2L2* in OSCC cells resulted in increased cell proliferation, while overexpression of exogenous wild-type *NFE2L2* in OSCC cells significantly reduced cell proliferation [136]. Moreover, silencing mutant *NFE2L2* decreased cell proliferation in OSCC cell lines [136] and enhanced the sensitivity of OSCC cells to chemotherapy [221]. Mutant *NFE2L2* might disrupt its tumour suppressive role and was associated with attachment-independent cell survival, correlating with lymph node metastasis, tumour progression, poor prognosis, and could potentially serve as a prognostic biomarker [136, 216, 221, 222].

Several other driver genes have been identified in OSCC, including *FAM135B*, *ADAM29*, *FAT1*, *FAT2*, *TGFBR2*, *RB1*, *ERBB2*, *SOX2*, *CREBBP*, *NAV3*, *LRP1B*, *EP300*, *TET2*, *PTCH1*, *USP8*, *RIPK4*, *KEAP1*, *PTEN*, *AJUBA*, *FBXW7*, *CUL3* and *CASP8*. While some studies have provided insights into their characteristics, their exact roles in OSCC remain poorly understood [77, 124, 125, 127, 128, 131, 133, 136, 154, 156].

1.4.3 Transition to transversion ratio in OSCC

Substitution mutations in genomes, such as transition (ti) and transversion (tv), hold significant importance [223]. Among 12 substitution mutations, 8 are transversions (A→T, T→A, A→C, C→A, G→T, T→G, G→C, and C→G), while transitions consist of 4 mutations (A→G, G→A, T→C and C→T) [224]. Notably, transition mutations tend to occur at higher rates compared to transversions [225, 226]. Analysis of the mutational spectrum of OSCC revealed the predominance of C:G>T:A transitions, followed by C:G>A:T and C:G>G:C transversions [131, 138]. The predominance of C:G> T:A transitions suggests spontaneous cytosine deamination to thymine, being a major mutagenic process in cancer [227], and also observed

in OSCC [77, 122, 123, 125, 128, 131, 137, 138, 142, 154]. Compared to OAC [122], head and neck squamous cell carcinoma [228] and lung squamous cell carcinoma [229], OSCC exhibits a mutation spectrum pattern similar to head and neck squamous cell carcinoma, but differs from lung squamous cell carcinoma, primarily characterised by C:G>A:T transversions [125]. While the most prevalent mutations in OAC are C:G>A:T transitions, the second most frequent mutations are T:A>G:C transversions [125]. Additionally, OSCC has a mutational rate higher than breast carcinoma and glioblastoma multiforme, but lower than head and neck squamous cell carcinoma, OAC and lung squamous cell carcinoma [122, 125, 228, 229].

1.4.4 Mutation signatures in OSCC

Mutation signatures provide insights into the underlying mechanisms of mutagenesis and DNA repair processes that contribute to cancer development. They enhance our understanding how different exposures and genetic defects contribute to the initiation and progression of cancer [144-146]. These process are either active throughout a patient's life or sporadically active, often influenced by lifestyle choices [144].

Each mutation process typically involves DNA damage or modification, repair mechanisms, and DNA replication [230]. For example, exogenous mutagens like UV light in skin cancer and tobacco smoke in lung cancer [231], or endogenous mutagens such as the spontaneous deamination of 5-methylcytosines to thymine [232] and DNA maintenance abnormalities such as defective DNA mismatch repair in some colorectal cancers contribute to these signatures [233]. These processes involve DNA damage that triggers repair mechanisms. Defects in these repair processes can leave distinct genomic imprints that are detectable through sequencing methods [146]. Each mutation signature represents a distinct combination of mutation types and their relative frequencies across the genome (Figure 1.4).

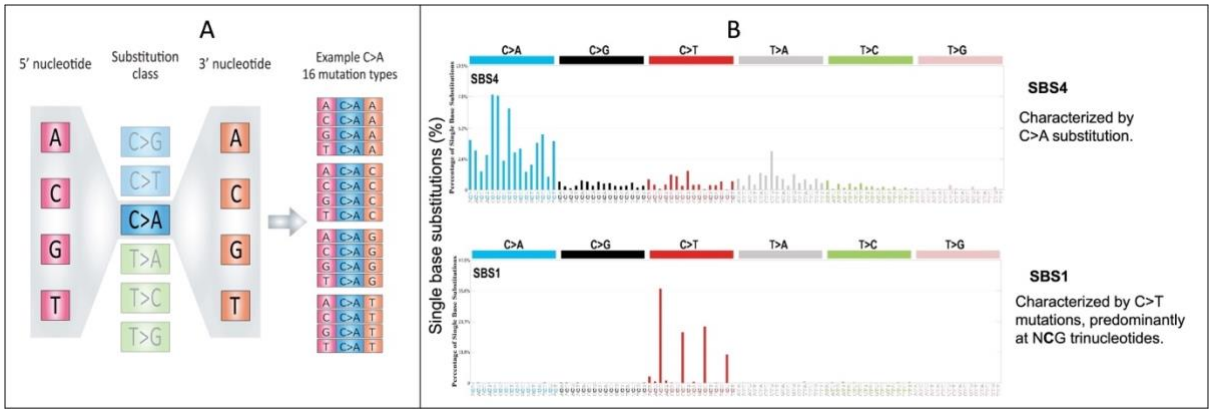


Figure 1.4 Mutation signatures.

(**A**) The 96 single base substitution (SBS) types. Considering the 5' flanking base (A, C, G, T), the 6 substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) and 3' flanking base (A, C, G, T) leads to a 96 mutation types classification (six types of substitution × four types of 5′ base × four types of 3′ base, 6 x 4 x 4 = 96). The 16 possible mutation types of the substitution class C>A are shown as an example. (**B**) Examples of mutation signatures, SBS4 characterized by C>A substitutions, associated with tobacco-smoking, and SBS1 characterized by C>T mutations, predominantly at N**C**G trinucleotides, associated with aging. Modified from [145, 230].

The primary classification for single-base substitutions (SBSs) consists of 96 classes, including the six somatic base substitutions: C>A, C>G, C>T, T>A, T>C, and T>G, with the mutated base is represented by the pyrimidine of the base pair, as well as their flanking 5′ and 3′ bases (Figure 1.4A) [144, 146, 230, 234, 235]. The 96-substitution classification differentiate between mutation signatures that cause the same substitution in different sequence contexts, resulting over 60 SBS mutation signatures [230].

For example, lung cancer, OAC, liver cancer and head and neck cancers exhibit mutation signature SBS4, characterized by higher prevalence of C>A substitutions, known to be caused by carcinogens in tobacco. Additionally, mutation signature SBS1 is found in nearly all cancer types, characterized by C > T mutations associated with aging (Figure 1.4B) [146, 230, 236, 237]. In most cancer types, at least two mutation signatures are observed, with a maximum of six signatures in cancers of the liver, uterus and stomach [146].

Previous studies have identified SBS1, SBS2 and SBS13 mutation signatures as commonly occurring in OSCC across various regions [22, 77, 123-125, 128, 131, 135, 136, 138, 139, 142, 238-240]. SBS1 is the result of an endogenous mutation process initiated by spontaneous deamination of 5-methylcytosine and exhibits clock-like behaviour [146, 241]. On the other hand, SBS2 and SBS13 have been attributed to the AID/APOBEC family of cytidine deaminases. APOBEC genes encode DNA deaminases that lead to mutations in C > G and C

> T [242, 243], with SBS2 characterized by C>T mutations, and SBS13 primarily resulting in C>G mutations [230]. The APOBEC enzymes can deaminate cytosine to uracil, leading to a cluster of mutations and potentially playing a role in various types of cancer types, including OSCC [77, 125, 244].

Despite tobacco smoking being a well-established risk factor for OSCC, the smoking-associated mutation signatures including SBS4, were notably absent in several OSCC studies [22, 77, 138, 139, 155]. Furthermore, some studies found no significant differences in mutation rates or composition between smokers and non-smokers, suggesting that smoking may contribute to OSCC risk through mechanisms distinct from other smoking-related cancer, such as lung squamous cell carcinoma, small-cell lung cancer, or lung adenocarcinoma [123]. However, in a study of Indian patients with OSCC, higher mutation loads and an enrichment of smoking-associated SBS4 were observed in tobacco chewers compared to smokers and non-tobacco users, suggesting a direct exposure of oesophageal tissue to tobacco mutagens (chewing tobacco) may increase OSCC risk [135].

Several OSCC studies have identified additional mutation signatures beyond SBS1, SBS2 and SBS13 in their cohorts [22, 77, 131, 135, 136, 138, 139, 142]. Particularly interesting findings in Malawian, Chinese and Korean studies indicating the presence of signatures with unknown aetiologies that did not match any of the known COSMIC mutation signatures [22, 77, 131, 136, 139]. For example, Du et al., [131] identified an unknown mutation signature in their cohort, previously reported in OSCC [77]. Similarly, Cui et al., [136] found signature S7 and S10, which did not match any COSMIC mutation signatures [245]. These findings suggest the existence of unidentified carcinogenic mechanisms contributing to the malignancy of OSCC [19, 77].

1.4.5 Commonly altered pathways in OSCC

Major signalling pathways known to be important in cancer are disrupted in OSCC, including the cell cycle, PI3K-AKT, histone modification, epigenetic regulation pathway, NFE2L2/KEAP pathway, DNA repair, Hippo, and NOTCH pathways [77, 122-134, 136-142]. The dysregulation of these pathways play a significant role in OSCC development and progression. Understanding these aberrations could potentially lead to the development of targeted therapies or diagnostic biomarkers for OSCC management.

1.4.5.1 Cell cycle

In OSCC, the cell cycle-related genes including *TP53*, *CDKN2A*, *CCND1*, *NFE2L2*, and *RB1*, undergo frequent alterations, indicating a significant disruption in cell cycle progression and genome instability [128, 131, 136, 138, 140, 142, 143, 156, 246]. These genes play crucial roles in controlling proliferation, apoptosis and cell survival [247]. Frequent mutations in these genes suggests that targeting this pathway could be a promising therapeutic strategy in managing OSCC [158].

1.4.5.2 NOTCH signalling pathway and Hippo pathway

The NOTCH pathway is also a frequent mutation target not only in OSCC but also in several other cancer types [205, 208, 248-250]. Furthermore, the Hippo pathway, another frequently targeted pathway in OSCC, interacts with the NOTCH pathway across various cell types [251, 252]. Both pathways regulate cell proliferation, fate determination, and cell survival [253-255]. Loss-of-function mutations, such as nonsense mutations or frameshift indels, affect several genes within these pathways, including *AJUBA*, *PTCH1*, *FAT1*/*2*/*3*/*4* in the Hippo signalling pathway and *FBXW7*, *NOTCH1*/*2*/*3*/*4* in the NOTCH signalling pathway. Inactivation of these signalling pathways promote the development of squamous cell carcinoma in other tissues, including cutaneous, head-and-neck and oesophageal squamous tumours [77, 122, 123, 128, 136, 156, 256-258]. This suggests that the loss activity in these pathways may be critical for the growth of differentiating squamous tumour cells [207]. The *FBXW7* gene, acting as a negative regulator of the NOTCH pathway, demonstrates an oncogenic role for NOTCH signalling when its function is compromised [259]. The involvement of the NOTCH and Hippo pathways, along with their crosstalk, further highlights the complexity of OSCC pathogenesis and the potential for targeted treatments.

1.4.5.3 PI3K-AKT pathway

The dysregulation of the PI3K-AKT pathway has been directly linked to several human cancers including OSCC [260-265]. This pathway is involved in regulating cell functions such as proliferation, differentiation, apoptosis, cell survival, cell growth, and angiogenesis, often implicated in OSCC development through driver mutations in the *PIK3CA* gene [123, 124, 260, 262, 263, 265-268]. Among the most significant genetic alterations within this pathway is the loss of the tumour suppressor *PTEN*, as well as activating point mutations, and

amplification of *PIK3CA*, which encodes the catalytic subunit of phosphatidyl 3-kinase (PI3K), specifically p110α, and amplification of AKT (protein kinase B) [266, 269, 270]. Similarly, genes involved in the PI3K-AKT, such as *PIK3CA*, *ERBB2*, *PTEN*, *EGFR*, and *AKT1/2/3*, frequently undergo alterations in OSCC [22, 123, 124, 128, 136, 140, 142, 154, 156]. These mutations, particularly activating point mutations and amplification of *PIK3CA*, are associated with a poor prognosis in OSCC, suggesting PI3Kα is a promising target for OSCC treatment [138, 158, 264].

1.4.5.4 Histone modification

Histone modification enzymes control chromatin structure and regulate gene expression, playing an important role in cancer initiation and progression [211]. OSCC frequently exhibits alterations in histone modifier genes including *KMT2D*, *KMT2C* (also called *MLL3*), *KDM6A*, *EP300* and *CREBBP* [22, 123, 125, 128, 154, 156]. While the role of these alterations remains unclear, drug candidates targeting epigenetic modulators have shown potential activity against OSCC [158]. Moreover, these alterations have the potential to disrupt epithelial homeostasis, contributing to OSCC tumorigenesis and progression [158]. Notably, mutations in *EP300* have been associated with poorer survival outcomes in OSCC [123].

1.4.5.5 NFE2L2/KEAP1 (NRF2) pathway

Frequent mutations in the NFE2L2/KEAP1 (NRF2) pathway have been consistently reported in OSCC [22, 77, 123, 124, 127, 128, 136, 140, 142, 155, 158], suggesting that mutations and dysfunction within this pathway may contribute to OSCC development by increasing the resistance to oxidative stress [131]. Key genes involved in the NFE2L2/KEAP1 pathway, such as *NFE2L2*, *KEAP1* and *CUL3*, frequently undergo alterations in OSCC [131, 136, 156]. The *NFE2L2* gene encodes the NFE2L2 protein, a transcription factor regulating the expression of antioxidant proteins that protect against damage caused by injuries and inflammation [214]. Genetic deletion of NFE2L2 increases the susceptibility to cancer development, allowing tumour cells to survive the oxidative stress induced by chemoradiation, leading to resistance to treatment [215, 216, 221, 271-274]. Driver mutations in NFE2L2 are believed to occur as late events in OSCC development [130], while mutations such as p.R34Q and p.E79K hinder NFE2L2's tumour suppressive function, promoting an oncogenic role in OSCC and significantly correlating with worse prognosis [136].

Several pathways besides those previously discussed have been linked to OSCC development, including DNA repair, the RTK-RAS pathway, WNT pathway, cell-cell adhesion, apoptosis and *TP53* pathway [125, 128, 132, 136, 142]. The involvement of these pathways suggests a multifaceted and complex aetiology underlying OSCC, indicating that OSCC arises from the interplay of various molecular mechanisms rather than being driven by a single pathway. Understanding the roles of these pathways could guide the development of more comprehensive diagnostic and therapeutic strategies targeting the diverse molecular alterations present in OSCC.

1.5 The expression of frequently mutated genes in OSCC

Differential gene expression between tissues allows the comparison of gene expression patterns across various tissues and conditions, offering insights into genes potentially contributing to different phenotypes. For example, the comparison of healthy versus diseased tissues can unveil genetic factors involved in disease pathology [275]. Among the methods used for gene expression analysis, mRNA Quantitative PCR (qPCR) remains the most widely used technique [276]. Other methods such as RNA-Sequencing (RNA-Seq) [277] and Microarrays [278] have also been instrumental in cancer research, particularly in identifying biomarkers for clinical endpoints such as diagnosis, prognosis, and treatment response prediction [279-282].

In OSCC, gene expression changes have been linked to disease progression [283]. Differential gene expression analysis has identified several key genes involved in OSCC. For instance, in an analysis of 76 paired tumour-normal OSCC samples, Gao et al., [123] observed that the expression level of *AJUBA* tended to be lower in *AJUBA*-mutant tumours compared to those tumours with wild-type *AJUBA*. Song et al., [125] identified *FAM135B* amplifications in 35 of 140 of OSCC samples. Additionally, they noted high expression of *FAM135B* in all nine OSCC cell lines (KYSE2, KYSE30, KYSE70, KYSE140, KYSE150, KYSE180, KYSE450, KYSE10 and COLO680N) compared with immortalized normal oesophageal cells (NE3). A recent study by Cui et al., [136] detected lower expression of NFE2L2 in *NFE2L2* mutant tumours compared to normal tissues. Additionally, RNA expression analysis on 59 Malawian OSCC samples revealed increased expression of *TP63*, the squamous cytokeratins such as *KRT5*, *KRT15* and keratinocyte-specific transcripts including *BNC1*, *DSC3*, and *DSG3* [22].

The results suggest that OSCC involves complex changes in gene expression that might influence various aspects of tumour biology, including growth, differentiation, and response to treatment. Understanding these changes can provide insights into the mechanisms of OSCC progression. Further research is needed to validate these findings and explore their implications for OSCC development and progression.

1.6 Project significance

OSCC presents a significant health challenge in Sub-Saharan Africa [27]. Despite its severity, the underlying biological mechanisms driving OSCC's lethality and aggressiveness remain poorly understood [136]. Although several studies have examined the genetic alterations associated with OSCC in other populations such as Chinese, Japanese, African American, European and Indian populations [123, 124, 127, 128, 132, 135-137, 139-141, 143, 213], the genomic landscape of OSCC, particularly in high-risk regions like Sub-Saharan Africa is not well-characterized [19]. Furthermore, the genomic alterations, including mutation signatures, and driver genes associated with OSCC in South African are not well-characterized, both at whole-genome and whole-exome levels. Thus, there is a need to identify cancer driver genes associated with OSCC in South African patients, potentially serving as diagnostic markers and therapeutic targets. This study employed whole-genome sequencing of thirty-one (31) and whole-exome sequencing of sixty-seven (67) OSCC tumours and matched blood from individuals in South African population, to characterize the mutational landscape of OSCC within this population, by identifying frequently mutated genes, driver genes, altered signalling pathways, and distinctive mutation signatures prevalent in OSCC in the South African population.

1.7 Aim and Objectives

The study aims to identify frequently mutated genes, driver genes, mutation signatures, and further characterise selected mutations within OSCC in South African patients, the objectives are:

1. Use whole-genome sequencing of 31 tumours and matched blood samples and whole-exome sequencing of 67 OSCC tumours and matched blood samples, to identify significantly mutated genes, distinct driver gene mutations and mutation signatures in OSCC in the South African patients.

2. Investigate the expression patterns of selected differentially expressed genes in OSCC biopsies that are involved in specific molecular pathways associated with OSCC to evaluate their potential impact on tumour progression.

3. Investigate *p16INK4a* and *p14ARF* in cultured oesophageal squamous cell carcinoma cells.

Chapter 2
Significantly mutated genes in OSCC in South African patients

2.1 Introduction

Genomic analyses of OSCC revealed a complex and diverse landscape of the disease, evident in both inter-tumour and intra-tumour forms, manifesting at both genomic and epigenomic levels. This heterogeneity significantly contributes to tumour development, drug resistance, and metastasis (reviewed in [284]). Numerous studies have demonstrated that the poor clinical outcomes associated with OSCC can, in part, be attributed to the substantial heterogeneity observed among tumours [285, 286]. Furthermore, the exact molecular mechanisms underlying OSCC aetiology are only partially understood, leading to limited targeted therapies and insufficient clinical management in OSCC patients [19, 136]. There is thus a need to identify driver genes associated with OSCC that could serve as either diagnostic markers and/or potential targets for therapeutic interventions [136].

Several studies using whole-genome sequence (WGS) and/or whole-exome sequence (WES) strategies have investigated genetic alterations in OSCC across both Western (American and European) and Eastern (Chinese, Japanese, Korean and Indian) populations. These studies reported frequent mutations in several known cancer genes and potential OSCC driver genes such as *TP53*, *CDKN2A, ZNF750*, *NOTCH1*, *KMT2D*, *NFE2L2*, *PIK3CA*, *AJUBA*, *RB1*, *FBXW7*, *TGFBR2* and *FAT1* [77, 124, 125, 127, 128, 131-133, 136, 142, 154-157]. These genes play crucial roles in cancer-associated pathways such as cell cycle, hippo pathway, epigenetic regulation, KEAP1-NFE2L2 pathway, PI3K-AKT signalling, DNA repair, and NOTCH signalling pathway [77, 123-125, 127, 131, 132, 136, 138, 139, 142, 143, 156, 239]. Furthermore, mutation signature analysis conducted on OSCC samples from different geographic regions, including those with lower incidence rates, have revealed three predominant mutation signatures. These include the age-dependent mutation signature (SBS1) and the mutation signatures associated with APOBEC enzymes (SBS2 and SBS13) [77, 131, 133, 135, 136, 138, 139, 142, 155, 287]. Interestingly, despite significant differences in OSCC incidence rates and the complexity of molecular mechanisms, these mutation signatures are consistently observed across diverse populations [287]. In addition, several studies have identified additional mutation signatures such as SBS4, SBS16, SBS6, SBS3, and SBS10 [77, 131, 133, 135, 136, 138, 139, 142, 155]. Mutation signatures SBS4 and SBS16 is associated

with smoking and alcohol consumption, respectively. Furthermore, mutation signature SBS6 is attributed to DNA mismatch repair deficiency, and mutation signature SBS3 is associated with homologous recombination deficiency, while SBS10 is associated with polymerase epsilon (*POLE*) exonuclease domain mutations [144-146, 230]. Importantly, several studies have found mutation signatures in OSCC with unknown aetiologies that have yet to be fully understood in terms of their origins [77, 131, 133, 136, 142].

Various studies have identified distinct molecular subtypes within their cohorts [130, 132, 137]. OSCC displays a complex mutational profile, with samples often clustered into high and low mutation groups. Asian OSCC patients, for instance, exhibit significantly higher mutation rates in genes such as *TP53*, *NFE2L2*, and *EP300* compared to Caucasian OSCC patients [130, 132]. In contrast, African American OSCC patients generally show less genetic deregulation related to the cell cycle and fewer *TP53* mutations [132, 137]. These molecular subtypes indicate genetic differences among OSCC samples from patients of Caucasian, Asian, or African American ethnic backgrounds [130, 132, 137]. In regions with high incidence of OSCC, like Asia, molecular alterations and genetic mechanisms are relatively well-characterized. However, in sub-Saharan Africa, these aspects remain poorly understood due to a limited number of genomic studies. A whole exome sequencing analysis of 59 Malawian OSCC samples revealed enrichment of mutations in genes such as *KMT2D*, *KMT2C*, *EP300*, *JAG1*, and *SERPINB4*. The study also found inactivating mutations in key tumour suppressor genes, including *TP53*, *CDKN2A*, *NOTCH1/3*, *FAT1/2/3/4*, and *FBXW7*, as well as activating mutations in *NFE2L2* and *PIK3CA* [22]. Furthermore, three distinct subtypes—1a, 1b, and 2—were identified in Malawian OSCC based on RNA expression levels [22].

Given the high incidence of oesophageal squamous cell carcinoma in sub-Saharan Africa and the underrepresentation of molecular data from African OSCC populations, alongside the observed diversity in the genomic landscape across different OSCC patients from various geographic regions, and the need to identify cancer driver genes and prognostic biomarkers associated with OSCC, we performed whole genome sequencing on 31 samples and whole exome sequencing on 67 samples from patients with OSCC. Our aim was to characterize the mutational landscape of OSCC in the South African population. Through this analysis, we have identified frequently mutated genes, molecular subtypes, OSCC cancer driver genes, altered signalling pathways and mutation signatures within OSCC in the South African patients.

2.2 Results

2.2.1 Study participants

Tumour, adjacent normal tissue, matched blood samples and clinical information were obtained from patients diagnosed with OSCC who had not undergone chemotherapy or radiotherapy. Patient recruitment took place at two clinical centres: Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa and at Charlotte Maxeke Johannesburg Academic Hospital, University of the Witwatersrand, Johannesburg, South Africa. Prior the enrolment in the study, written informed consent was obtained from all patients. Ethics approval was obtained from the University of Cape Town/Groote Schuur Hospital Human Research Ethics Committee (Cape Town, South Africa; approval number: HEC040/2005). The whole genome sequencing was performed on 31 pairs of OSCC tumours and matched blood samples (referred to as WGS cohort (Table 2.1)). The cohort included 20 cases from Groote Schuur Hospital cohort and 11 cases from Charlotte Maxeke Johannesburg Academic Hospital cohort (Table 2.1). The WGS cohort was made up of 13 men (42%) and 18 females (52%). Regarding smoking status, 10 cases were smokers (32%), 5 were ex-smokers (16%), 16 were non-smokers (52%). Except for 12 patients with unknown history of alcohol consumption, 13 patients (42%) had a history of alcohol consumption, and 6 patients (19%) had no history of alcohol consumption. Ethnically, the majority (29) were black patients (94%), with 2 mixed ancestry patients (6%) (Table 2.1). Additionally, whole exome sequencing was performed on 67 pairs of OSCC tumours and matched blood samples (referred to as WES cohort in Table 2.1). The WES cohort comprised 29 cases from Groote Schuur Hospital cohort and 38 cases from Charlotte Maxeke Johannesburg Academic Hospital cohort (Table 2.1). There were 39 men (58%) and 28 females (42%). Regarding smoking status in the WES cohort, 24 were smokers (36%), 16 were ex-smokers (24%), 23 were non-smokers (34%), and 4 patients with unknown smoking histories. There were 37 patients (55%) with a history of alcohol consumption (either drinking at time of diagnosis or in the past), 20 patients (30%) with no history of alcohol consumption and 10 patients (15%) with unknown history of alcohol consumption. Ethnically, there were 45 black patients (67%), 1 white (2%), 17 mixed ancestry (25%), and 4 of unknown race (6%). The average age of patients from Charlotte Maxeke cohort was higher than that of patients from GSH cohort (Table 2.1). Black subjects were mainly Xhosa or Zulu speakers from the Western Cape province of South Africa, who migrated from the Eastern Cape over the past 1–2 generations (in the cases from GSH), or from the KwaZulu-Natal province of South Africa (cases from Charlotte Maxeke), respectively. Meanwhile, those with mixed

ancestry subjects were from the Western Cape, representing an admixed population derived from various ethnic groups (indigenous Khoisan, Bantu-speaking Africans, Europeans, Indonesian and Malaysian) [32].

Table 2.1 Summary of the characteristics of patients with OSCC who were included in this study for WGS and WES.

	WGS cohort (n=31)		WES cohort (n=67)	
	GSH (n=20)	Charlotte M. (n=11)	GSH (n=29)	Charlotte M. (n=38)
Clinical factors				
Gender				
Female	10	8	15	13
Male	10	3	14	25
Age (years)				
Mean	56	61	56	65
Min-Max	37-71	38-81	28-86	52-89
0-25	-	-	-	-
26-50	6	2	10	-
51-75	13	7	17	32
76+	1	2	2	6
Smoking history				
Yes	8	2	14	10
Ex-smoker	5	-	7	9
No	7	9	8	15
No info	-	-	-	4
Alcohol				
Yes	13	-	16	2
No	6	-	6	14
In the past	-	-	1	18
No info	1	11	6	4
Ethnicity				
Black	18	11	7	38
Mixed ancestry	2	-	17	-
White	-	-	1	-
No info	-	-	4	-
Histology				
OSCC	20	11	29	38

Patients were recruited at two clinical centres: Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa and at Charlotte Maxeke Johannesburg Academic Hospital, University of the Witwatersrand, Johannesburg, South Africa. The whole genome sequencing was performed on 31 pairs of OSCC tumours and matched blood samples. The cohort included 20 cases from Groote Schuur Hospital cohort and 11 cases from Charlotte Maxeke Johannesburg Academic Hospital cohort. On the other hand, whole exome sequencing was performed on 67 pairs of OSCC tumours and matched blood samples. This cohort comprised 29 cases from Groote Schuur Hospital cohort and 38 cases from Charlotte Maxeke Johannesburg Academic Hospital cohort. WGS cohort - Whole Genome Sequencing cohort, WES cohort - Whole Exome Sequencing cohort, GSH - Groote Schuur Hospital, n – number of cases, OSCC – oesophageal squamous cell carcinoma.

To identify genomic alterations that contribute to the development of OSCC in the South African population, whole genome sequencing of 31 pairs of tumour and matched blood samples and whole exome sequencing of 67 pairs of tumour and matched blood samples was performed at the Wellcome Sanger Institute, United Kingdom. Somatic single nucleotide variations (SNVs) and short insertions/deletions (indels) (<200bp) were called (detailed in Materials and Methods section 6.2.2.1).

Somatic point mutations were identified and filtered using CaVEMan (Cancer Variants through Expectation Maximization) [288]. CaVEMan calls variants by comparing sequencing data from tumour samples with normal samples and then calculating the likelihood of a mutation at each base-pair position locus [288]. After calling the full set of variants, off-target variants and false positive variants were filtered with a set of standard CaVEMan filters outlined in Materials and Methods (Table 6.4). Variants which pass all filters were considered for functional analysis. Small insertions and deletions were called and filtered using cgpPindel [289]. The estimated contribution of mutation signatures (single base substitutions, SBS) was performed using MutationalPatterns [290], and confirmed with deconstructSigs [291] and hierarchical Dirichlet process (HDP) method [292], using the reference collection of COSMIC v.2 and version 3.3 SBS mutation signatures profiles [230] (Materials and Methods section 6.2.2.2). Genes under positive selection (also termed as driver genes in our study) were identified using the dNdSCV method [149] as described in Materials and Methods section 6.2.2.3. The dNdSCV approach analyses patterns of somatic mutations across tumour samples by calculating the ratio of non-synonymous to synonymous mutations (dN/dS) while considering the effects of mutation clonality and cancer-specific selection pressures. This approach differentiates between driver mutations and passengers mutations, providing insights into the genetic mechanisms underlying cancer development and progression [149]. For this analysis, genes with FDR q-value of ≤ 0.05 were considered to be significantly mutated. Using an FDR of $q \leq 0.05$ ensures that the expected fraction of false positives in our analysis does not exceed 5%. This well-established statistical procedure allows one to increase statistical power to detect true positives, while controlling the proportion of false positives [293]. Pathway enrichment analysis was done using Reactome database [294], (Materials and Methods section 6.2.2.6). Pathways affected were considered significant when $p < 0.05$ and FDR < 0.1. Mutational landscape of patients for both WGS and WES data was summarized and plotted

using an R package GenVisR (Genomic Visualizations in R) [295] (details in Materials and Methods section 6.2.2.4).

2.2.2 Profiling of OSCC in South African patients by Whole Genome Sequencing

2.2.2.1 Frequently mutated genes and driver genes

WGS was performed on matched normal-tumour samples from 31 South African patients diagnosed with OSCC. Analysis of these genomes led to the identification of a total 605 533 somatic events, of these 1 260 were synonymous mutations, 2 917 were missense, 231 were nonsense mutations, 4 were stop codon losses, 9 were start codon losses, 527 were splice-site mutations including essential splice and essential splice-region variants, and 426 exonic indels, including 86 in-frame indels and 340 frameshift indels. The median tumour mutation burden among the 31 OSCC patients was 2.5 mutations/Mb, ranging from 0.12–16.5 mutations/Mb (for non-silent variants) across the cohort (Figure 2.1A). The non-silent variants included missense, nonsense, stop codon loss, start codon loss, indels (frameshift and in frame) or splicing-site mutations. Missense mutations constituted the primary type of alterations in the coding region of genes in tumour samples (Figure 2.1C). The most frequently mutated genes in our cohort were *TP53* (26/31, 84%), *AHNAK2* (12/31, 39%), *MUC4* (11/31, 35%), *CDKN2A* (11/31, 35%), *TTN* (10/31, 32%), *AHNAK* (9/31, 29%), *NOTCH1* (8/31, 25%), *PCLO* (7/31, 23%), *KMT2D* (7/31, 22%), *PLEC* (6/31, 19%) and *FAT2* (6/31, 19%) (Figure 2.1B). Our list of frequently mutated genes included previously known oesophageal cancer associated oncogenes and tumour-suppressor genes, similar to the genes identified in different studies from Asian, Brazilian, and Malawian OSCC cohorts [22, 123-125, 128, 131, 135, 136, 139, 140, 143, 296].

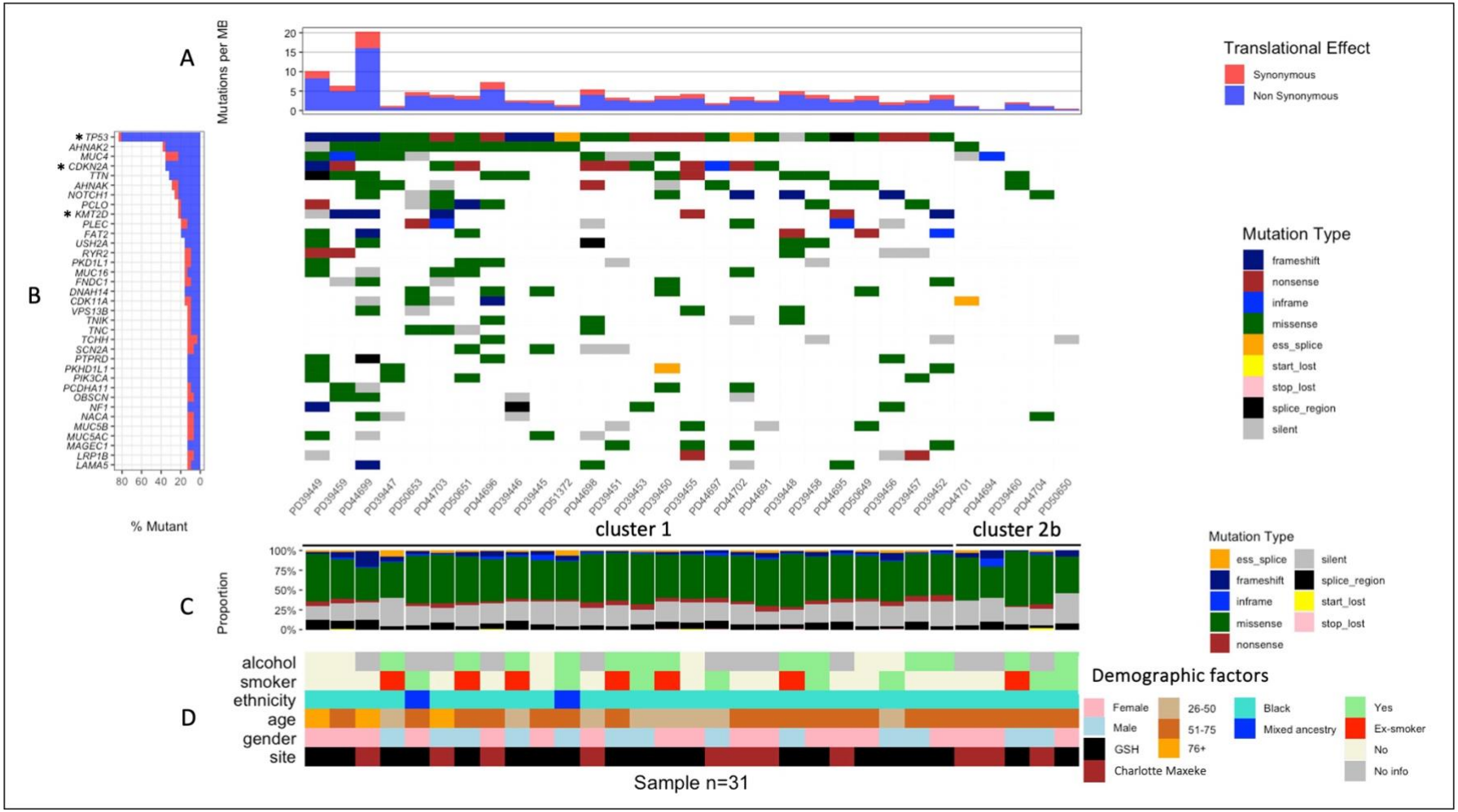


Figure 2.1 Genome alterations in OSCC patients by WGS.

Frequently mutated genes and distribution of mutations among 31 cases of OSCC. **A**) Mutation rates of synonymous and non-synonymous mutations in 31 OSCC tumours, presented as the number of mutations per megabase (Mb) of covered target sequence. Non-synonymous mutations included frameshift, nonsense, in frame, missense, indels and substitutions at splice sites; essential splice and splice region, start lost and stop lost mutations. Synonymous mutations represent silent mutations. **B**) The middle panel shows the mutation landscape across the analysed tumours, showing various mutation types, colour-coded differently. Essential splice (orange),

frameshift (navy blue), in-frame (blue), missense (green), nonsense (red), silent (grey), splice region (black), start lost (yellow) and stop lost (pink). Each row represents a gene, while columns depict individual samples, emphasizing mutual exclusivity among gene mutations. The left panel shows the percentages of tumours harbouring mutations in the top 35 frequently mutated genes. Three genes marked with an asterisk (*), *TP53*, *CDKN2A* and *KM2TD*, were identified as significantly altered genes (driver genes) ($q < 0.05$) using the dNdSCV method. Two classes of mutational profiles were identified, cluster 1 and cluster 2b. **C**) Percentage distribution of mutation types (essential splice, frameshift, in-frame, missense, nonsense, silent, splice region, start lost and stop lost) across the samples. **D**) Epidemiological data, including gender, site of collection, smoking status, ethnicity, alcohol consumption, and age of the OSCC patients.

Based on the mutation spectra analysis of 31 genomes, our samples separated into two distinct clusters labelled cluster 1 and cluster 2b (Figure 2.1B). (We refer to one of these as "cluster 2b" due to the presence of two distinct subgroups within cluster 2, which will be clear in section 2.2.3.1). These clusters were differentiated primarily by the presence of *TP53* alterations and the frequency of mutations per Mb. Within cluster 1, tumours shared common features such as a relatively high somatic mutations per Mb and mutations in *TP53*. The majority of our samples (26 out of 31) were in cluster 1. On the other hand, cluster 2b (comprising 5 out of 31 samples) showed no *TP53* mutations across all samples and displayed fewer genomic alterations. These five samples were consistently among tumours with the fewest somatic mutations per Mb (Figure 2.1B). Previous studies in Malawian and African American cohorts have reported molecular OSCC subtypes, notably, similar subtypes lacked *TP53* mutations and displaying fewer genomic alterations [22, 132, 137, 297].

In order to predict candidate cancer driver genes, i.e. genes under positive selection, we used the widely used dNdSCV approach, to calculate dN/dS ratio, which is the normalised ratio of non-synonymous to synonymous mutations [137, 141, 149, 155, 156, 293, 298, 299]. Based on this approach, *TP53*, *CDKN2A.p16INK4a*, *CDKN2A.p14ARF* and *KMT2D* were found to be statistically significantly mutated, with q values < 0.05 (Figure 2.1B, Table 2.2).

Table 2.2 Driver genes in 31 OSCC samples.

	q-value
TP53	0.000000e+00
CDKN2A.p16INK4a	0.000000e+00
CDKN2A.p14ARF	3.045938e-06
KMT2D	2.157983e-03

A q-value is a modified p-value and it gives the proportion of false positives among all the positive results in a hypothesis test. It is also interpreted as False Discovery Rate (FDR): the proportion of false positives among all positive results. Genes with FDR q-value of ≤ 0.05 were considered to be significantly mutated.

Most *TP53* mutations occurred predominantly in the DNA-binding domain (Figure 2.2A). *KMT2D* mutations clustered throughout the protein, and they are mainly protein truncating variants including in-frame insertions/deletions and nonsense variants (Figure 2.2B). Both p14ARF and p16INK4a proteins are encoded by *CDKN2A* and they both function in tumour

suppression [167]. These proteins have unique first exons, exon 1α for p16INK4a and exon 1β for p14ARF, spliced into common exons 2 and 3 [163, 164, 300]. Most of the *CDKN2A* mutations were located in exon 2, potentially affecting the functions of both p14ARF and p16INK4a proteins. Our analysis revealed differences in molecular consequences of *CDKN2A* variants on *p16INK4a* and *p14ARF*, reflecting the use of different open-reading frames [196, 301]. The majority of the p16INK4a exon 2 mutations truncate the protein (Figure 2.3A). Variant p.R58* (p16INK4a) was observed in three different patients (Figure 2.3A). This alteration results in a premature truncation of the p16INK4a protein, and is predicted to confer a loss of function of the protein, disrupting its tumour suppressive role and predispose individuals to cancer [177]. In comparison, our results show that the majority of p14ARF exon 2 mutations results mainly in missense mutations (Figure 2.3B).

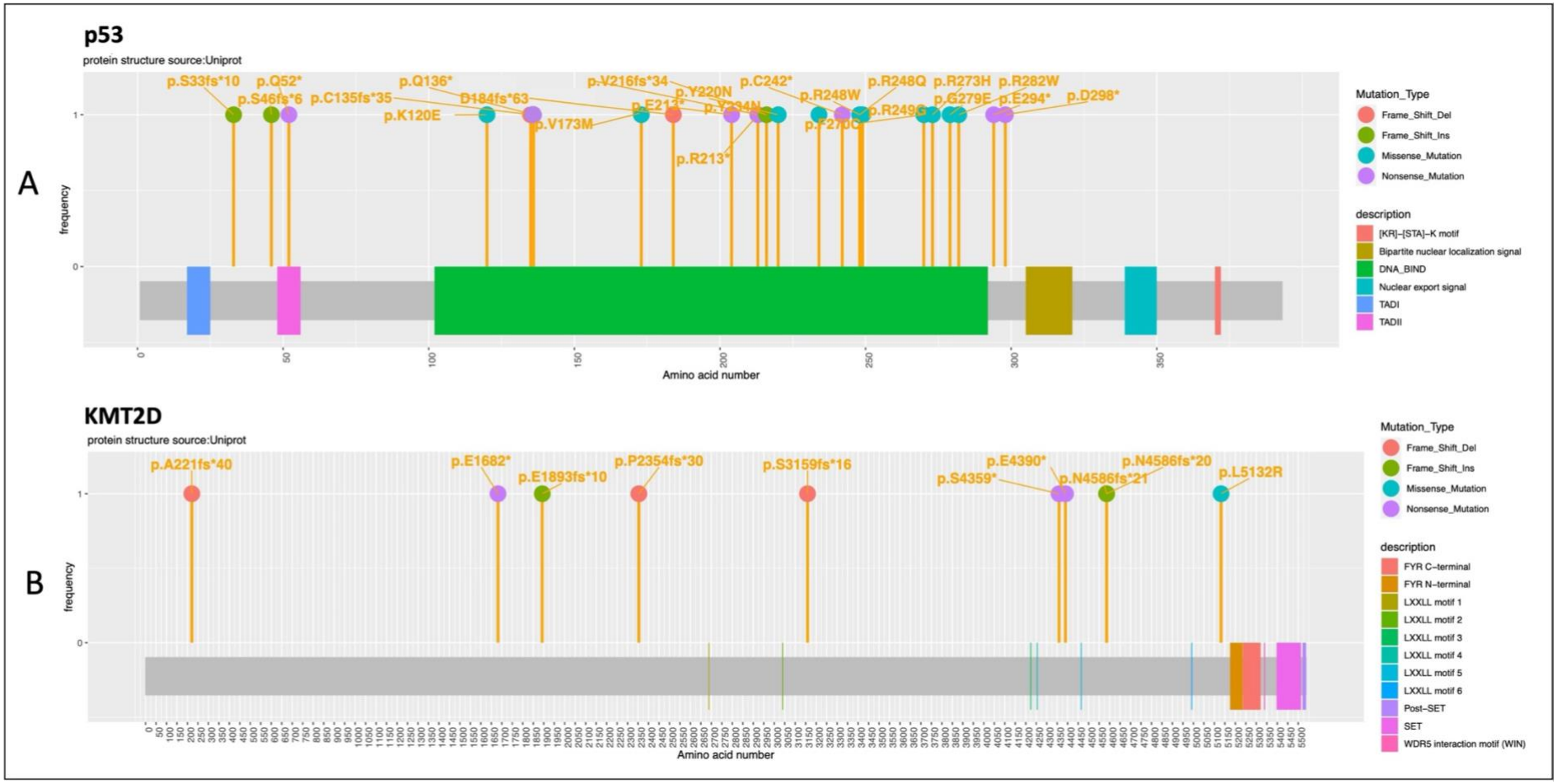


Figure 2.2 Distribution of mutations in driver genes (p53 and KMT2D) in OSCC patients.

A schematic illustration shows the domain structure of proteins p53 and KMT2D proteins, highlighting the distribution of mutation identified by our WGS. Pins on the figure indicate mutations. The Y-axis represents frequency of samples with the mutation, while the X-axis provides a schematic representation of the protein, indicating the location of important domains and regions for both p53 and KMT2D proteins. The right panel shows the type of mutations found in each protein and protein domains along the protein structure. **A**) p53 protein. Mutations are indicated by colour-coding ( ) - frameshift deletion = 2, ( ) - frameshift insertion = 3, ( ) - missense = 11 and ( ) - nonsense = 7, while **B**) KMT2D protein in ( ) -frameshift deletion = 3, ( ) - frameshift insertion = 3, ( ) - missense = 1 and ( ) - nonsense = 3.

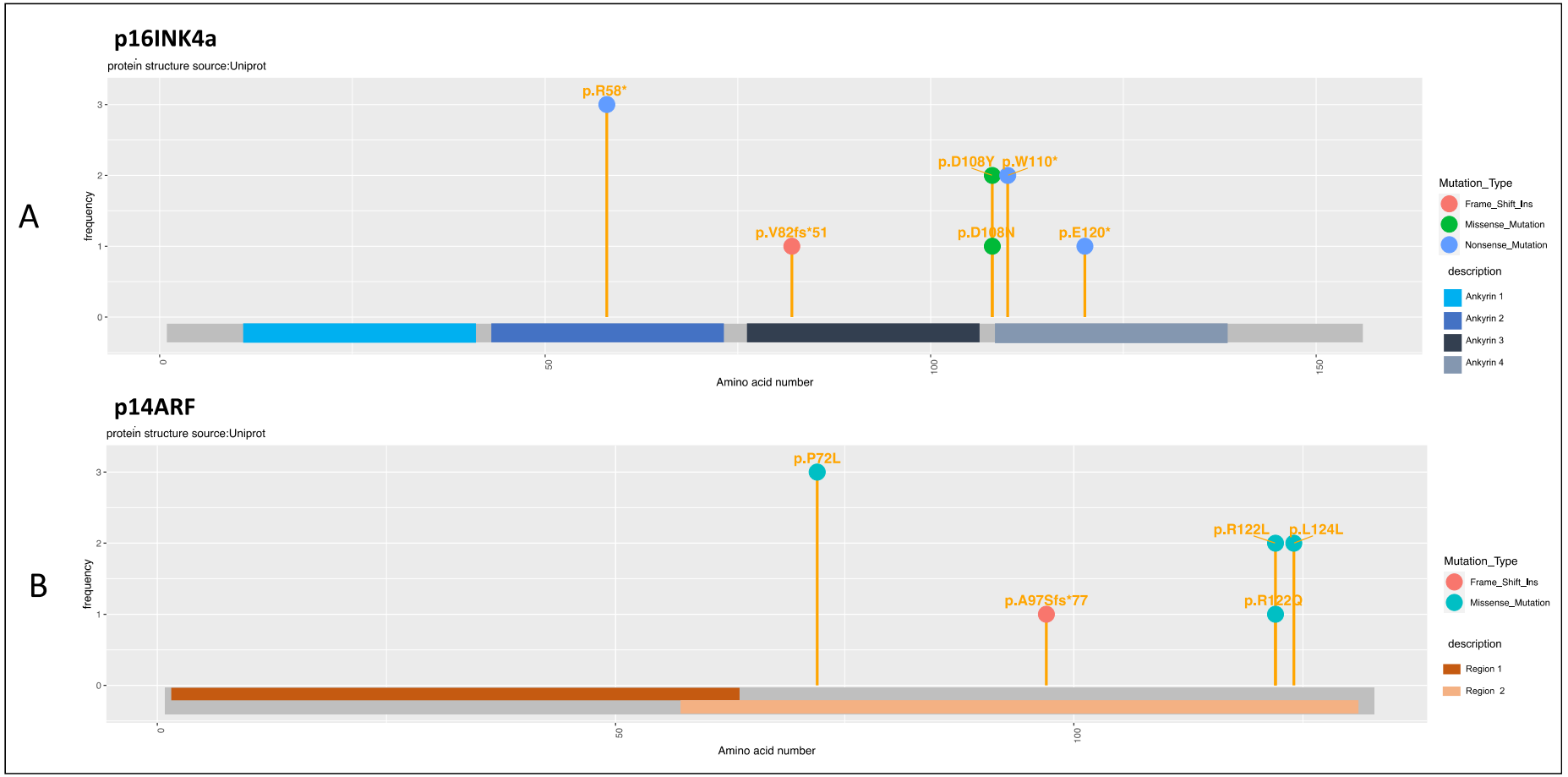


Figure 2.3 Distribution of mutations in driver genes (p16INK4a and p14ARF) in OSCC patients.

A schematic illustration shows the domain structure of proteins p16INK4a and p14ARF proteins, highlighting the distribution of mutation identified by our WGS. Pins on the figure indicate mutations. The Y-axis represents frequency of samples with the mutation, while the X-axis provides a schematic representation of the protein, indicating the location of important domains and regions for both p16INK4a and p14ARF proteins. The right panel shows the type of mutations found in each protein and protein domains along the protein structure. The right panel shows the type of mutations found in each protein and protein domains along the protein structure. **A**) p16INK4a protein. Mutations are indicated by colour-coding ( ) - frameshift insertion = 1, ( ) - missense = 2 and ( ) - nonsense = 3, while **B**) p14ARF protein in ( ) - frameshift insertion = 1, ( ) - missense = 4.

Figure 2.4 Transition and transversion mutations in OSCC patients.

Whole-genome mutation spectra in in coding regions of 31 oesophageal squamous cell carcinoma cases. **A)** Frequency and distribution of the six substitution subtypes across 31 samples. Single-nucleotide substitutions were classified into six subtypes, each category was represented by a different colour. T:A>G:C ( ), T:A>C:G ( ), T:A>A:T ( ), C:G> T:A ( ), C:G>G:C ( ), C:G>A:T ( ). **B)** Normalised proportion of substitution subtypes across 31 samples. C:G> T:A transitions were the most common mutations observed, followed by C:G>A:T and C:G>G:C transversions within our cohort.

The transition and transversion rate analysis in coding sequences of our samples showed that C:G> T:A transitions were the most common mutations, followed by C:G>A:T and C:G>G:C transversions (Figure 2.4), similar to previous studies of OSCC [131, 138]. The C:G> T:A transitions predominance is consistent with spontaneous cytosine deamination [227] being a major mutagenic process in oesophageal squamous cell carcinoma, as previously observed in OSCC [77, 122, 123, 125, 128, 131, 137, 138, 142, 154].

2.2.2.2 Mutation signatures

Mutation signatures are patterns of mutations found in DNA sequences, reflecting the DNA-damaging agents that the cells may have been exposed to [144-146]. Mutation signatures are often associated with specific mutagenic processes such as defective DNA repair, UV radiation, or tobacco smoking or exposure to other mutagens. Various mutation processes generate unique combinations of mutation types, referred to as 'signatures' [144-146]. Each mutation signature represents a distinct combination of mutation types and their relative frequencies across the genome, moreover, providing insights into the causes of mutations and aiding in understanding cancer development, environmental exposures, and evolutionary processes [146]. To better understand the mutational mechanisms in OSCC in the South African patients, single base substitution mutation signature analysis was done on the 31 genomes. The analysis identified eight mutation signatures, including known signatures SBS1, SBS2, SBS13, SBS5, and SBS6, as well as three novel, unknown signatures labelled as unknown A, unknown B, and unknown C (Figure 2.5). The age-dependent mutation signature, SBS1, is characterized by an enrichment of C>T transitions at NpCpG trinucleotides due to the spontaneous deamination of 5-methyl-cytosine [146, 302]. Signature SBS1 was detected in 52% of the analysed samples, occurring in 16 out of 31 cases. SBS5, like SBS1, represents a clock-like mutation signature. Clock-like mutation signatures refer to specific patterns of mutations observed in genomes that accumulate over time and correlate with the age of the individual [241]. These signatures are thought to reflect the natural aging processes and the cumulative exposure of cells to endogenous mutation processes, such as DNA replication errors or spontaneous chemical changes within cells [241]. SBS5 is characterized by a flat pattern, predominantly featuring C>T and T>C transitions [241]. This signature was found in 2 of 31 samples (Figure 2.5).

SBS2 and SBS13 are characterized by C>G mutations and C>T mutations, respectively, occurring at TpCpN trinucleotides [146] (Figure 2.5). These mutation signatures (SBS2 and SBS13) are associated with APOBEC enzymatic deamination of cytidine to uracil in the RNA, leading to the formation of a premature stop codon and the synthesis of a truncated protein [146, 303]. Mutation signature SBS2 is usually found in the same samples as SBS13 [146, 230]. Our findings are consistent with this pattern, as SBS2 and SBS13 mutation signatures accounted for mutations in 77% (24 out of 31) of the samples (Figure 2.5).

Mutation signature SBS6 is characterized predominantly by C>T at NpCpG trinucleotides, but is distinct from SBS1 [146] (Figure 2.5). SBS6 is associated with defective DNA mismatch repair and microsatellite instability [146]. Two patients presented a mutational pattern of SBS6 (Figure 2.5).

Separating mutations produced by two correlated signatures or those that closely resemble one another presents a challenge in mutation signature extraction [230, 287]. Signatures SBS1 and SBS5 serve as good examples [304], as their mutation loads show a positive correlation with patient age and, in some cases, with each other [241]. Furthermore, in most cancer types, at least two mutation signatures were observed, with a maximum of six signatures in cancers of the liver, uterus and stomach [146]. Similarly, our results revealed samples exhibiting combinations of two or three individual mutation signatures. Among them, fifteen samples (48%) showed enrichment of mutations from both SBS1 and APOBEC l signatures (SBS2 and SBS13), characterized by patterns enriched with C>G mutations at TpCpN trinucleotides and C>T mutations at NpCpG and TpCpN trinucleotides (Figure 2.5). Additionally, one sample displayed both SBS5 and APOBEC signatures (SBS2 and SBS13) defined by C>G and C>T at TpCpN trinucleotides and T>C mutations (Figure 2.5).

The mutation signatures labelled as unknown A-C showed low similarity to any of the COSMIC mutation signatures and, to our knowledge, have not yet been previously identified. These signatures appear to represent novel mutation patterns in OSCC (Figure 2.5). Each of these novel mutation signatures was observed in only one tumour sample. Mutation signature unknown A is defined by T>G at ApTpA, ApTpT, TpTpA and TpTpT trinucleotides. Mutation signature unknown B predominantly shows C>T transitions. Mutation signature unknown C shows a flat pattern, with a predominant presence of both C>T and T>C transitions.

Figure 2.5 Mutation signatures in OSCC patients.

Characterization of eight mutation signatures in OSCC patients. Eight mutation signatures (SBS1, SBS2, SBS13, SBS5, and SBS6, as well as three novel, unknown signatures labelled as unknown A, unknown B, and unknown C) were identified across the OSCC genomes. Each mutation signature is displayed according to the 96-substitution classification, defined by substitution class (C>A, C>G, C>T, T>A, T>C, and T>G), as well as the nucleotides immediately 5' and 3' to the mutated base. The six substitution classes are indicated on the top of the plot, color-coded as follows (C>A (blue), C>G (black), C>T (red) T>A (grey) T>C (green) and T>G (pink). Each vertical bar indicates the proportion of mutations of a particular mutation type—a single base mutation from a C or T in the context of its immediately preceding and following bases. The y-axes indicate the proportion of mutations of each type, while the x-axes display mutation types of the 96 trinucleotide substitutions.

Next, we correlated different mutation signatures with patients' epidemiological data. However, due to the limited sample size of our WGS cohort and the small subgroup sizes per identified signature, we could not make definitive conclusions. Samples displaying signature SBS6 comprised non-smoking females (n = 2) and exhibited a high mutation burden per Mb (Figure 2.1). These findings are consistent with previous studies linking SBS6 to defective DNA mismatch repair and microsatellite instability [146]. Microsatellite instability is characterized by genetic alterations in specific repetitive DNA sequences called microsatellites. These microsatellites, also called short tandem repeats (STRs), consist of repeated sequences of 1–6 nucleotides. Due to their repetitive nature, microsatellites are particularly prone to errors during the DNA replication process [305]. SBS6 is associated with high numbers of short indels (shorter than 3bp) at mono/polynucleotide repeats [146]. Indeed, we observed multiple small indels in samples with SBS6.

An unknown signature, labelled as signature unknown A (unknown signatures showed no similarity to any of the COSMIC mutation signatures) was found in a 79-year-old, non-smoker female. This tumour exhibited a mutation spectrum completely different from that observed in other tumours, with high T > G transversions and ranking among the samples with the highest tumour mutation burden per Mb (Figure 2.1, Figure 2.4). In contrast, the two samples; PD51372 and PD44694, which displayed the other two unknown signatures (unknown B and unknown C, respectively), were observed with the lowest tumour mutation burden per Mb among other samples (Figure 2.1). Additionally, sample PD44694 had the fewest C:G>A:T mutations and the highest T:A>G:C mutations in our cohort (Figure 2.4). The implications of these observations are currently uncertain.

In our cohort of 31 OSCC cases, we found that 48% of patients were smokers or had a smoking history, while 42% reported alcohol consumption. Despite this, we did not observe the smoking-related mutational signatures (SBS4 and SBS29 typically associated with C>A mutations) nor the alcohol-related signature (SBS16, associated with T>C substitutions) [146, 230, 306]. Furthermore, we found no difference when comparing the proportion of C>A transversions between smoking and non-smoking individuals. These findings may be attributed to the relatively small sample size in our WGS cohort, which could have lacked sufficient statistical power to detect subtle associations between exposures and mutational signatures.

2.2.2.3 Significantly affected biological pathways

To examine the biological functions of the frequently mutated genes, we conducted pathway enrichment analysis using the Reactome pathway database webtool (version 86) [294]. We integrated the top 70 frequently mutated genes from our WGS cohort (Table 2.3) into pathway analysis to identify the pathways significantly affected in OSCC. This analysis revealed several signalling pathways that were significantly affected (p-value < 0.05, FDR <0.05) (Figure 2.6). Notably, pathways such as "cellular responses to stimuli", "gene expression (transcription)", "disease pathway", "metabolism of proteins", "extracellular matrix organization", "signal transduction" and "immune system" were the most affected (Figure 2.6). Among these pathways "cellular responses to stimuli", "NOTCH signal transduction", "programmed cell death", "cell cycle" and "gene expression" have all been previously shown to be associated with the development and progression of OSCC [77, 123, 124, 127, 128, 132, 135, 136, 138].

Table 2.3 The top 70 frequently mutated genes in 31 OSCC samples.

1. TP53	24. PTPRD	47. FLG
2. AHNAK2	25. PKHD1L1	48. DMD
3. MUC4	26. PIK3CA	49. CSMD3
4. CDKN2A	27. PCDHA11	50. CSMD1
5. TTN	28. OBSCN	51. CNTNAP2
6. AHNAK	29. NF1	52. CNNM2
7. NOTCH1	30. NACA	53. CHD8
8. PCLO	31. MUC5B	54. CACNA1C
9. KMT2D	32. MUC5AC	55. C3
10. PLEC	33. MAGEC1	56. AMY2B
11. FAT2	34. LRP1B	57. ADGRV1
12. USH2A	35. LAMA5	58. ADGRB3
13. RYR2	36. IGFN1	59. ZNF316
14. PKD1L1	37.HSPG2	60. ZNF268
15. MUC16	38. HMCN2	61. ZFHX3
16. FNDC1	39. HECTD4	62. ZC3H4
17. DNAH14	40. GOLGA3	63. VWA5A
18. CDK11A	41. FSIP2	64. UBR5
19. VPS13B	42. FLT4	65. TRIOBP
20. TNIK	43. FLG	66. TOX4
21. TNC	44. DMD	67. TMEM266
22. TCHH	45. CSMD3	68. TMEM132D
23. SCN2A	46. FLT4	69. TG
		70. TAS2R43

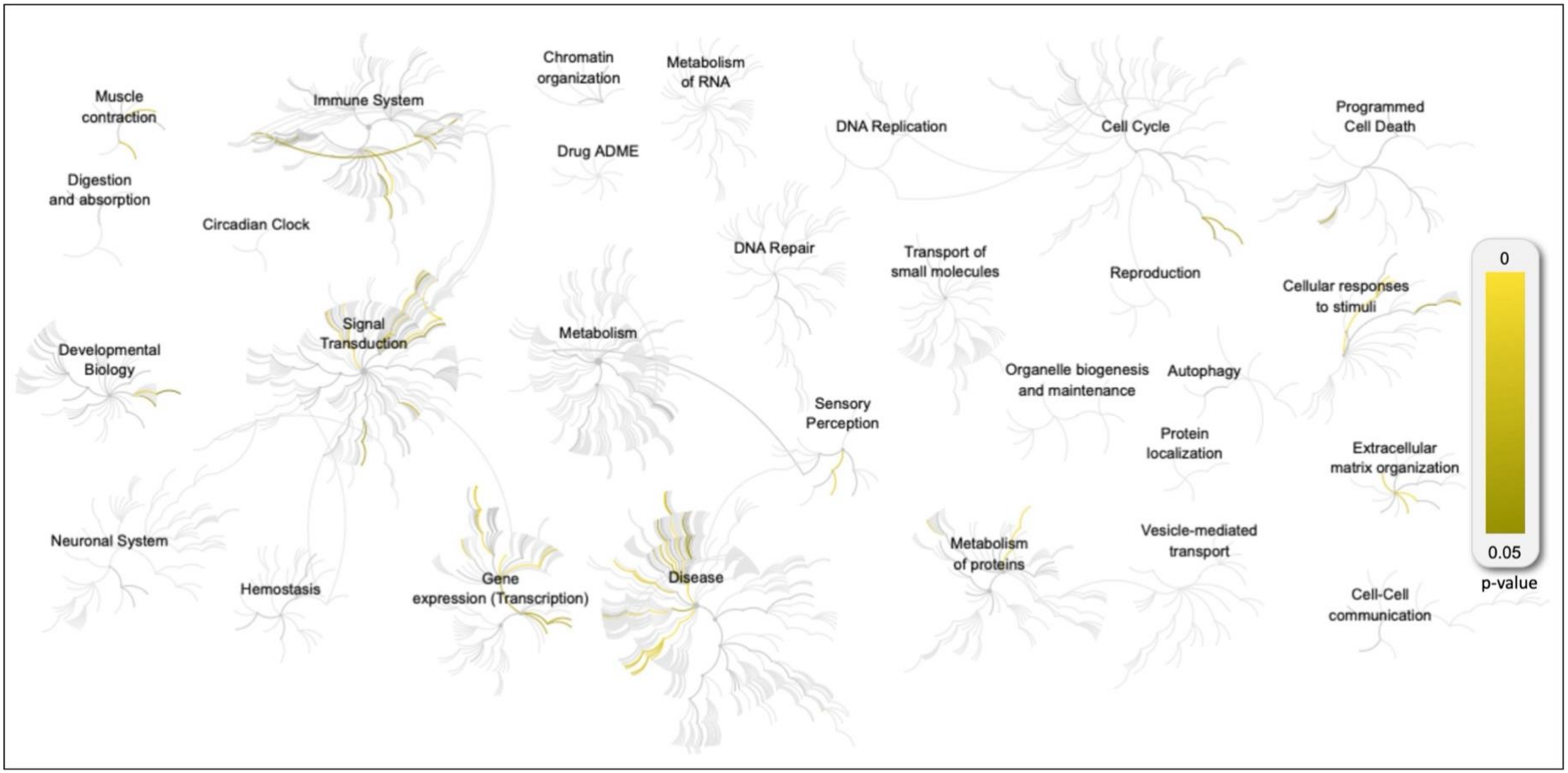


Figure 2.6 Significantly affected pathways in OSCC.

Pathways significantly affected from our gene list, depicted with a colour-coded scale indicating their respective p-values. Analysis excluded protein-protein interactors from the IntAct database to ensure pathways identified were curated by Reactome and held biological significance. Significant pathways (p-value < 0.05, FDR <0.05) included "cellular responses to stimuli", "gene expression (transcription)", "disease pathway", "metabolism of proteins", "extracellular matrix organization", "signal transduction" and "immune system". Notably, eighteen Reactome pathways, such as "digestion and absorption", "neuronal system", "circadian clock", "hemostasis", "chromatin organization", "drugADME", "metabolism", "metabolism of RNA", "DNA replication", "DNA repair", "transport of small molecules", "organelle biogenesis and maintenance", "reproduction", "autophagy", "protein localization", "vesicle-mediated transport" and "cell-cell communication", showed no association with the altered genes in our cohort.

A

Cellular responses to stimuli	Oncogene Induced Senescence Oxidative Stress Induced Senescence Cellular Senescence
Gene expression (Transcription)	Transcriptional Regulation by *VENTX* Transcriptional regulation by *RUNX3* Regulation of *TP53* Expression *RUNX3* regulates *CDKN1A* transcription Regulation of *TP53* Expression and Degradation
Disease	Defective *GALNT12* causes CRCS1 Defective *GALNT3* causes HFTC Defective *C1GALT1C1* causes TNPS *FGFR1* mutant receptor activation Diseases associated with O-glycosylation of proteins Signalling by *FGFR1* in disease Diseases of cellular response to stress Diseases of Cellular Senescence Signalling by *FGFR1* amplification mutants Diseases of glycosylation
Metabolism of proteins	Termination of O-glycan biosynthesis O-linked glycosylation of mucins
Extracellular matrix organization	Non-integrin membrane-ECM interactions
Signal transduction	*NOTCH4* Intracellular Domain Regulates Transcription PI-3K cascade: *FGFR1* Pre-NOTCH Transcription and Translation Downstream signalling of activated *FGFR1*
Immune system	Dectin-2 family

B

ZFHX3	*ZNF268*	*DMD*	*FLG-P11362-19*	*FLG-P11362-1*	*FLG-P11362*	*FLT4*	*HSPG2*	*LAMA5*	*MUC5AC*	*MUC5B*	*PIK3CA*	*TNC*	*TNIK*	*MUC16*	*NOTCH1*	*CDKN2A-Q8N726*	*CDKN2A-P42771*	*MUC4*	*TP53*	
																✖	✖		✖	Oncogene Induced Senescence
													✖			✖	✖		✖	Oxidative Stress Induced Senescence
																	✖		✖	Transcriptional Regulation by *VENTX*
									✖	✖				✖				✖		Defective *GALNT12* causes CRCS1
									✖	✖				✖				✖		Defective *GALNT3* causes HFTC
									✖	✖				✖				✖		Defective *C1GALT1C1* causes TNPS
													✖			✖	✖		✖	Cellular Senescence
									✖	✖				✖				✖		Termination of O-glycan biosynthesis
	✖														✖	✖			✖	Transcriptional regulation by *RUNX3*
			✖	✖	✖						✖									*FGFR1* mutant receptor activation
									✖	✖				✖	✖			✖		Diseases associated with O-glycosylation of proteins
			✖	✖	✖						✖									Signalling by *FGFR1* in disease
																			✖	Regulation of *TP53* Expression
																✖	✖			Diseases of cellular response to stress
																✖	✖			Diseases of Cellular Senescence
			✖	✖																Signalling by *FGFR1* amplification mutants
		✖					✖	✖				✖								Non-integrin membrane-ECM interactions
						✖									✖					*NOTCH4* Intracellular Domain Regulates Transcription
									✖	✖				✖				✖		Dectin-2 family
			✖	✖							✖									PI-3K cascade: *FGFR1*
									✖	✖				✖				✖		O-linked glycosylation of mucins
✖																			✖	RUNX3 regulates *CDKN1A* transcription
							✖		✖	✖				✖	✖			✖		Diseases of glycosylation
															✖				✖	Pre-NOTCH Transcription and Translation
															✖				✖	Downstream signalling of activated *FGFR1*
																✖			✖	Regulation of *TP53* Expression and Degradation

Figure 2.7 Gene involvement in significantly affected pathways in OSCC.

The altered genes involvement within various significantly affected pathways. **A**) Sub-pathways or molecular events involved in the seven pathways significantly affected ("cellular responses to stimuli", "gene expression (transcription)", "disease pathway", "metabolism of proteins", "extracellular matrix organization", "signal transduction" and "immune system") **B**) Genes involved in the affected sub-pathways, with ✘ indicating the corresponding gene (X-axis) and the corresponding sub-pathway (Y-axis) involved. Gene names and their corresponding uniport IDs and: protein names, matched to Reactome pathways are listed as follows: *TP53* (P04637: Tumour suppressor p53), *MUC4* (Q99102: Mucin-4), *CDKN2A* (Q8N726: Tumour suppressor ARF (*p14ARF*) and P42771: Cyclin-dependent kinase inhibitor 2A (*p16INK4A*)), *NOTCH1* (P46531: Neurogenic locus notch homolog protein 1), *MUC16* (Q8WXI7: Mucin-16), *TNIK* (Q9UKE5: TRAF2 and NCK-interacting protein kinase), *TNC* (P24821: Tenascin), *PIK3CA* (P42336: Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform), *MUC5B* (Q9HC84: Mucin-5B), *MUC5AC* (P98088: Mucin-5AC), *LAMA5* (O15230: Laminin subunit alpha-5), *HSPG2* (P98160: Basement membrane-specific heparan sulphate proteoglycan core protein), *FLT4*/*VEGFR3* (P35916: Fms-like tyrosine kinase 4/Vascular endothelial growth factor receptor 3), *FLG* (P11362, P11362-1, P11362-19: Fibroblast growth factor receptor 1), *DMD* (P11532: Dystrophin), *ZNF268* (Q14587: Zinc finger protein 268) and *ZFHX3* (Q15911: Zinc finger homeobox protein 3). Statistical significance is indicated by p-value < 0.05 and False Discovery Rate (FDR) <0.05.

The "cellular responses to stimuli" pathway constituted the most significantly affected category (Figure 2.7, Table 2.4). Within this pathway, several molecular events and sub-pathways were implicated, including oncogene induced senescence, oxidative stress induced senescence and cellular senescence, and were mainly involved in response to oncogene and oxidative stress induced senescence (Figure 2.7A). Notably, the altered genes that were associated with this pathway include *TP53, TNIK* and *CDKN2A* encoded proteins, i.e. p14ARF (Q8N726) and p16INK4a (P42771) (Figure 2.7B). Similarly, these genes are involved in the "gene expression (transcription)" pathway. Molecular events and sub-pathways associated with the "gene expression (transcription)" pathway, such as transcriptional regulation by *VENTX,* transcriptional regulation by *RUNX3,* regulation of *TP53* expression, *RUNX3* regulates *CDKN1A* transcription, and regulation of *TP53* expression and degradation, primarily regulate *TP53* expression and activity and the transcription of *TP53* target genes such as *CDKN1A* (p21) (Figure 2.7, Table 2.4).

Table 2.4 Pathways and molecular events ranked by the significance levels: p-value and FDR analysis.

Pathway name	p-value	FDR*
Oncogene Induced Senescence	1.68e-07	5.86e-05
Oxidative Stress Induced Senescence	2.84e-07	5.86e-05
Transcriptional Regulation by *VENTX*	3.64e-07	5.86e-05
Defective *GALNT12* causes colorectal cancer (CRCS1)	5.72e-06	5.53e-04
Defective *GALNT3* causes Hyperphosphatemic *familial tumoral calcinosis* (HFTC)	5.72e-06	5.53e-04
Defective *C1GALT1C1* causes *TN polyagglutination syndrome* (TNPS)	6.92e-06	5.53e-04
Cellular Senescence	1.78e-05	0.001
Termination of O-glycan biosynthesis	2.12e-05	0.001
Transcriptional regulation by *RUNX3*	5.88e-05	0.003
FGFR1 mutant receptor activation	7.62e-05	0.004
Diseases associated with O-glycosylation of proteins	8.56e-05	0.004
Signalling by *FGFR1* in disease	1.82e-04	0.007
Regulation of *TP53* Expression	2.47e-04	0.008
Diseases of cellular response to stress	2.47e-04	0.008
Diseases of Cellular Senescence	2.47e-04	0.008
Signalling by *FGFR1* amplification mutants	3.84e-04	0.012
Non-integrin membrane-ECM interactions	4.15e-04	0.012
NOTCH4 Intracellular Domain Regulates Transcription	4.53e-04	0.012
Dectin-2 family	5.26e-04	0.013
PI-3K cascade: *FGFR1*	5.62e-04	0.013
O-linked glycosylation of mucins	8.09e-04	0.019
RUNX3 regulates *CDKN1A* transcription	9.72e-04	0.021
Diseases of glycosylation	0.001	0.021
Pre-NOTCH Transcription and Translation	0.002	0.03
Downstream signalling of activated *FGFR1*	0.002	0.03
Regulation of *TP53* Expression and Degradation	0.002	0.04

(p-value < 0.05, FDR <0.05). * False Discovery Rate.

The “disease pathway” involved sub-pathways and molecular events, such as defective *GALNT12* which is associated colorectal cancer (CRCS1), defective *GALNT3* which causes hyperphosphatemia familial tumoral calcinosis (HFTC), defective *C1GALT1C1* which causes TN polyagglutination syndrome (TNPS), diseases associated with O-glycosylation of proteins, diseases of glycosylation, signalling by *FGFR1* in disease, diseases of cellular response to stress and diseases of cellular senescence (Figure 2.7A, Table 2.4). The majority of these molecular events are associated with cancer development. Genes altered within the “disease pathway” include the Mucins (*MUC4*, *MUC5B*, *MUC16* and *MUC5AC*), *NOTCH1*, *PIK3CA*, *FLG* and *CDKN2A* (Figure 2.7B).

The NOTCH signalling pathway plays a significant role in various cellular processes, including cell proliferation, differentiation, and apoptosis. Dysregulation of this pathway has been implicated in the development and progression of several types of cancer including OSCC [207]. Within the NOTCH signalling pathway, two sub-pathways have been identified: the pre-NOTCH transcription and translation sub-pathway, and the *NOTCH4* intracellular domain transcriptional regulation sub-pathway. The pre-NOTCH transcription and translation sub-pathway includes the initial stages before activation of the NOTCH signalling cascade. Once activated, NOTCH signalling regulates gene expression to control various cellular processes, including differentiation, proliferation, and apoptosis [207, 307]. Altered genes involved in the NOTCH pathway were *TP53*, *NOTCH1* and *FLT4* (Figure 2.7).

The “extracellular matrix organization”, “metabolism of proteins” and “immune system” pathways were among the significantly disrupted pathways (Figure 2.7A). The tumour microenvironment plays a key role in tumorigenesis and tumour progression [308]. This dynamic environment involves several cellular components, including tumour cells, immune cells, fibroblasts, and one of the most important constituents of the tumour microenvironment; the extracellular matrix (ECM) [308, 309]. The ECM not only provides the biochemical and mechanical support for tumour progression [309], but also serves crucial roles in maintaining tissue integrity, regulating cell functions in normal cellular and tissue biology, such as providing structural support, facilitating cell adhesion and signalling, and guiding tissue development and homeostasis [310]. In our pathway analysis, the non-integrin membrane-ECM interactions molecular event was enriched in the extracellular matrix organisation pathway (Figure 2.7 A). Moreover, altered genes involved in this pathway included *DMD*, *LAMA3*, *TNC* and *HSPG2* (Figure 2.7B). Apart from their involvement in “disease pathways”

the Mucins (*MUC4*, *MUC5B*, *MUC16* and *MUC5AC*) were also implicated in the "metabolism of proteins" pathway. Molecular events within the "metabolism of proteins" pathway included the termination of O-glycan biosynthesis and the O-linked glycosylation of mucins (Figure 2.7A).

Abnormal changes in these biological processes have the potential to alter cell fate, disrupt cellular responses to external stimuli, accumulation of senescent cells, promote tissue inflammation and create a microenvironment conducive to tumourigenesis. All these events play critical roles in cancer development.

Our WGS analysis revealed 35 frequently mutated genes potentially associated with OSCC in the South African population. Among these, *TP53*, *CDKN2A.p16INK4a*, *CDKN2A.p14ARF* and *KMT2D* were identified as cancer driver genes within our cohort of 31 samples analysed. Based on the mutation spectra analysis, we identified two distinct clusters: cluster 1 and cluster 2b, primarily distinguished by the presence of *TP53* alterations and the frequency of mutations per Mb sequenced in the samples. Notably, C:G> T:A transitions and C:G>A:T transversions were the dominant mutation types across our cohort. Mutation signature analysis identified eight distinct mutation signatures (SBS1, SBS2, SBS13, SBS5, and SBS6, and three novel, unknown signatures labelled as unknown A, unknown B, and unknown C). The contributions of mutation signatures SBS1, SBS2 and SBS13 were relatively high within our cohort. Signature SBS6 suggests possible defects in DNA mismatch repair, while the three unknown signatures indicate mutation processes unique to South African OSCC (Figure 2.5). Furthermore, pathway enrichment analysis revealed that our altered genes were associated with seven biological pathways, including "cellular responses to stimuli", "disease pathways", "gene expression (transcription)", "metabolism of proteins", "immune system", "extracellular matrix organization" and "signal transduction" pathways.

To enhance the validation of mutations, driver genes, signalling pathways, and mutation signatures implicated in somatic mutations in OSCC, we conducted another analysis on a larger sample size. This involved WES of 67 pairs of matched normal and tumour samples, distinct from samples used in the initial WGS, aiming to provide a more robust understanding of the molecular landscape underlying OSCC development.

2.2.3 Profiling of OSCC in South African patients by Whole Exome Sequencing

2.2.3.1 Frequently mutated genes and driver genes

RNA was extracted from tumour tissue, matched blood samples and adjacent normal tissue from a total of 67 subjects, as outlined in Material and Methods section 6.2.1.4.

Analysis of matched exome sequences from 67 normal/tumour pairs led to the identification of a total 30699 somatic events, comprising 29642 single nucleotide substitutions. Of these 2471 were synonymous mutations, 5900 were missense and 421 were nonsense mutations. Additionally, there were 9 stop codon losses, 13 start codon losses, 814 essential splice and essential splice-region variant, and 998 indels, including 619 frameshift and 234 in-frame mutations. The remaining somatic events included mutations with no impact on the amino acid sequence, such as silent mutations. The median tumour mutation burden across the samples was 2.5 mutations per Mb, ranging 0.1–24 mutations/Mb for non-silent variants per Mb across 67 OSCC patients (Figure 2.8A).

As observed in our WGS data, missense mutations predominated as the primary mutation types in the coding region of genes (Figure 2.8C). The most frequently mutated genes among the 67 genomes were *TP53* (54 out of 67, 81%), *TTN* (25 out of 67, 37%), *NOTCH1* (18 out of 67, 27%), *MUC16* (16 out of 67, 24%), *NFE2L2* (12 out of 67, 18%), *KMT2D* (12 out of 67, 18%), *CSMD3* (12 out of 67, 18%), *PCLO* (10 out of 67, 15%), *AHNAK2* (10 out of 67, 15%) and *ZFHX4* (9 out of 67, 13%) (Figure 2.8B).

Based on the mutation spectra observed within our cohort, the samples clustered into three distinct groups: cluster 1, cluster 2a and cluster 2b (Figure 2.8B), expanding upon the two clusters identified in our WGS analysis: cluster 1 and cluster 2b. These clusters were differentiated by the frequency of *TP53* alterations and the mutations per Mb sequenced. Consistent with our WGS findings, cluster 1 tumours were characterized by *TP53* mutations and a relatively high somatic mutation rate per Mb. Most of our samples (54 out of 67 samples) were in cluster 1. Cluster 2 (13 out of 67 samples) showed no *TP53* mutations in all samples and were predominantly black female patients. In addition, cluster 2 was further subdivided into two subclusters; cluster 2a and cluster 2b. Cluster 2a showed a high mutation rate per Mb. In contrast, cluster 2b samples exhibited fewer genomic alterations, with the fewest somatic mutations per Mb (Figure 2.8B). This pattern of high or low mutation rates and presence or

absence of *TP53* mutations is consistent with observations in Malawian OSCC patients and African American OSCC populations [22, 137, 297].

Table 2.5 Driver genes in 67 OSCC samples.

	q-value
TP53	0.000000e+00
NFE2L2	3.201418e-06
CDKN2A.p16INK4a	1.714431e-03
ZNF750	8.080599e-03
NOTCH1	3.608327e-02

A q-value is a modified p-value and it gives the proportion of false positives among all the positive results in a hypothesis test. It is also interpreted as False Discovery Rate (FDR): the proportion of false positives among all positive results. Genes with FDR q-value of ≤ 0.05 were considered to be significantly mutated.

The analysis of driver genes in the 67 tumours, using the dNdSCV approach revealed *TP53*, *NFE2L2*, *CDKN2A*.p16INK4a, *ZNF750* and *NOTCH1* as statistically significant, with q values < 0.05 (Figure 2.8B, Table 2.5). Despite *CDKN2A* and *ZNF750* being significantly mutated (driver genes), their mutation frequencies within our cohort were low, thus excluding them from the top 35 frequently mutated genes observed. As expected, the majority (83%) of *TP53* non-synonymous mutations were located predominantly in the DNA-binding domain, at mutation hotspots such as R273 and R282W, similar to our WGS finding and previous reports [311] (Figure 2.9A). Additionally, other previously identified mutation hotspots in OSCC, including the mutations in *KEAP1* binding motifs (ETGE and DLG) of *NFE2L2* [219], *CDKN2A.p16INK4a* R80 and D108 [156], were observed in our cohort (Figure 2.9B and Figure 2.9C). Compared to our WGS analysis of driver genes, mutations in *p14ARF* were not identified as driver of OSCC in WES cohort, possibly due to the fewer mutations identified in *CDKN2A* within this cohort, with only eight mutations in 7 out of 67 samples. Among these mutations, two affected *p16INK4a*: [p.Q50* and p.L16fs*6], both occurring in exon 1α which encodes for *p16INK4a*. Additionally, one of the mutations detected in exon 2 of *CDKN2A*, resulted in a missense mutation in *p16INK4a*: [p.L130P]. However, in *p14ARF*, the same mutation occurred outside its coding region.

Positive selection was also observed in other OSCC-associated genes; *ZNF750* and *NOTCH1* [124] (Figure 2.10A and Figure 2.10B). The non-silent mutation rate of *ZNF750* was 7.9%, with the majority mutations identified being inactivating, which aligns with previous findings in OSCC [77, 122].

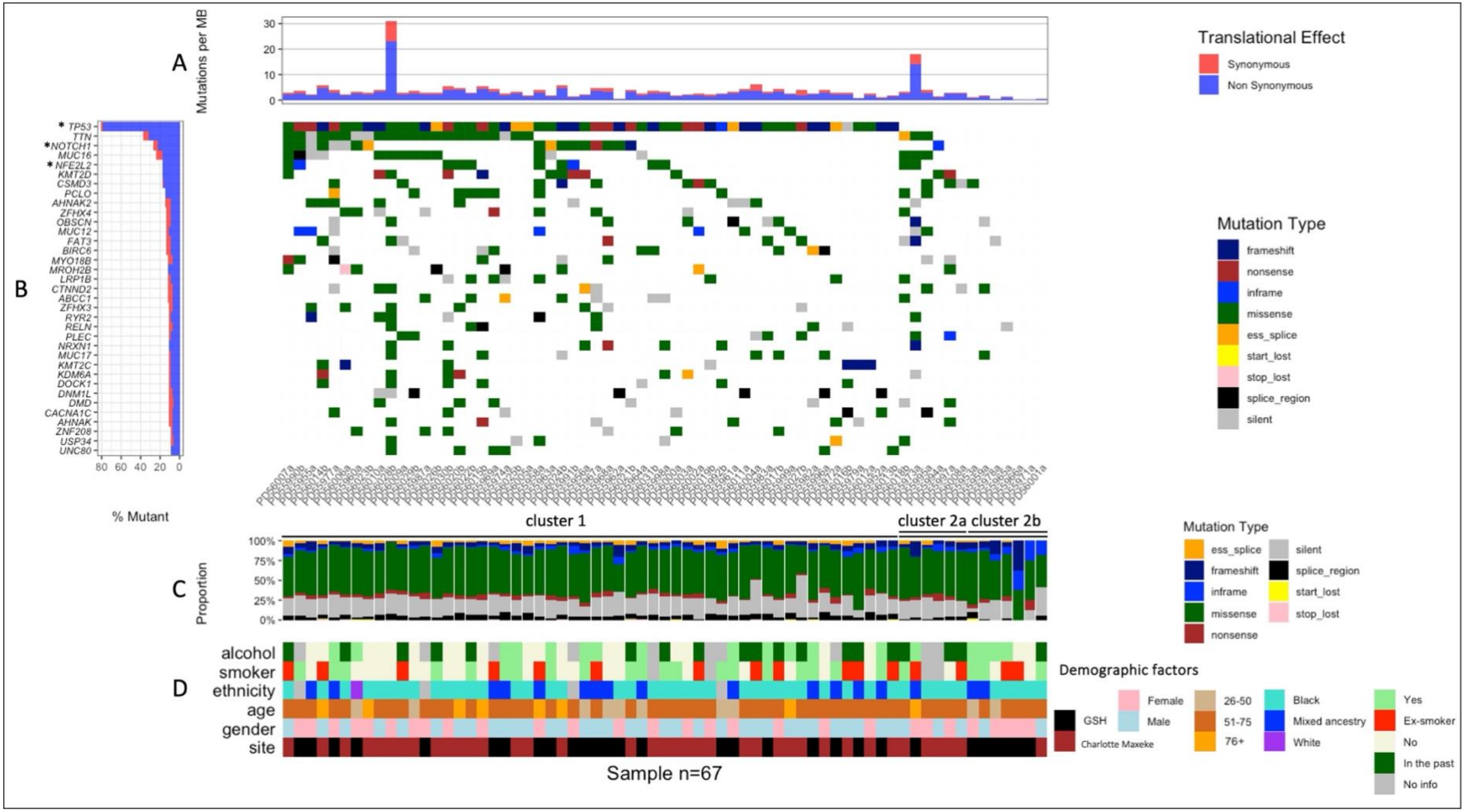


Figure 2.8 Genome alterations in OSCC patients by WES.

Frequently mutated genes and distribution of mutations in 67 OSCC exomes. **A**) Mutation rates of synonymous and non-synonymous mutations in 67 tumours, presented as the number of mutations per megabase (Mb) of covered target sequence. Non-synonymous mutations consist of frameshift, nonsense, in frame, missense, essential splice and splice region, start lost and stop lost mutations. Synonymous mutations consist of silent mutations. **B**) The middle panel illustrates the mutation landscape

across analysed tumours, showcasing various mutation types, colour-coded differently. Essential splice (orange), frameshift (navy blue), in-frame (blue), missense (green), nonsense (red), silent (grey), splice region (black), start lost (yellow) and stop lost (pink). Each row represents a gene, while columns depict individuals' samples, emphasizing mutual exclusivity among gene mutations. The left panel displays percentages of tumours harbouring mutations in the in the top 35 frequently mutated genes. In addition to *ZNF750* and *CDKN2A*, three genes marked with an asterisk (*), *TP53*, *NOTCH1* and *NFE2L2*, were identified as significantly altered genes (driver genes) ($q < 0.05$) using the dNdSCV method. Notably, *CDKN2A* and *ZNF750* were not part of the 35 frequently mutated genes within our cohort. Three classes of mutational profiles were identified, cluster 1, cluster 2a and cluster 2b, differentiated by the frequency of *TP53* alterations and the mutations per Mb sequenced. **C**) Percentage distribution of mutation types (essential splice, frameshift, in frame, missense, nonsense, silent, splice region, start lost and stop lost) across the samples. **D**) Epidemiological data, including gender, site of collection, smoking status, ethnicity, alcohol consumption, and age of the OSCC patients.

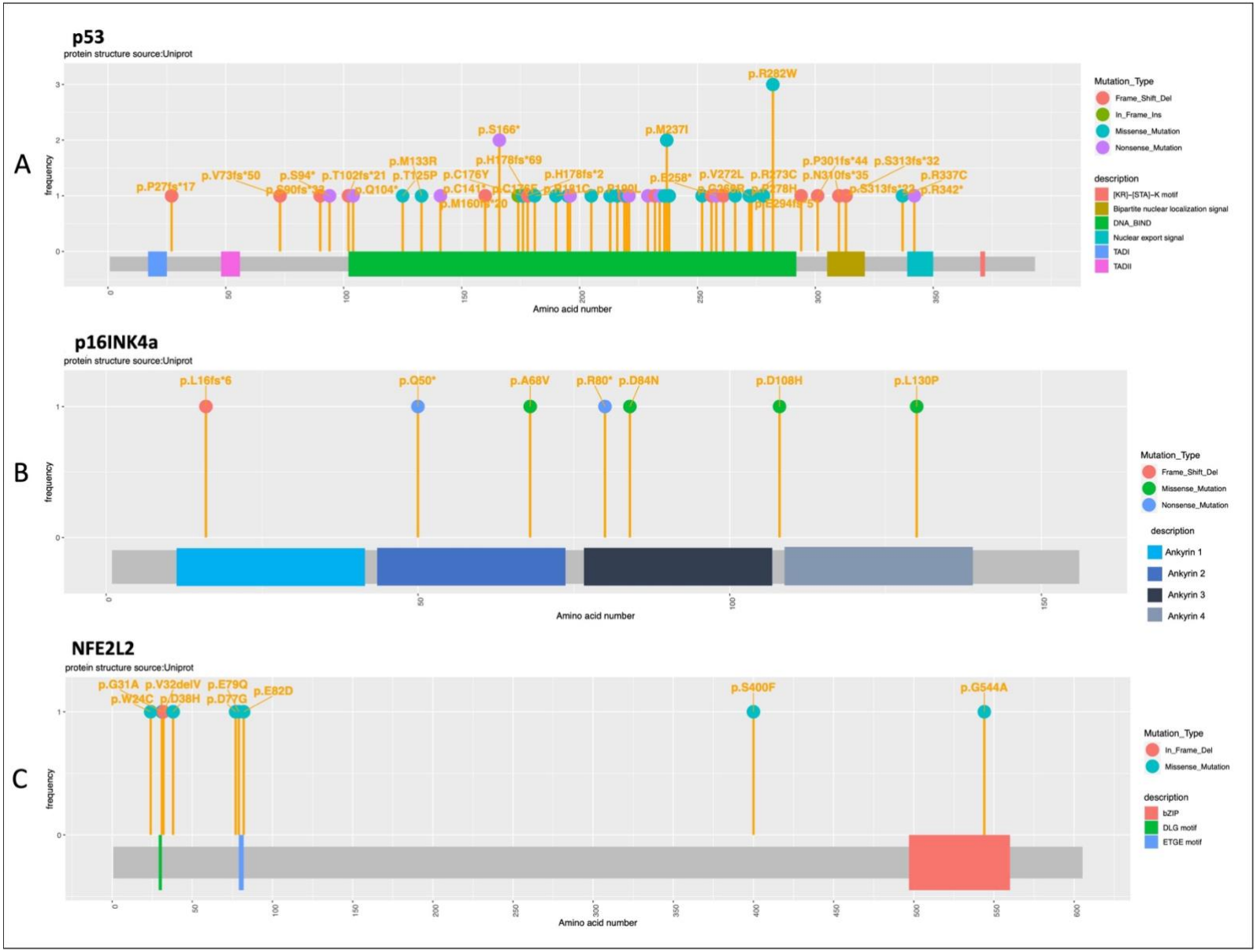

A
p53
protein structure source:Uniprot
frequency
Amino acid number
p.R282W
p.S166*
p.M237I
p.M133R
p.H178fs*69
p.H178fs*2
p.P27fs*17
p.V73fs*50
p.S94*
p.T102fs*21
p.Q104*
p.T125P
p.C176Y
p.M160fs*20
p.E258*
p.V272L
p.R273C
p.P301fs*44
p.N310fs*35
p.S313fs*32
p.R337C
p.R342*
Mutation_Type
Frame_Shift_Del
In_Frame_Ins
Missense_Mutation
Nonsense_Mutation
description
[KR]-[STA]-K motif
Bipartite nuclear localization signal
DNA_BIND
Nuclear export signal
TADI
TADII
B
p16INK4a
protein structure source:Uniprot
frequency
Amino acid number
p.L16fs*6
p.Q50*
p.A68V
p.R80*
p.D84N
p.D108H
p.L130P
Mutation_Type
Frame_Shift_Del
Missense_Mutation
Nonsense_Mutation
description
Ankyrin 1
Ankyrin 2
Ankyrin 3
Ankyrin 4
C
NFE2L2
protein structure source:Uniprot
frequency
Amino acid number
p.G31A
p.V32delV
p.E79Q
p.E82D
p.W24C
p.D38H
p.D77G
p.S400F
p.G544A
Mutation_Type
In_Frame_Del
Missense_Mutation
description
bZIP
DLG motif
ETGE motif

Figure 2.9 Distribution of mutations in driver genes (p53, p16INK4a, NFE2L2) in OSCC.

A schematic illustration represent the domain structure of p53, p16INK4a and NFE2L2 proteins, highlighting the distribution of mutation identified by our WES. Pins on the figure indicate mutations. The Y-axis represents the frequency of samples found with the mutation, while the X-axis is a schematic representation of the protein, indicating the location of important domains and regions of p53, p16INK4a and NFE2L2 proteins. The right panel shows the type of mutations found in each protein and protein domains along the protein structure. **A**) p53 protein: Mutations are indicated by colour-coding, ( ) - frameshift deletion = 16, ( ) - in-frame insertion = 1, ( ) - missense = 26 and ( ) - nonsense = 12, while, **B**) p16INK4a protein ( ) - frameshift deletion = 1, ( ) - missense = 4 and ( ) - nonsense = 2, and **C**) NFE2L2 protein ( ) - in-frame deletion = 2, and ( ) - missense = 10.

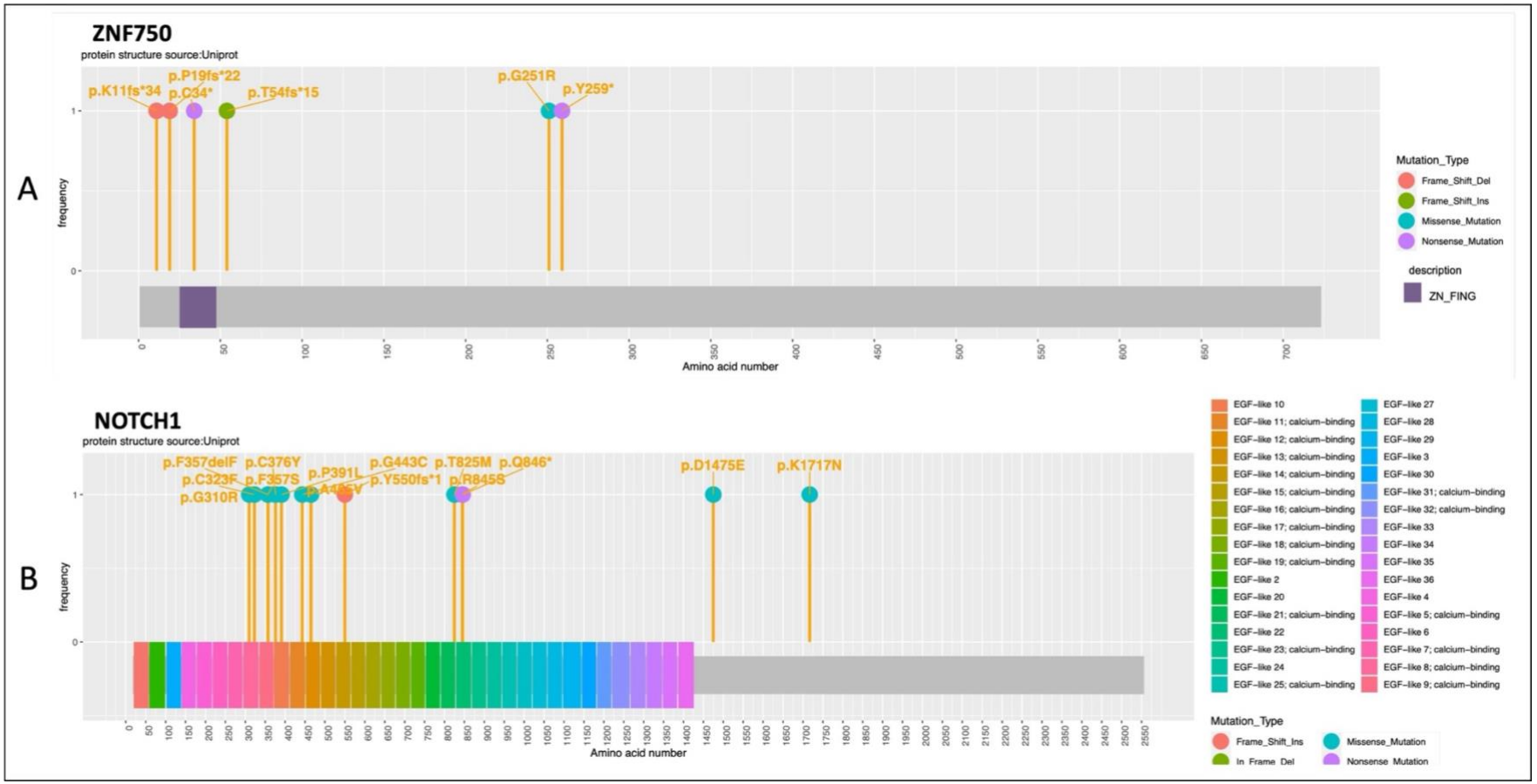


Figure 2.10 Distribution of mutations in driver genes (ZNF750 and NOTCH1) in OSCC.

A schematic illustration represents the domain structure of ZNF750 and NOTCH1 proteins, highlighting the distribution of mutation identified by our WES. Pins on the figure denote mutations. The Y-axis represents the frequency of samples found with the mutation, while the X-axis is a schematic representation of the protein, indicating the location of important domains and regions of ZNF750 and NOTCH1 protein. The right panel shows the type of mutations found in each protein and protein domains along the protein structure. **A**) ZNF750 protein. Mutations are indicated by colour-coding, ( ) - frameshift insertion = 1, ( ) - frameshift deletion = 2, ( ) - missense = 1 and ( ) - nonsense = 2, while **B**) NOTCH1 protein, in ( ) - frameshift insertion = 1, ( ) - in-frame deletion = 1, ( ) - missense = 11 and ( ) - nonsense = 1.

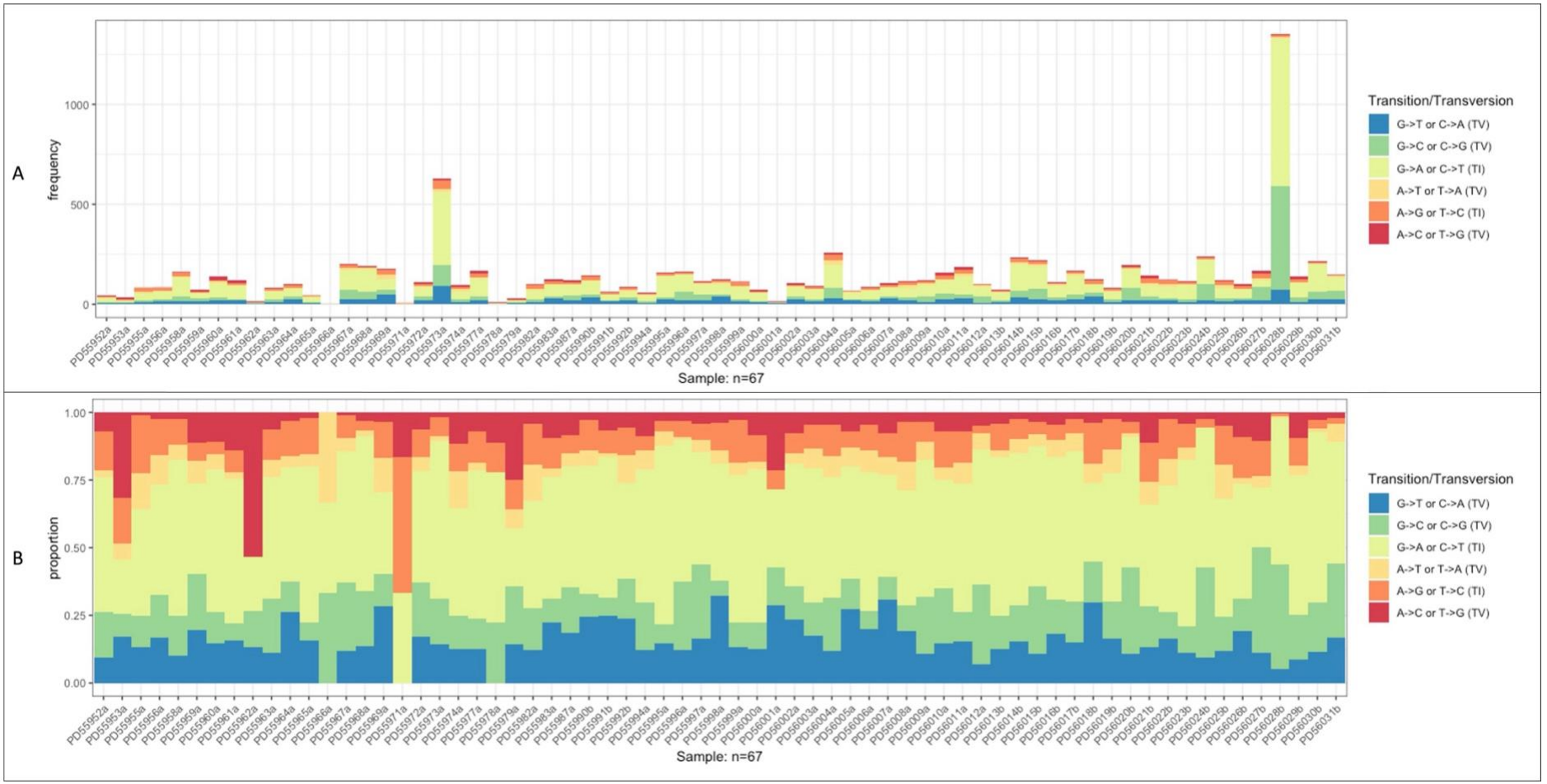


Figure 2.11 Transition and transversion mutations in OSCC patients.

The whole-exome mutation spectra in the coding regions of 67 oesophageal cancer genomes. **A**) Frequency and distribution of the six substitution subtypes in 67 samples. Single-nucleotide substitutions were classified into six subtypes and each category was represented by different colours. Substitution type (T:A>G:C ( ), T:A>C:G ( ), T:A>A:T ( ), C:G> T:A ( ), C:G>G:C ( ), C:G>A:T ( ). **B**) Normalised proportion of substitution subtypes in 67 samples.

Consistent with our WGS findings, the transition and transversion rate analysis in coding sequences of our samples revealed a high frequency of C:G> T:A transitions, followed by C:G>A:T and C:G>G:C transversions (Figure 2.11). The high C:G> T:A mutations observed in our cohort is consistent with previous reports of OSCC [77, 122, 123, 125, 128, 131, 137, 138, 142, 154].

2.2.3.2 Mutation signatures

The analysis of mutation signatures revealed seven mutation signatures (SBS1, SBS2, SBS5, SBS10b, SBS13, SBS15 and SBS55) across the 67 samples (Figure 2.12). Notably, SBS1 and SBS5 were found in all samples, and solely contributors to mutations in 45% (30 out of 67) of the samples (Figure 2.12). Both SBS1 and SBS5 are age-related mutation signatures, indicating that the number of mutations in tumours correlates with the age of the individual at diagnosis [241]. Additionally, APOBEC-mediated mutation signatures; SBS2 and SBS13 were detected in 37% (25 out of 67) and 40% (27 out of 67) of the samples, respectively (Figure 2.12). The prevalence of age-related (SBS1 and SBS5) and APOBEC-mediated mutation signatures, (SBS2 and SBS13) in our WES cohort is consistent with the significant presence of SBS1, SBS2 and SBS13 observed in our WGS cohort. Similarly, previous studies have found APOBEC-mediated mutation signatures and age-related (clock-like) signatures as most prevalent in OSCC samples [22, 77, 123-125, 128, 131, 135, 136, 138, 139, 142, 238-240], suggesting their role in OSCC mutagenesis and development.

The mutation signature SBS10b, characterized by C>T mutations (Figure 2.12C), was detected in 6% (4/67) samples. This signature is associated with altered activity of the error-prone polymerase epsilon exonuclease domain (*POLE*) [230]. On the other hand, SBS15 predominantly characterized by C>T mutations and is associated with defective DNA mismatch repair [146]. SBS15 was detected in 9% (6 out of 67) of the samples (Figure 2.12). The mutation signatures SBS10b and SBS15 associated with DNA proofreading defects observed in our WES analysis, were not found in our WGS analysis. However, within our WGS cohort, we found one signature, SBS6, which is associated with defective DNA mismatch repair [146]. These results suggest a potential association between defects in DNA repair pathways and OSCC, highlighting roles of defects DNA repair mechanisms in OSCC development. On the other hand, signature SBS55 was detected in 9% (6 out of 67) of the samples. Currently, SBS55 is considered a non-validated signature possibly arising from a

sequencing artifact [230]. Despite being detected only in our WES analysis and not in our WGS analysis, SBS55 was previously reported by Alexandrov et al., [230] using both WGS and WES techniques.

Similar to our observations in the WGS analysis, our results revealed samples exhibiting combinations of two or more individual mutation signatures, a pattern consistently observed across all samples. Notably, among them, 30 out of 67 of the samples (45%) showed enrichment of mutations from both age-related signatures, SBS1 and SBS5 (Figure 2.12). Furthermore, 28% (19 out of 67) of the samples displayed four different mutations, including SBS1 and SBS5 and APOBEC signatures (SBS2 and SBS13) (Figure 2.12). Some samples displayed three mutation signatures (SBS1, SBS5 and SBS15), while others exhibited five (SBS1, SBS2, SBS5, SBS13 and SBS15 or SBS1, SBS2, SBS5, SBS10b, SBS13 and SBS15) (Figure 2.12). The presence of combinations of mutation signatures suggests a complex interaction of various mutation processes existing within these samples. The enrichment of mutations from both age-related signatures, SBS1 and SBS5, in a large portion of the samples emphasise the influence of aging on the mutational profile in our cohort. Moreover, the co-occurrence of age-associated signatures (SBS1 and SBS5), APOBEC signatures (SBS2 and SBS13), along with other signatures within samples, highlights the diverse mechanisms contributing to mutagenesis in our cohort.

We further explored the association between different mutation signatures and patients' epidemiological data as well as their mutational profiles. The somatic mutations observed in our cohort were primarily attributed to age-related mutation signatures; SBS1 and SBS5 (Figure 2.12). Consistent with this, the majority of patients (94%, 63 out of 67) were aged from 45 to 89, with 18% (12 out of 67) of the population being 70 years old or older, while only 6% (4 out of 67) were younger than 45 years old. Additionally, all of the samples exhibiting a predominant presence of these two age-related mutation signatures were from individuals older than 45 years old, except for two samples (PD55978a and PD55961a), both from female patients aged 41 years old (Figure 2.8, Figure 2.12).

The age-related mutation signature SBS5 is characterized primarily by C>T and T>C transitions (Figure 2.12C), and its mutational burden is reported increased in many cancer types associated with tobacco smoking, such as head and neck cancer, colorectal carcinoma and lung squamous cell carcinoma [241]. Indeed, the majority of the samples (41 out of 67) displaying

high SBS5 mutational activities were either smokers or ex-smokers. Furthermore, sample PD56028b, a non-smoker female, showed the lowest contribution of SBS5 mutational burden (Figure 2.8, Figure 2.12).

SBS10b is caused by *POLE* proofreading defects. *POLE* is responsible for recognition and excision of mismatched base mutations and can consequently lead to DNA hypermutation [230]. As expected, the four patients (PD56028b, PD56024b, PD56017b, PD55996a) showing this signature tended to have a higher mutation burden (ranging from 3.4-27.2 mutations/Mb) (Figure 2.8 and Figure 2.12). Additionally, one of the four patients (PD56017b) showed a missense mutation (D2218Y) within *POLE*, although this was not observed in the other samples.

Notably, similar to our WGS analysis, we did not detect smoking-associated mutation signatures such as SBS4 and SBS29 or alcohol consumption (SBS16) in this cohort of samples. This is consistent with prior studies that also did not find these mutation signatures in their cohorts [22, 77, 128]. These signatures are characterized by C>A mutations [146, 241]. Our further analysis found no difference in the overall mutation load between the smokers and non-smokers (Figure 2.8), consistent with findings from other OSCC cohorts [77, 123-125, 127, 128, 135, 155]. However, when comparing the proportion of C>A transversions between smoking and non-smoking individuals, we found that C>A substitutions were found less frequent in non-smokers group compared to smokers (Figure 2.11). Mangalaparthi and colleagues [135] similarly found a higher frequency of C > A transversions in individuals who used chewing tobacco, followed by smokers, with the lowest frequency of C > A transversions in non-smokers [135]. Moreover, two samples from non-smokers (PD55971a and PD55978a) were observed with no C>A substitutions (Figure 2.11). On the other hand, we did not observe a similar distinction for T>C substitutions between alcohol consumers and non-consumers, further highlighting the complexity of OSCC mutagenesis and the potential influence of factors beyond smoking and alcohol consumption.

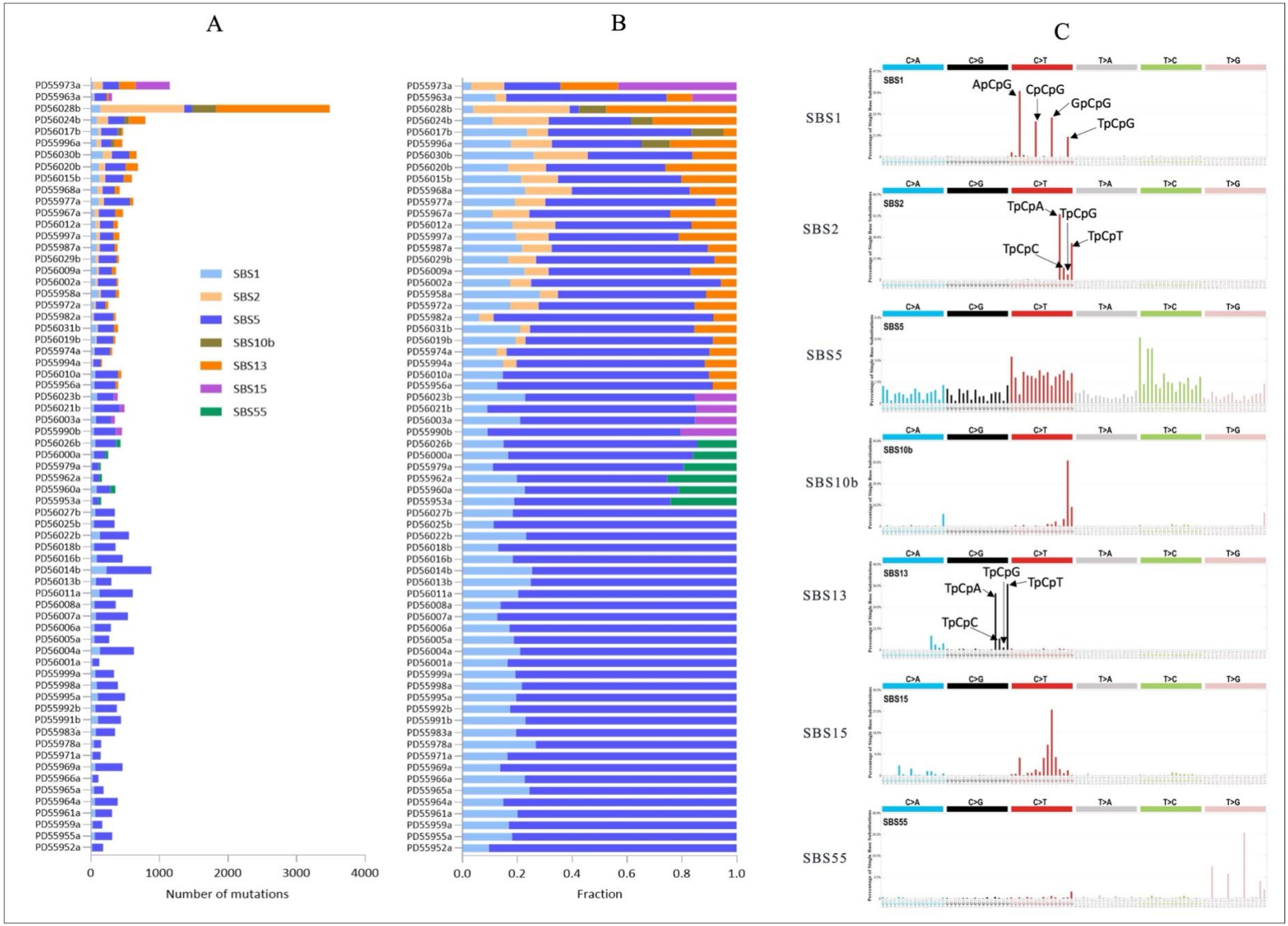

A
B
C
SBS1
SBS2
SBS5
SBS10b
SBS13
SBS15
SBS55
Number of mutations
Fraction
C>A
C>G
C>T
T>A
T>C
T>G
ApCpG
CpCpG
GpCpG
TpCpG
TpCpA
TpCpT
TpCpC

Figure 2.12 Mutation signatures in OSCC patients.

The contribution of seven mutation signatures (SBS1, SBS2, SBS5, SBS10b, SBS13, SBS15 and SBS55) to individual tumours within our cohort. **A**) Each bar on the vertical axis represents a sample, ordered by the proportion of mutation signatures observed in samples. The mutation signatures are colour-coded: SBS1 (sky blue), SBS2 (peach), SBS5 (blue), SBS10b (olive green), SBS13 (orange), SBS15 (purple), SBS55 (green). The distribution and contribution of each mutation signatures to the total number of mutations in each tumour. **B)** A normalised plot displaying the contribution of each mutation signatures c in each tumour. **C**) Patterns of the seven mutation signatures retrieved from the Catalogue Of Somatic Mutations In Cancer (COSMIC) for SBS mutation signatures (v3.3) (https://cancer.sanger.ac.uk/signatures/sbs/). Accessed on October 24, 2023.

2.2.3.3 Significantly affected biological pathways in OSCC

Pathways analysis was performed using the Reactome pathway database webtool (version 86) [294], to examine the biological functions of the frequently mutated genes in our cohort. Based on our pathway enrichment analysis using the top 70 frequently mutated genes identified through WES (Table 2.6), we found that eight biological pathways were significantly affected (p-value < 0.05, FDR <0.05). These pathways included "cellular responses to stimuli", "disease pathway", "gene expression (transcription)", "extracellular matrix organization", "metabolism of proteins", "signal transduction", "neuronal system" and "cell-cell communication" (Figure 2.13). Notably, many of these pathways were previously reported to be involved in cancer-associated pathways including "cell cycle", "histone modification pathway", "NOTCH signalling pathway", "*KEAP1-NFE2L2* pathway" [77, 122-134, 136-143], consistent with our findings from our WGS data analysis.

Table 2.6 The top 70 frequently mutated genes in 67 OSCC samples.

1. *TP53*	24. *NRXN1*	47. *PTCH1*
2. *TTN*	25. *MUC17*	48. *LRP1*
3. *NOTCH1*	26. *KMT2C*	49. *LAMA3*
4. *MUC16*	27. *KDM6A*	50. *KIAA1244*
5. *NFE2L2*	28. *DOCK1*	51. *KCNA4*
6. *KMT2D*	29. *DNM1L*	52. *HSPG2*
7. *CSMD3*	30. *DMD*	53. *FCGBP*
8. *PCLO*	31. *CACNA1C*	54. *FBN2*
9. *AHNAK2*	32. *AHNAK*	55. *FAT1*
10. *ZFHX4*	33. *ZNF208*	56. *DSP*
11. *OBSCN*	34. *USP34*	57. *DPP6*
12. *MUC12*	35. *UNC80*	58. *DNAH5*
13. *FAT3*	36. *ST18*	59. *DCAF4L2*
14. *BIRC6*	37. *SI*	60. *ASXL3*
15. *MYO18B*	38. *SHANK2*	61. *ANK3*
16. *MROH2B*	39. *RYR1*	62. *ACE*
17. *LRP1B*	40. *RTEL1*	63. *ZNF236*
18. *CTNND2*	41. *RP1*	64. *XIRP2*
19. *ABCC1*	42. *PTPRQ*	65. *USH2A*
20. *ZFHX3*	43. *PTCH1*	66. *UBR4*
21. *RYR2*	44. *LRP1*	67. *TRPM3*
22. *RELN*	45. *LAMA3*	68. *TRANK1*
23. *PLEC*	46. *PTPRQ*	69. *TENM2*
		70. *SLITRK4*

Similar to our WGS findings, the pathway most significantly disrupted was “cellular responses to stimuli” (Figure 2.14A, Table 2.7). Within this pathway, several molecular events and sub-pathways were significantly enriched, including regulation of *NFE2L2* gene expression, *NFE2L2* regulating multi-drug resistance (MDR) associated enzymes, *NFE2L2* regulating tumorigenic genes, nuclear events mediated by *NFE2L2* and KEAP*1-NFE2L2* pathway (Figure 2.14, Table 2.7). All of these pathways and molecular events were mainly involved in regulation of *NFE2L2* expression, *NFE2L2* targeted genes, and the *KEAP1-NFE2L2* pathway.

Although there was overlap in the pathways significantly disrupted between WGS and WES analysis, including “disease pathway”, “gene expression (transcription)”, “extracellular matrix organization”, “metabolism of proteins” and “signal transduction”, two additional pathways, “neuronal system” and “cell-cell communication”, were exclusively found in the WES cohort. These pathways play crucial roles in mediating interactions between tumour cells and their surrounding microenvironment, potentially influencing tumour development and progression [308, 309].

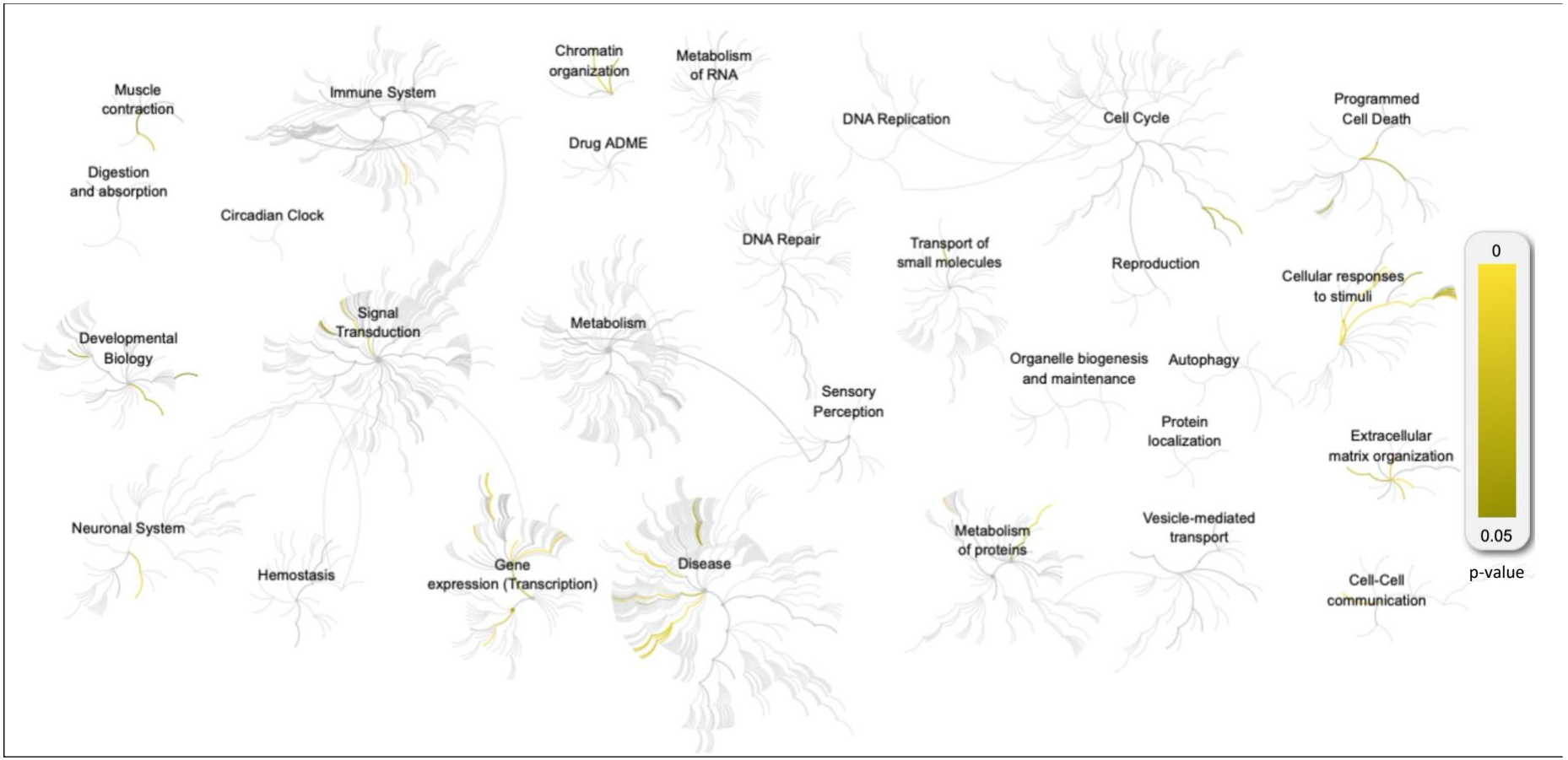


Figure 2.13 Significantly affected pathways in OSCC.

Pathways significantly affected from our gene list, depicted with a colour-coded scale indicating their respective p-values. Analysis excluded protein-protein interactors from the IntAct database to ensure pathways identified were curated by Reactome and held biological significance. Significant pathways (p-value < 0.05, FDR <0.05) included "cellular responses to stimuli", "disease pathway", "gene expression (transcription)", "extracellular matrix organization", "metabolism of proteins", "signal transduction", "neuronal system" and "cell-cell communication". Note that fourteen Reactome pathways, including "autophagy", "protein localization", "vesicle-mediated transport", "organelle biogenesis and maintenance", "reproduction", "DNA replication", "DNA repair", "metabolism of RNA", "drugADME", "sensory perception", "metabolism", "digestion and absorption", "circadian clock" and "hemostasis", showed no association with the altered genes in our cohort.

A

Cellular responses to stimuli	Regulation of *NFE2L2* gene expression *NFE2L2* regulating MDR associated enzymes *NFE2L2* regulating tumorigenic genes Nuclear events mediated by *NFE2L2* *KEAP1-NFE2L2* pathway
Disease	Defective *GALNT12* causes CRCS1 Defective *GALNT3* causes HFTC Defective *C1GALT1C1* causes TNPS Diseases associated with O-glycosylation of proteins
Gene expression (Transcription)	Regulation of *TP53* Expression *RUNX3* regulates *CDKN1A* transcription Formation of *WDR5*-containing histone-modifying complexes
Extracellular matrix organization	Non-integrin membrane-ECM interactions Extracellular matrix organization
Metabolism of proteins	Termination of O-glycan biosynthesis
Signal transduction	Pre-NOTCH Transcription and Translation
Neuronal System	Protein-protein interactions at synapses
Cell-Cell communication	Type I hemidesmosome assembly

B

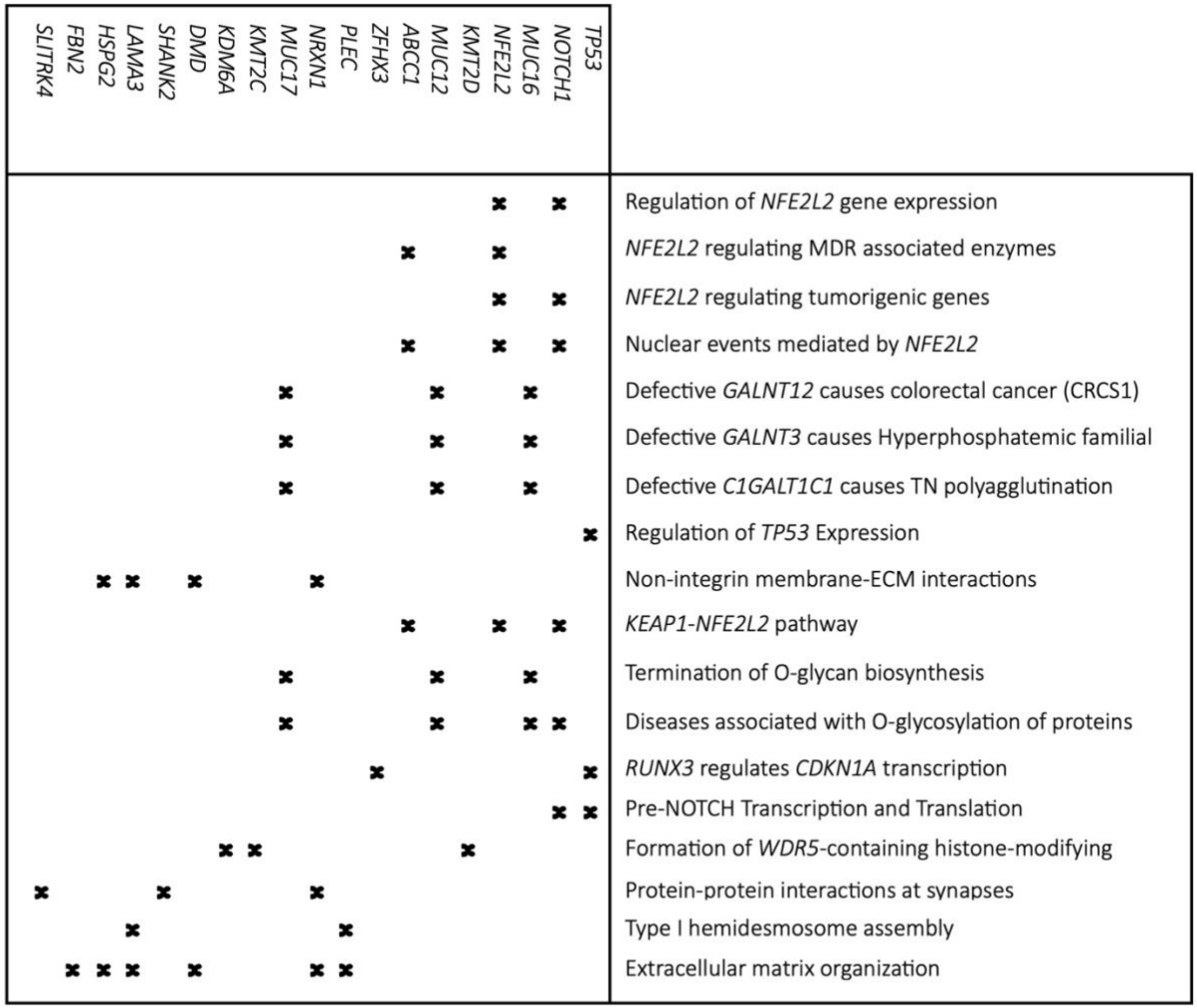


Figure 2.14 Gene involvement in significantly affected pathways in OSCC.

The altered genes involvement within various significantly affected pathways. **A**) Sub-pathways or molecular events involved in the seven pathways significantly affected ("cellular responses to stimuli", "gene expression (transcription)", "disease pathway", "metabolism of proteins", "extracellular matrix organization", "signal transduction" and "immune system") **B**) Genes involved in the affected sub-pathways, with ✘ indicating the corresponding gene (X-axis) and the corresponding sub-pathway (Y-axis) involved. Gene names and their corresponding uniport IDs and: protein names, matched to Reactome pathways are listed as follows: *TP53* (P04637: Tumour suppressor p53), *NOTCH1* (P46531: Neurogenic locus notch homolog protein 1), *MUC16* (Q8WXI7: Mucin-16), *NFE2L2* (Q16236: NFE2-related factor 2), *KMT2D* (O14686: Histone-lysine N-methyltransferase 2D), *MUC12* (Q9UKN1:Mucin-12), *ABCC1* (P33527: Multidrug resistance-associated protein 1), *ZFHX3* (Q15911: Zinc finger homeobox protein 3), *PLEC* (Q15149: Plectin), *NRXN1* (Q9ULB1: Neurexin I-alpha), *MUC17* (Q685J3: Mucin-17), *KMT2C* (Q8NEZ4: Histone-lysine N-methyltransferase 2C), *KDM6A* (O15550:Lysine-specific demethylase 6A), *DMD* (P11532: Dystrophin), *SHANK2* (Q9UPX8: SH3 and multiple ankyrin repeat domains protein 2), *LAMA3* (Q16787: Laminin subunit alpha-3), *HSPG2* (P98160: Basement membrane-specific heparan sulphate proteoglycan core protein), *FBN2* (P35556: Fibrillin-2) and *SLITRK4* (Q8IW52: SLIT and NTRK-like protein 4). Statistical significance is indicated by p-value < 0.05 and False Discovery Rate (FDR) <0.05.

Involvement of the *KMT2D*, *KMT2C* and *KDM6A* were enriched in the "gene expression (transcription)" pathway, particularly through the formation of *WDR5*-containing histone-modifying complexes (Figure 2.14). Histone modification enzymes, such as these, play crucial roles in cancer development and progression by modifying chromatin structure and regulating gene expression [211]. Both *KMT2D* and *KMT2C* facilitate transcriptional activation of target genes by modifying Histone H3 Lysine 4 Trimethylation (H3K4me3) [312].

Table 2.7 Significantly enriched pathways ranked according to their p-value and FDR.

Pathway name	p-value	FDR*
Regulation of *NFE2L2* gene expression	1.63e-05	6.6e-03
NFE2L2 regulating MDR associated enzymes	3.81e-05	7.73e-03
NFE2L2 regulating tumorigenic genes	1.06e-04	1.07e-02
Nuclear events mediated by *NFE2L2*	1.12e-04	1.07e-02
Defective *GALNT12* causes colorectal cancer (CRCS1)	1.71e-04	1.07e-02
Defective *GALNT3* causes Hyperphosphatemic *familial tumoral calcinosis* (HFTC)	1.71e-04	1.07e-02
Defective *C1GALT1C1* causes *TN polyagglutination syndrome* (TNPS)	1.97e-04	1.07e-02
Regulation of *TP53* Expression	2.14e-04	1.07e-02
Non-integrin membrane-ECM interactions	3.16e-04	1.42e-02
KEAP1-NFE2L2 pathway	3.83e-04	1.53e-02
Termination of O-glycan biosynthesis	4.56e-04	1.64e-02
Diseases associated with O-glycosylation of proteins	7.9e-04	2.61e-02
RUNX3 regulates *CDKN1A* transcription	8.44e-04	2.62e-02
Pre-NOTCH Transcription and Translation	1.23e-04	3.57e-02
Formation of *WDR5*-containing histone-modifying complexes	1.46e-03	3.63e-02
Protein-protein interactions at synapses	1.51e-03	3.63e-02
Type I hemidesmosome assembly	1.58e-03	3.63e-02
Extracellular matrix organization	1.71e-03	3.76e-02

MDR - multi-drug resistance, HFTC - Hyperphosphatemic Familial Tumoral Calcinosis, CRCS1 - colorectal cancer 1, TNPS - Tn polyagglutination syndrome, ECM - Extracellular matrix (p-value < 0.05, FDR <0.05). FDR - False Discovery Rate.

Our WES analysis revealed 35 frequently mutated genes, among these, *TP53*, *NFE2L2*, *CDKN2A*.p16INK4a, *ZNF750* and *NOTCH1* were identified as cancer driver genes. Based on the mutation spectra analysis, the samples clustered into three distinct groups: cluster 1, cluster 2a and cluster 2b, expanding upon the two clusters (cluster 1 and cluster 2b) identified in our WGS analysis. Consistent with WGS findings, these clusters were differentiated by the frequency of *TP53* alterations and the mutations per Mb sequenced in samples. In both WGS and WES analyses, cluster 1 tumours had *TP53* mutations and a relatively high somatic mutation rate per Mb, comprising the majority of samples in our cohort compared to other clusters. Cluster 2, observed in both WES and WGS analyses, showed no *TP53* mutations and was more prevalent among black female patients. However, in the WES analysis, cluster 2 was further divided into subclusters 2a and 2b. Cluster 2a displayed a high mutation rate per Mb, while cluster 2b displayed fewer genomic alterations. Across these clusters, the predominant

mutation types included C:G> T:A transitions and C:G>A:T transversions, similar to the that observed in our WGS analysis.

Mutation signature analysis identified seven mutation signatures: SBS1, SBS2, SBS5, SBS10b, SBS13, SBS15 and SBS55 across the 67 samples. Notably, SBS1, SBS5, SBS2 and SBS13 made significant contributions to the mutation burden in our cohort, consistent with our WGS findings. This suggests diverse mechanisms contributing to mutagenesis, with aging and APOBEC activation being crucial in OSCC tumour development. Furthermore, combinations of these mutation signatures were consistently observed across all samples, with 45% of the samples showing mutations from both age-related signatures, SBS1 and SBS5. Additionally, 28% of the samples displayed a mix of age-related signatures and APOBEC-associated mutation signatures. Some samples displayed three mutation signatures, while others showed five mutation signatures. This highlights the complex interplay of various mutation processes within the samples and the diverse mechanisms contributing to mutagenesis in our cohort. Interestingly, although unknown signatures were found in our WGS analysis, none were detected in the WES analysis. These unknown signatures, identified solely in the WGS analysis, likely arise from mutations in non-coding regions.

Pathway enrichment analysis revealed eight biological pathways associated with our altered genes, including “cellular responses to stimuli”, “disease pathway”, “gene expression (transcription)”, “extracellular matrix organization”, “metabolism of proteins”, “signal transduction”, “neuronal system” and “cell-cell communication”. Many of these pathways were previously reported to be involved in cancer-associated pathways including “cell cycle”, “histone modification pathway”, “NOTCH signalling pathway”, KEAP1-NFE2L2 pathway [15, 71, 115-126, 128-135]. Consistent with our WGS findings, “cellular responses to stimuli” pathway was the most disrupted pathway in our cohort, with various molecular events and sub-pathways, including the regulation of *NFE2L2* gene expression and the KEAP1-NFE2L2 pathway. These findings suggest the role of *NFE2L2* and its regulatory network in mediating “cellular responses to stimuli”, implicating its involvement in the molecular mechanisms driving tumorigenesis. Furthermore, while there was overlap in the disrupted pathways significantly disrupted between WGS and WES analysis, two additional pathways, “neuronal system” and “cell-cell communication”, were exclusively found in the WES cohort. These pathways contribute to the interactions between tumour cells and their surrounding microenvironment, playing a crucial roles in tumour development and progression [308, 309]

2.2.4 Validation of bioinformatics analysis by re-sequencing

Our sequencing data identified variants of significant impact that could drive the development of oesophageal squamous cell carcinoma, as outlined earlier in this chapter (section 2.2.2.1 and section 2.2.3.1). Given the significance of our results, we sought to validate variants identified through WGS and WES to ensure the accuracy and sensitivity of our bioinformatics data. To validate the somatic mutations obtained through sequencing, we performed experimental validation through PCR-based Sanger sequencing on a selected subset of genes from the whole genome sequencing data. We examined six genes mutated in at least two samples with non-silent mutations. Among these genes, three are well-established tumour suppressors/oncogenes frequently mutated in OSCC (*CDKN2A*, *NFE2L2* and *PIK3CA*), while the remaining three (*TOPBP1*, *ERCC6* and *C20orf196*) are less characterized in OSCC.

Due to limitations in samples availability, the DNA content of some of the biopsy samples were insufficient for additional PCR and sequencing analysis. Therefore, only the samples listed in Table 2.8 (below) were used for further analysis for the selected mutations as indicated.

Table 2.8 List of somatic mutations validated.

Gene	Mutation	Exon	Case(s)
CDKN2A	c.172C>T [p.R58*]	2	PD39455, PD39451
	c.322G>A [p.D108N]	2	PD39453
	c.329G>A [p.W110*]	2	PD39459
	c.244_245insGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCG [p.V82fs*51]	2	PD39449
PIK3CA	c.1633G>A [p.E545K]	9	PD39449, PD39457
NFE2L2	c.101G>A [p.R34Q]	2	PD39449
	c.229G>C [p.D77H]	2	PD50653
	c.241G>A [p.G81S]	2	PD39455
TOPBP1	c.3792A>T [p.L1264F]	22	PD39451
	c.1429T>A [p.F477I]	9	PD39449
ERCC6	c.3254C>G [p.S1085C]	17	PD39448
	c.388G>T [p.E130*]	1	PD39460
C20orf196	c.161C>T [p.S54F]	1	PD39457
	c.211T>A [p.S71T]	2	PD39459

All *CDKN2A* mutations were confirmed in all samples tested. These mutations included *CDKN2A* c.172C>T [p.R58*], c.322G>T [p.D108Y] and c.329G>A [p.W110*] as well as the frameshift insertion of 40 bp c.244_245insGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCG [p.V82fs*51] (Figures 2.15 and 2.16A). Of note, the 40 bp frameshift insertion was heterozygous in patient PD39449a, as indicated by the presence of bands for both mutant and wild-type sequences (Figures 2.15, Figure 2.16B).

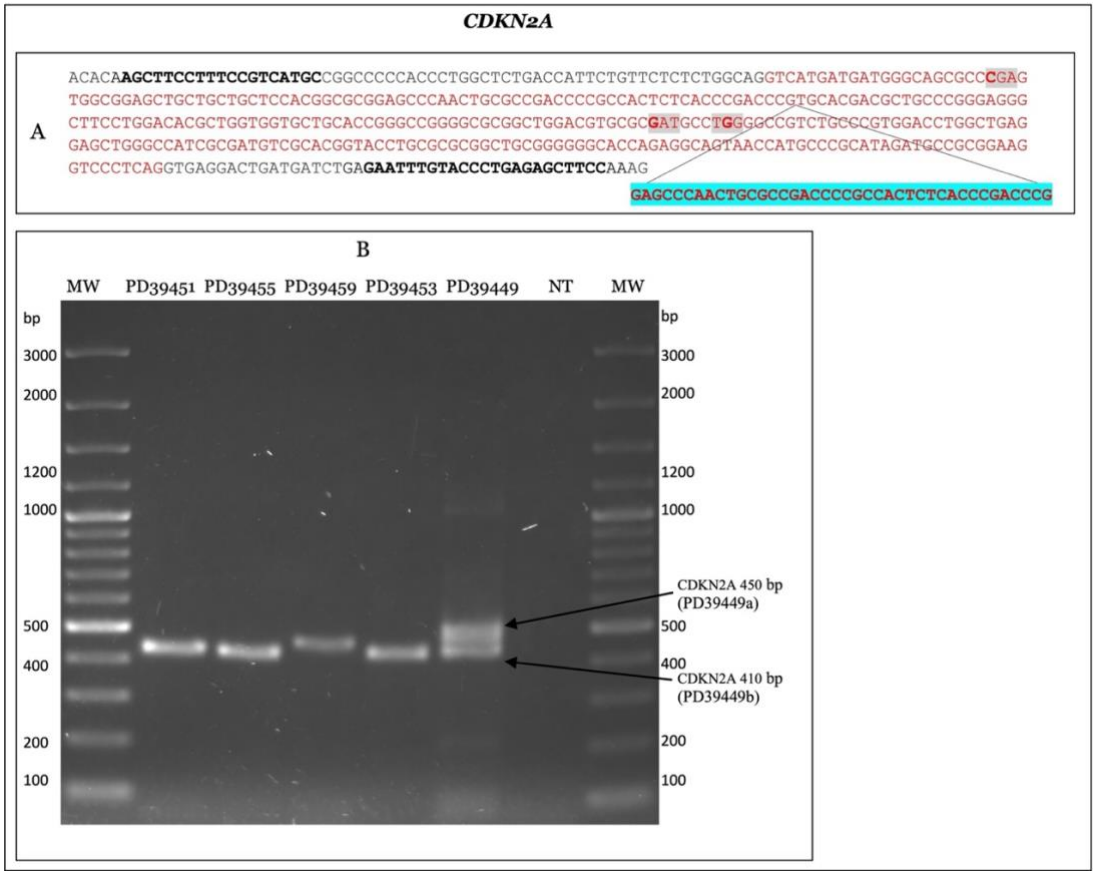
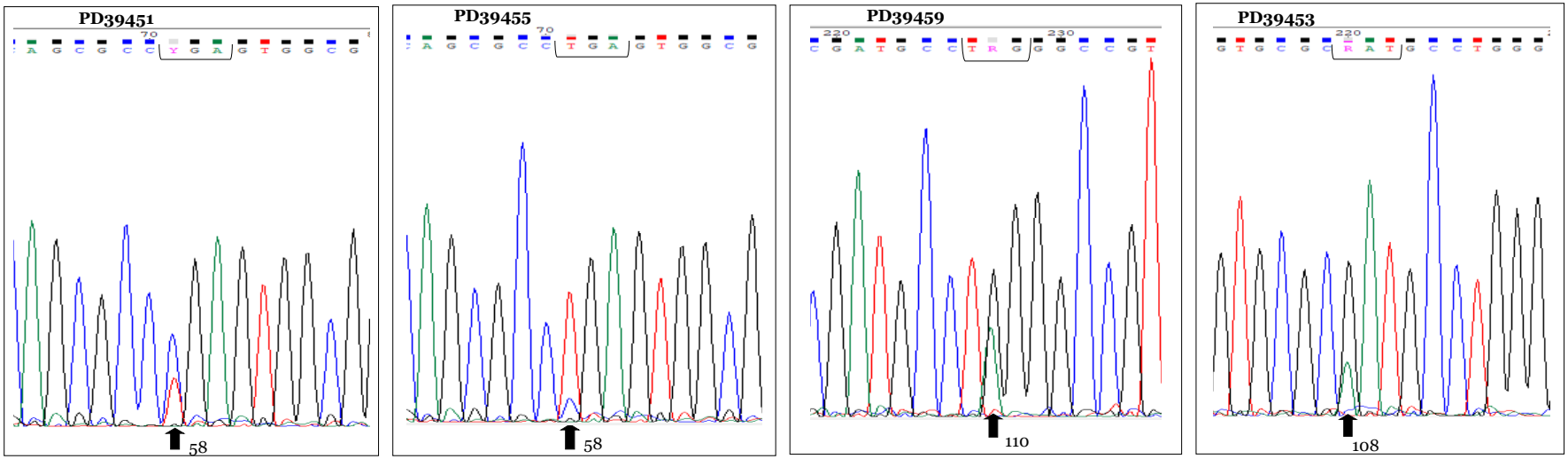


Figure 2.15 Primer design and PCR amplification for *CDKN2A* mutations validation.

A) Sequence of the PCR amplified region encompassing *CDKN2A* exon 2 (red) flanked by *CDKN2A* intron 1 and 2 (black). Two amplicons were observed for *CDKN2A*: a wild-type amplicon of 410 bp and a mutant 450 bp amplicon. The 40 nucleotides insertion is indicated in a blue highlighter. The three codons mutated by single base substitutions are highlighted in grey. The specific primer pairs used for *CDKN2A* mutational analysis are indicated in Materials and Methods in section 6.2.2.7, Table 6.4. The positions of the two primers used for PCR amplification are shown in bold. **B**) Agarose gel analysis of PCR products of DNA from tumour samples with mutations. DNA was extracted from the tumour samples as described in Material and Methods in section 6.2.1.3 and 50ng genomic DNA was used for PCR reaction mixture with 5% DMSO as listed in Table 6.6. PCR conditions are listed in Material and Methods in section 6.2.4, Table 6.7. PCR products were separated on a 1% agarose gel for 120 minutes. Sample PD39451, PD39455, PD39459 and PD394593 PCR products as well as agarose gel slices PD39449a and PD39449b were subjected to DNA sequencing to validate the presence of *CDKN2A* mutations.

CDKN2A

A


PD39451
58
PD39455
58
PD39459
110
PD39453
108

B

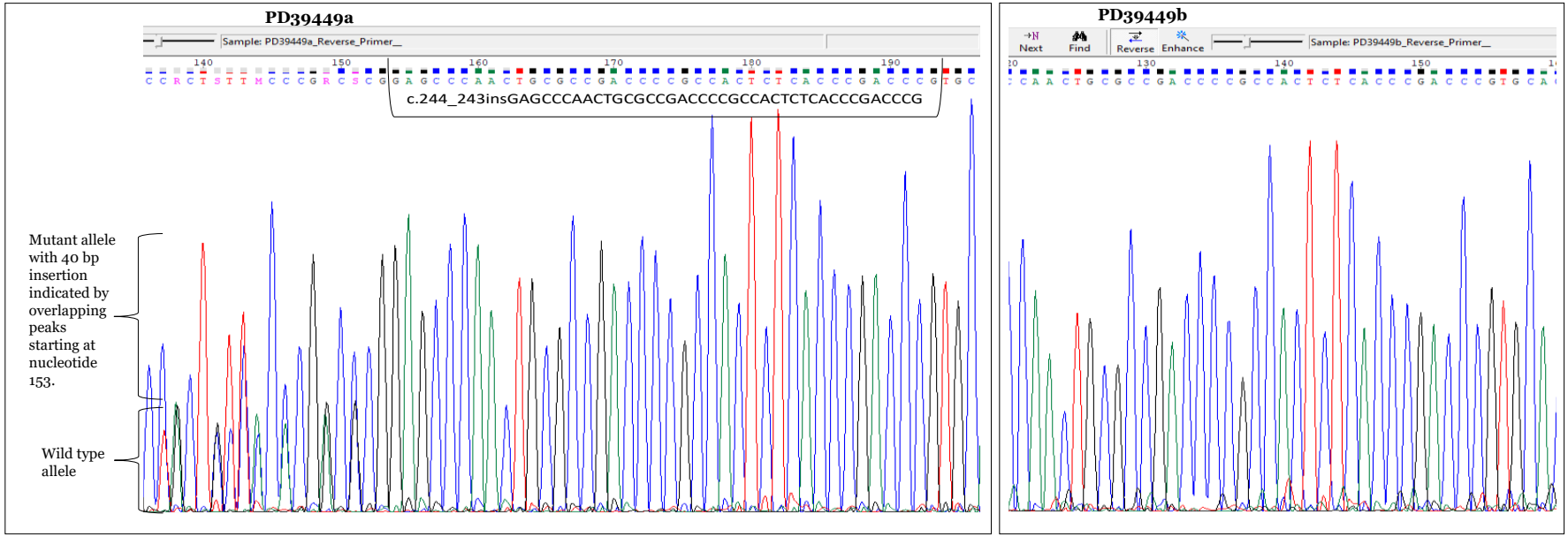

PD39449a
Sample: PD39449a_Reverse_Primer_
c.244_243insGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCG
Mutant allele with 40 bp insertion indicated by overlapping peaks starting at nucleotide 153.
Wild type allele
PD39449b
Next
Find
Reverse
Enhance
Sample: PD39449b_Reverse_Primer_

Figure 2.16 Sanger sequencing of *CDKN2A* PCR products.

A) The position of nucleotide substitution is indicated by a black arrow. The codons mutated by single base substitutions are indicated with a bracket. In sample PD39451 and PD39455 at codon 58, (CGA > TGA), the mutation changed an arginine into a stop codon [p.R58*]. In sample PD39459 at codon 110 (TGG > TAG), the mutation changed a tryptophan into a stop codon [p.W110*]. In sample PD39453 at codon 108 (GAT > AAT), the mutation changed aspartic acid into asparagine [p.D108N]. In sample PD39449, the insertion of 40 nucleotides in *CDKN2A* was heterozygous. This was also confirmed by the sequences of the two gel bands. **B**) In sample PD39449, the frameshift insertion of 40 nucleotides in *CDKN2A* was heterozygous. PD39449a gel band (mutant) on the left showed overlap peaks of both mutant and wild-type sequences starting at nucleotide 153 while PD39449b gel band (wild-type) on the right had wild-type peaks only.

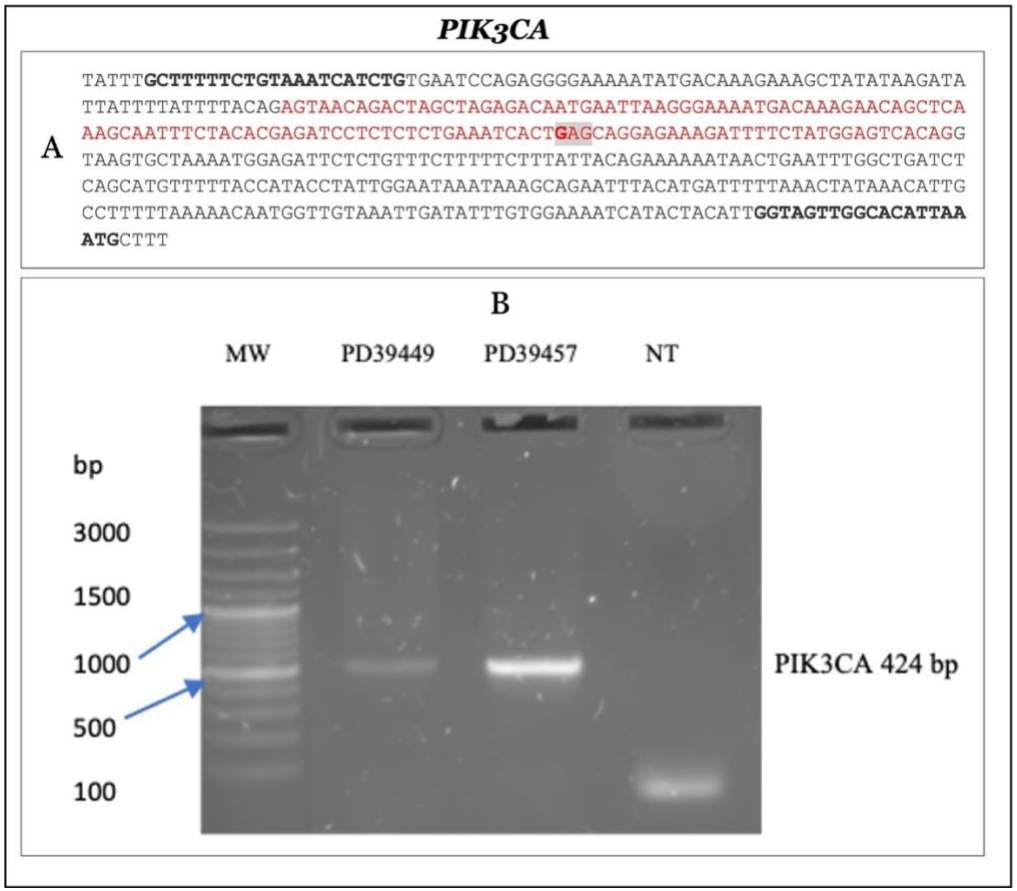


Figure 2.17 Primer design and PCR amplification for *PIK3CA* mutation validation.

A) Sequence of the PCR amplified region encompassing *PIK3CA* exon 9 (red) flanked by *PIK3CA* intron 8 and 9 (black). The codon mutated is highlighted in grey. The specific primer pairs used for *PIK3CA* mutational analysis are indicated in Materials and Methods in section 6.2.2.7, Table 6.4. The positions of the two primers used for PCR amplification are shown bolded in black. Amplicon size was 424bp. **B**) Agarose gel analysis of PCR products of DNA from tumour samples with mutations. DNA was extracted from the tumour samples as described in Material and Methods in section 6.2.1.3 and 50ng genomic DNA was used for PCR reaction mixture as listed in Table 6.5. PCR conditions are listed in Material and Methods in section 6.2.4, Table 6.7. PCR products were subjected to DNA sequencing to validate the presence of *PIK3CA* mutations.

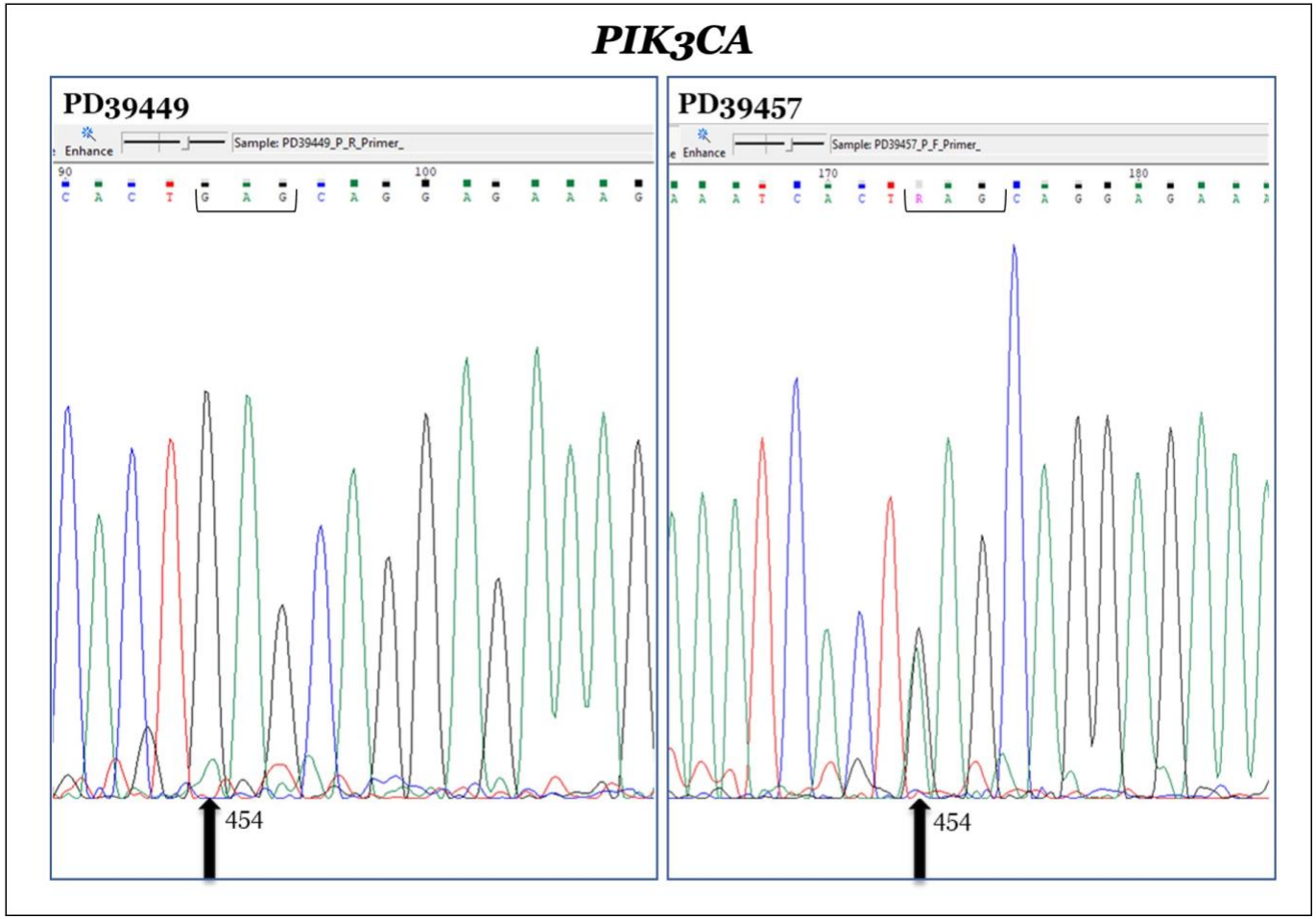


Figure 2.18 Sanger sequencing of *PIK3CA* PCR products.

Sequencing chromatograms showing the mutations of *PIK3CA* exon 9 as identified in our cohort. The position of nucleotide substitution is indicated by a black arrow. The codon mutated by single base substitution is indicated with a bracket. In samples PD39449 and PD39457 at codon 454, (GAG > AAG) the mutation changed glutamic acid into lysine [E454K].

PIK3CA c.1633G>A [p.E545K] missense mutation was validated in the one of the two samples (Figures 2.17 and Figure 2.18). In the case PD39449, Sanger sequencing failed to clearly confirm the presence of the mutation. The chromatogram displayed a very low mutation peak; specifically, the A (mutant) peak was considerably smaller than the G (wild-type) peak, suggesting a possible false negative result. (Figure 2.18).

All three *NFE2L2* mutations (*NFE2L2* c.101G>A [p.R34Q], c.229G>C [p.D77H], c.241G>A [p.G81S]) were preset in exon 2 (Table 2.8) and were validated in all examined samples (Figures 2.19, Figure 2.20 and Figure 2.21). Mutations in *C20orf196* were present in exon 1 (*C20orf196* c.161C>T [p.S54F]) for sample PD39457 and in exon 2 (*C20orf196* c.211T>A [p.S71T]) for sample PD39459 (Table 2.8). Although all *C20orf196* mutations were validated, the peak heights of mutant alleles were smaller than the wild-type alleles in both samples (Figures 2.19, Figure 2.20 and Figure 2.21).

Two different mutations were examined for *TOPBP1* and *ERCC6* (Table 2.8); however, only one mutation was validated for each gene: *TOPBP1* c.3792A>T [p.L1264F] and *ERCC6* c.3254C>G [p.S1085C] (Figures 2.19, Figure 2.20 and Figure 2.21).

The majority of mutations (12/14, 86%) examined have been validated, thus demonstrating the specificity and precision of bioinformatics data. Additionally, our analysis of mutation validation revealed the presence of overlapping peaks, where both wild-type and mutant bases are present at mutated sites across most of the genes analysed (Figures 2.16, Figure 2.18, Figure 2.21). This phenomenon could be due to the following: the heterozygous nature of mutations within genes (where one allele exhibits a mutation while the other retains the wild-type sequence), the presence of polyclonal tumour cells in our samples, or contamination from normal tissue in the tumour samples.

In sample PD39449, where the tumour percentage was 54% (Table 2.9), despite the observation that more than half of the cells in this were tumour cells, the mutant allele's peak height (green) was less than 1/4 of the wild-type allele's peak height (black) (Figure 2.18). This observation suggests that only one of the alleles in the tumour tissue is mutated. Furthermore, it is possible that in some samples with low tumour tissue percentages, mutations may fall below the limit of detection, a known pitfall of tissue-based next generation sequencing assays [313]. This

limitation potentially explains the false negative result for the *ERCC6* c.3254C>G [p.S1085C] mutation in sample PD39460 (tumour percentage below 20%) (Figure 2.21, Table 2.9).

In summary, our analysis of WGS and WES revealed frequently mutated genes and driver genes such as *TP53*, *NFE2L2*, *CDKN2A.p16INK4a*, CDKN2A.p14ARF, *KMT2D*, *ZNF750* and *NOTCH1*. We also identified mutation signatures and affected pathways within our cohort. Based on the mutation spectra analysis, we found three distinct groups: cluster 1, cluster 2a and cluster 2b, characterized by the frequency of *TP53* alterations and the mutations per Mb sequenced in samples. Age-related signatures (SBS1 and SBS5) and APOBEC-associated (SBS2 and SBS13) were predominant in our cohort, suggesting diverse mechanisms contributing to mutagenesis, with aging and APOBEC activation being crucial in OSCC tumour development. In addition, three novel mutation signatures were detected by WGS, although the relevance of this signature for OSCC in South African population is uncertain given the low incidence observed. Sanger sequencing validated 12 of the 14 selected mutations, comprising one insertion and 13 synonymous mutations across six genes, including well-known cancer genes and new identified recurrently mutated genes in OSCC. Collectively, our findings provide very interesting insights into frequently mutated genes, driver genes, molecular subtypes in OSCC, disrupted pathways, mutation signatures in OSCC.

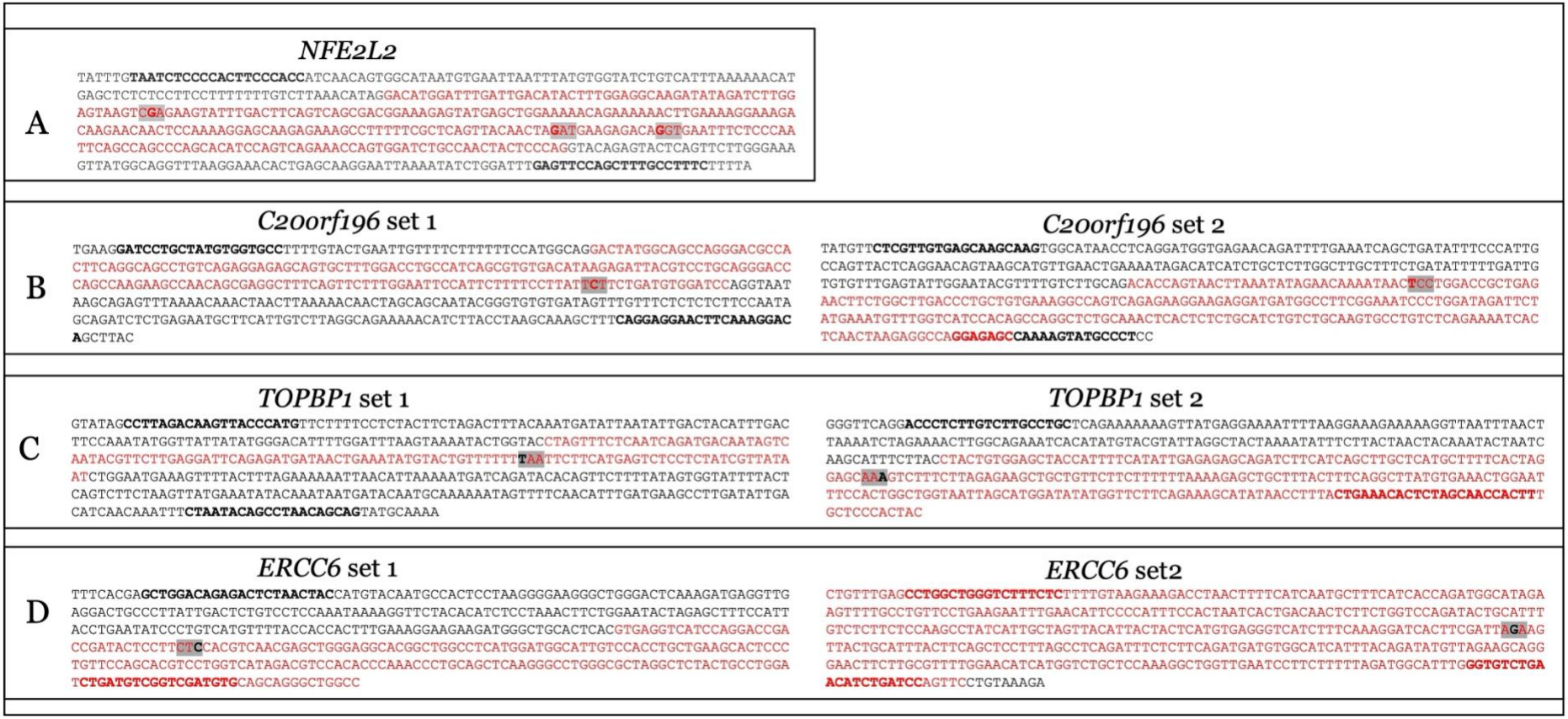


Figure 2.19 Primer design of *NFE2L2, C20orf196, TOPBP1*, and *ERCC6*.

Sequence of the PCR amplified region for *NFE2L2, C20orf196, TOPBP1*, and *ERCC6*. Exons (red) flanked introns (black). The codons are highlighted in grey. The specific primer pairs used for each gene mutational analysis are indicated in Materials and Methods in section 6.2.2.7, Table 6.4. The positions of the two primers used for PCR amplification are shown in bold. **A**) *NFE2L2*, **B**) *C20orf196*, **C**) *TOPBP1*, and **D**) *ERCC6*.

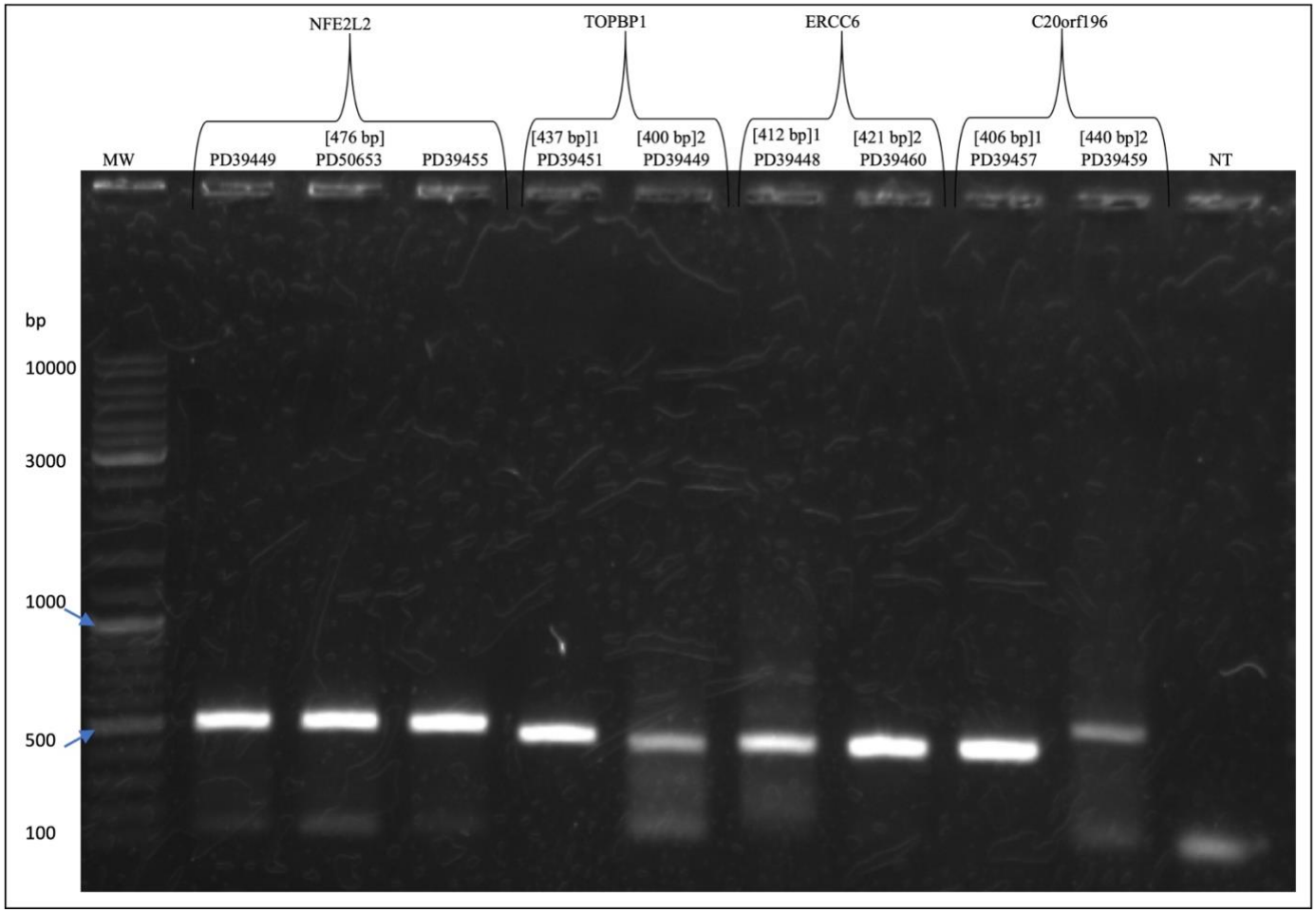


Figure 2.20 PCR amplification of *NFE2L2, C20orf196, TOPBP1*, and *ERCC6*.

Agarose gel analysis of PCR products of DNA from tumour samples with mutations. DNA was extracted from the tumour samples as described in Material and Methods in section 6.2.1.3 and 50ng genomic DNA was used for PCR reaction mixture as listed in Table 6.5. PCR conditions are listed in Material and Methods in section 6.2.4, Table 6.7. PCR products were separated on a 1% agarose gel for 55 minutes. PCR products were subjected to DNA sequencing to validate the presence of *NFE2L2, TOPBP1*, *ERCC6,* and *C20orf196* mutations.

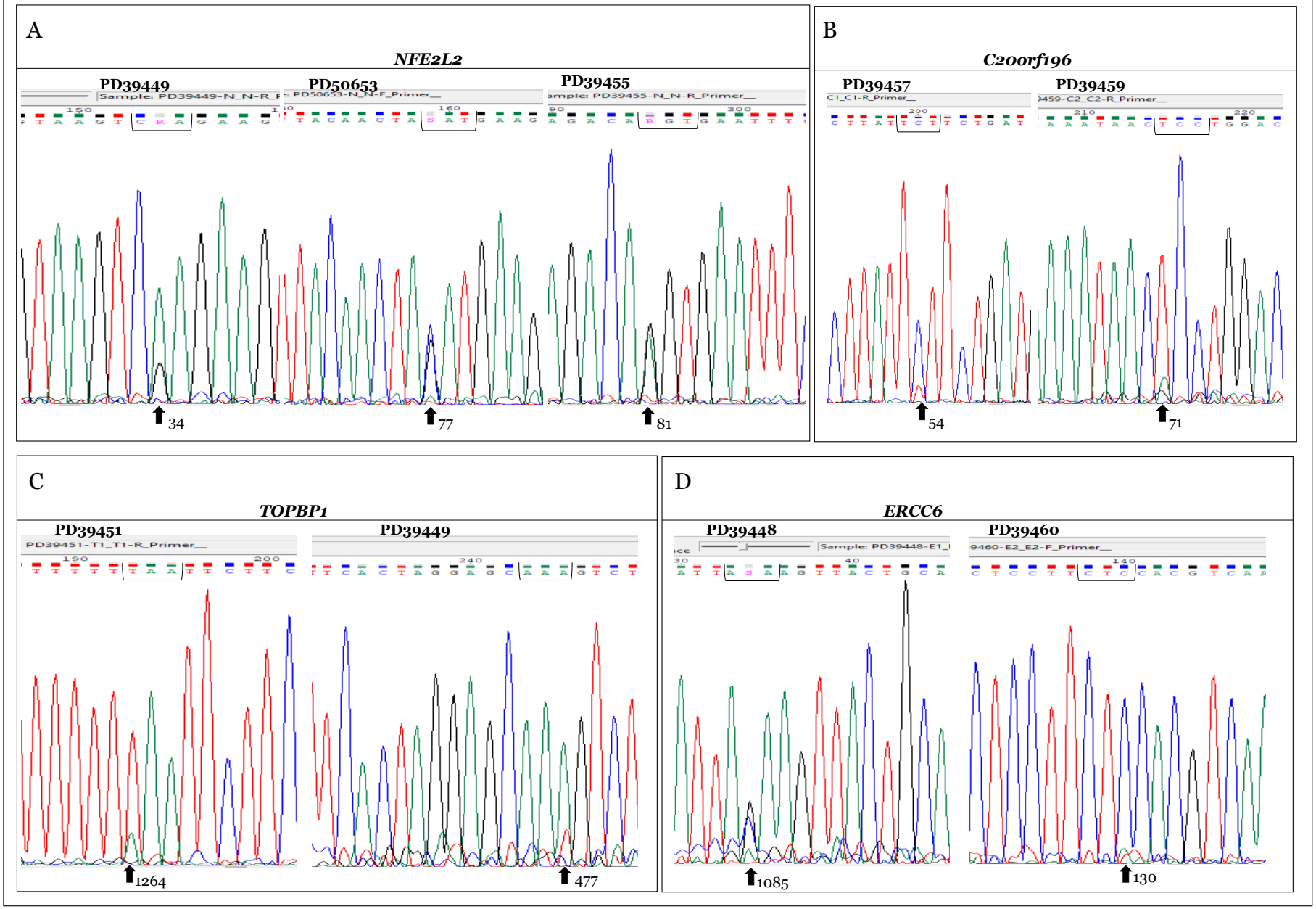

A
NFE2L2
PD39449
PD50653
PD39455
34
77
81
B
C20orf196
PD39457
PD39459
54
71
C
TOPBP1
PD39451
PD39449
1264
477
D
ERCC6
PD39448
PD39460
1085
130

Figure 2.21 Sanger sequencing of *NFE2L2*, *TOPBP1*, *ERCC6*, and *C20orf196* PCR products.

Sequencing chromatograms showing the mutations of *NFE2L2*, *TOPBP1*, *ERCC6*, and *C20orf196* mutations. The curves are colour coded to indicate the four DNA nucleotides: G (black), C (blue), A (green), and T (red). The position of nucleotide substitution is indicated by a black arrow. The codon mutated by single base substitution is indicated with a bracket. **A)** For *NFE2L2*, in sample PD39449 at codon 34, the CGA > CAA mutation changed arginine into glutamine [p.R34Q], while in sample PD50653 at codon 77, the GAT > CAT mutation changed aspartic acid into histidine [p.D77H] and in sample PD39455 at codon 81, the GGT > AGT mutation changed glycine into serine [p.G81S]). **B**) For *C20orf196* mutations in sample PD39457 at codon 54, the TCT > TTT mutation changed serine into phenylalanine [p.S54F] and in sample PD39459 at codon 71, the TCC > ACC mutation changed serine into threonine [p.S71T]). **C**) Validated *TOPBP1* mutation (TTA > TTT, -strand) in sample PD39451 at codon 1264 changed leucine into phenylalanine [p.L1264F]). **D**)Validated *ERCC6* mutation (TCT > TGT, -strand) in sample PD39448 at codon 1085 changed serine into cysteine [p.S1085C].

2.3 Discussion

In this study, we performed whole genome sequencing and whole exome sequencing analysis on 31 and 67 OSCC samples, respectively, to investigate the mutational landscape in South African oesophageal squamous cell carcinoma patients. Several novel findings have been identified in this study including driver genes, molecular subtypes of OSCC, mutation signatures and several frequently mutated genes associated with known cancer pathways.

The observed median tumour mutation burden of non-silent variants, at 2.5 mutations per Mb, was similar to the reported mutation load of 1.9–3 mutations per Mb in OSCC [123-125, 128, 135, 139]. We identified frequently mutated genes including *TP53*, *AHNAK2*, *TTN*, *AHNAK*, *NOTCH1*, *KMT2D*, *NFE2L2*, *CDKN2A*, *PCLO*, *CSMD3*, *PIK3CA*, *LRP1B*, *FAT3*, *FAT2*, *KMT2C*, *RYR2* and *MUC16* across both the WGS and WES cohorts. Most of these genes had recurrent loss-of-function mutations including missense, nonsense, and frameshift variants indels, or splicing variants, suggesting their roles tumour suppressor [139]. Furthermore, our study found a considerable overlap between frequently mutated genes in our cohort and other high-risk regions for oesophageal squamous cell carcinoma including Malawi and Asia [22, 77, 123-125, 127, 128, 131, 135, 136, 138, 139, 142, 143, 156].

We identified several driver genes in our analysis including *TP53*, *CDKN2A* (*p14ARF* and *p16INK4a*), and *KMT2D* in our WGS analysis, and *TP53*, *CDKN2A* (*p16INK4a*), *NFE2L2*, *ZNF750* and *NOTCH1* in our WES analysis. These findings suggest deregulation in genomic stability and cell cycle progression in OSCC [158], consistent with previously reported significantly mutated genes (driver genes) in OSCC cohorts [77, 124, 125, 127, 128, 131-133, 136, 142, 154-157]. Although there was some overlap in the driver genes identified by both WGS and WES, certain driver genes were unique to each analysis, *NFE2L2*, *ZNF750* and *NOTCH1* in the WES cohort, and *p14ARF* and *KMT2D* in the WGS cohort. The difference in detected mutations can partly be attributed to the difference in sample size, with WGS involving on 31 samples, and WES involving 67 samples. A larger sample size in WES increases statistical power and enhances the cohort's diversity, potentially detecting mutations that are missed by WGS.

Most *TP53* mutations occurred predominantly in the DNA-binding domain, particularly at *TP53* mutation hotspots such as R273 and R282W, similar to previous reports [311]. *KMT2D*

mutations clustered throughout the protein, mainly consisting of protein truncating variants, including in-frame indels and nonsense variants. *NFE2L2* mutations clustered in *NFE2L2* mutation hotspots, including mutations within *KEAP1* binding motifs (ETGE and DLG) of *NFE2L2* (p.W24, p.V32, p.R34, p.D77, p.E79 and p.E82), which had been previously reported in OSCC [136, 156, 219]. Notably, mutations such as p.R34Q and p.E79K hinder the tumour suppressive activity of NFE2L2 protein and exert an oncogenic role in OSCC [136].

Another key finding was the identification of *CDKN2A*-encoded proteins (*p14ARF* and *p16INK4a*) as driver genes in our cohort. However, *p14ARF* mutations were identified as a driver in our WGS analysis but not in our WES analysis, possibly due to the fewer mutations in *p14ARF* within this cohort. Furthermore, two of the mutations in *CDKN2A* (*p16INK4a* [p.Q50* and p.L16fs*6]) occurred in exon 1α encoding for *p16INK4a*, while one mutation in exon 2 of (*p16INK4a* [p.L130P]) occurred outside the *p14ARF* coding region. Previous studies frequently report driver mutations of *p16INK4a* in OSCC rather than *p14ARF* [156]. Additionally, mutations in exon 2 of *p14ARF* and *p16INK4a* almost exclusively inactivate p16INK4a protein only [314]. In our WGS analysis, most of the *CDKN2A* mutations were located in exon 2 of *p14ARF* and *p16INK4a*. The majority of these mutations resulted in a premature truncation of the p*16INK4a* protein. Several of these mutations occurred at *p16INK4a* mutational hotspot (p.R58*, p.W110* and p.D108X), which were reported in two or more samples, consistent with frequencies previously observed in other OSCC populations [124, 128, 174]. These alterations are predicted to confer a loss of function of the protein, disrupting its tumour suppressive role and predisposing individuals to cancer [177]. In comparison, the majority of *p14ARF* exon 2 mutations resulted in mainly missense mutations. Point mutations in exon 1β (encoding for *p14ARF*) are infrequent [175], with *p16INK4a* genomic alterations being more common than those of *p14ARF* in OSCC [176]. Similarly, in our WES analysis, we observed two mutations in exon 1α encoding for *p16INK4a* (*p16INK4a* [p.Q50* and p.L16fs*6]), and none in exon 1β of *p14ARF*. Moreover, one mutation in exon 2 of *p16INK4a* [p.L130P] occurred outside the *p14ARF* coding region, potentially explaining why *p14ARF* mutations were not significant drivers in our WES analysis. Mutations in these driver genes could confer growth advantages to cancer cells and promote the development and progression of oesophageal squamous cell carcinoma [315].

Various studies revealed molecular evidence suggesting the presence of sub-structuring of OSCC (reviewed in [284]). These studies have identified distinct molecular subtypes within

their cohorts [22, 128, 132, 137, 297]. In our study, both of our WGS and WES cohorts revealed distinct clusters based on the mutation profile of our samples. Our WGS analysis identified two distinct clusters labelled (in our study) cluster 1 and cluster 2b, primarily distinguished by the presence of *TP53* alterations and the frequency of mutations per Mb sequenced. Meanwhile, within our WES cohort, the samples clustered into three distinct groups labelled cluster 1, cluster 2a and cluster 2b, expanding upon the two clusters identified in our WGS analysis. In both WGS and WES analysis, cluster 1 tumours were characterized by *TP53* mutations and a relatively high somatic mutation rate per Mb, representing the majority of our samples in both WGS and WES cohorts.

Cluster 2, on the other hand, showed no *TP53* mutations across all samples and was predominantly composed of black female patients in both WGS and WES cohorts. In addition, cluster 2 was subdivided into two subclusters in WES analysis: cluster 2a showed a high mutation rate per Mb while cluster 2b samples exhibited fewer genomic alterations, with the fewest somatic mutations per Mb. This pattern of high or low mutation rates and presence or absence of *TP53* mutations is consistent with observations in Malawian OSCC patients and African American OSCC populations [22, 137, 297].

A study of African American OSCC patients reported a complex mutation profile, dividing their samples into high and low mutation groups, reflecting similar patterns [137]. Another study identified three molecular subtypes within their cohort labelled subtype OSCC1, OSCC2 and OSCC3 [132]. OSCC1 subtype displayed alterations in the NFE2L2/KEAP1 pathway, OSCC2 showed higher rates of mutation of *NOTCH1* or *ZNF750*, while subtype 3 (OSCC3) is present among African Americans patients [132], exhibited a lower likelihood of *TP53* mutations. Similarly, a whole exome sequencing analysis of Malawian OSCC samples revealed three subtypes based on RNA expression analysis: subtype 1a, 1b and 2 [22]. Their subtype 1b as well as subtype 3 (OSCC3) shared characteristics with our cluster 2b, featuring fewer genomic alterations, the fewest somatic mutations per Mb and lower prevalence to no mutations in *TP53*.

Cluster 1 in our study appeared to align with previously reported molecular subtypes found in Caucasian and Asian patients, subtype 2 (OSCC2) and subtype 1 (OSCC1), respectively [132].

The identification of molecular subtypes, such as cluster 2b characterized by fewer genomic alterations and allow prevenance of *TP53* mutations, suggests specific biological pathways or mechanisms involved in the development and progression of OSCC, especially among black females, who predominantly exhibited this subtype. Furthermore, these observations suggest potential ethnic or geographic influences on OSCC molecular profiles. These findings underscore the diverse nature of OSCC and highlight the presence of distinct molecular profiles that may significantly influence clinical outcomes and guide therapeutic strategies.

The transition and transversion rate analysis in coding sequences of our samples showed that C:G> T:A transitions were the most common mutations, followed by C:G>A:T and C:G>G:C transversions, similar to previous studies of OSCC [131, 138]. The C:G> T:A transitions predominance is consistent with spontaneous cytosine deamination to thymine [227] being a major mutagenic process in oesophageal squamous cell carcinoma, as previously observed in OSCC [77, 122, 123, 125, 128, 131, 137, 138, 142, 154].

Our mutation signatures analyses revealed distinct signature profiles in OSCC and their correlation with patients' demographic variables, behavioural or lifestyle factors as well as their mutational profiles. Our analysis identified eight mutation signatures in WGS including SBS1, SBS2, SBS13, SBS5, and SBS6, and three novel, unknown signatures labelled as unknown A, unknown B, and unknown C. Similarly, in our WES analysis, seven mutation signatures were identified including SBS1, SBS2, SBS5, SBS10b, SBS13, SBS15 and SBS55 across the 67 samples. Importantly, SBS1, SBS5, SBS2 and SBS13 significantly contributed to the mutation burden in our cohort, consistent with previous OSCC studies [22, 77, 123-125, 128, 131, 135, 136, 138, 139, 142, 238-240]. The enrichment of mutations from both age-related signatures, SBS1 and SBS5, in a large portion of the samples highlights the influence of aging on the mutational profile in our cohort. Consistently, a significant association was shown between signatures SBS1 and SB5 and the age of OSCC patients, as previously reported in OSCC [156]. SBS 2 and SBS13 were also detected in several samples, in both WGS and WES analyses. These signatures are associated with the hyperactivity of the APOBEC family enzyme, cytidine deaminase [146], resulting in deamination of cytidine to uracil within RNA and DNA [244]. Additionally, SBS10b suggests polymerase epsilon exonuclease domain mutations in our samples, while SBS6 and SBS15 suggests possible defects in mismatch repair and microsatellite instability [146]. Interestingly, SBS10b contributes to very high numbers of mutations in samples exhibiting this signature, consistently, four samples with SBS10b had a

higher mutation burden. On the other hand, the identification of three unknown signatures through WGS suggests that there may be unique mutational processes that contributes to oesophageal carcinogenesis in South Africa. These unknown signatures, identified solely in the WGS analysis, likely arise from mutations in non-coding, although it's unclear how this may arise.

In most cancer types, at least two mutation signatures were observed, with a maximum of six signatures in cancers of the liver, uterus and stomach [146]. Similarly, across our samples, our results revealed samples exhibiting combinations of two or more mutation signatures. In WES analysis, 45% of the samples both age-related signatures, while 28% displayed a mixture of age-related signatures and APOBEC-associated mutation signatures. Some samples displayed 3-5 mutation signatures. These results highlights the complex interplay of various mutation processes within the samples and the multifaceted nature of mutagenesis in our cohort.

Interestingly, smoking-associated mutation signatures such as SBS4 and SBS29 were not detected in OSCC, aligning with previous studies [22, 77, 128]. These signatures are characterized by C>A mutations [146, 241]. Further analysis found no difference in the overall mutation load between the smokers and non-smokers, consistent with findings from other OSCC cohorts [77, 123-125, 127, 128, 135, 155]. Similarly, we did not observe the mutation signature typically associated with alcohol consumption (SBS16), further, we found no difference in T>C substitutions between alcohol consumers and non-consumers. These findings highlight the complexity of OSCC mutagenesis and the potential influence of factors beyond smoking and alcohol consumption.

Pathway enrichment analysis revealed six biological pathways associated with our altered genes, including “cellular responses to stimuli”, “disease pathway”, “gene expression (transcription)”, “extracellular matrix organization” and “metabolism of proteins” across our two cohorts. Several of these pathways were previously implicated in cancer-associated pathways including “cell cycle”, PI3K/AKT pathway “NOTCH signalling pathway”, Histone modification, *KEAP1-NFE2L2* pathway [77, 122-134, 136-143]. Among these, “cellular responses to stimuli” was the most disrupted pathway in our cohort, exhibiting various molecular events and sub-pathways, notably involving regulation of *NFE2L2* gene expression and the *KEAP1-NFE2L2* pathway. *NFE2L2* is a transcriptional factor that regulates and activates expression of antioxidant response genes involved in the pathway [214]. Under

normal conditions, *NFE2L2* binds to *KEAP1* via its DLG and ETGE domains, leading to its proteasomal degradation and low cellular levels [215, 216]. However, upon exposure to stresses, inactivation of KEAP1 stabilizes NFE2L2, resulting in elevated expression of its target genes, thereby enhancing stress resistance and cell proliferation [214, 217].

Other genes that were involved in regulating cellular stress responses and cellular senescence include *TP53* and *p14ARF* and *p16INK4a*. Previous studies have demonstrated that the expression of *p14ARF* and *p16INK4a* typically increases in highly proliferative cells affected by oncogenic stimuli, provided the genes are intact [316, 317]. Our pathway analysis revealed these genes as participants in the "cellular responses to stimuli" pathway, mainly in response to oncogene-induced senescence. In addition, *p14ARF* and *p16INK4a* were implicated in other pathways significantly disrupted in our cohort, such as the "gene expression (transcription)" pathway, with molecular events mainly involved in the regulation of *TP53* expression and the transcription of *TP53* target genes such as *CDKN1A* (p21). These findings suggest that *p14ARF* and *p16INK4a* regulate the oncogene-induced stress and senescence by regulating gene expression and activating *TP53* downstream genes like *CDKN1A* (p21), crucial for maintaining genomic stability, regulating cell cycle progression and promoting senescence and possibly preventing tumourigenesis [318].

Components of the "NOTCH signalling pathway" have been reported to interact with p53 [319]. In our cohort, the NOTCH pathway was mainly enriched in "signal transduction", involving the regulation of pre-NOTCH transcription and translation, expression, and processing. Altered genes that were found involved in the NOTCH pathway included *NOTCH1* and *TP53*, with *NOTCH1* frequently disrupted by loss-of-function mutations. This supports the role of NOTCH pathway activity in the growth of tumour cells with squamous differentiation characteristics [207].

While there was overlap in the significantly disrupted pathways identified in both WGS and WES analyses, the WES cohort revealed two additional pathways, "neuronal system" and "cell-cell communication". These pathways contribute to the interactions between tumour cells and their surrounding microenvironment, playing a crucial roles in tumour development and progression [308, 309].

In summary, this study characterized the genomic landscape of the South African oesophageal squamous cell carcinoma, and we identified frequently mutated genes and driver genes, disrupted signalling pathways, mutation signatures, and OSCC subtypes. Within our cohort, we identified three distinct groups based on mutation spectra labelled as cluster 1, cluster 2a and cluster 2b, characterized by varying frequency of *TP53* alterations and mutations per Mb sequenced. We have reported driver genes crucial to OSCC occurrence and development in South African population, including inactivating mutations in *TP53, CDKN2A* products; *p14ARF* and *p16INK4a*, *KMT2D*, *NFE2L2*, *NOTCH1* and *ZNF750*, along with genomic aberrations targeting the *NFE2L2-KEAP* pathway, cell cycle and senescence and NOTCH signalling pathways. Furthermore, we identified the predominant mutation processes linked to OSCC such as clock-like signature and the APOBEC-mediated mutation signatures. Additionally, our cohort revealed three unknown mutation signatures. Smoking-associated mutation signatures did not strongly correlate with OSCC in the South African population, suggesting that mutation processes in our cohort may not be primarily driven by smoking or tobacco use, despite epidemiological evidence pointing to such a link in patients with OSCC in South Africa [6].

Chapter 3
Expression profiles and pathway analysis of selected differentially expressed genes in OSCC

3.1. Introduction

Genetic modifications and genomic instability are two of the mechanisms that can result in abnormal proteins and altered expression, thereby potentially contributing to development and progression of tumours [320]. In various cancer types, such as oesophageal squamous cell carcinoma, researchers have simultaneously investigated genomic alterations and gene expression patterns [22, 77, 124, 125, 127, 131, 132, 136, 138, 321]. Expression profiling has proven its usefulness in identifying clinical biomarkers in different cancers [322, 323], and uncovering the genomic alterations responsible for changes in gene expression profiles can aid in predicting driver genes [324]. It is recognised that gene expression changes indicative of patient outcomes are subtle and individual genes alone are unlikely to effectively predict clinical behaviour [323]. Thus, analysis of gene expression and biological pathways associated with cancer could provide a better understanding of molecular alterations and mechanisms of carcinogenesis, thereby identifying potential targets for cancer diagnosis, prognosis, and treatment [320, 322, 323, 325].

Various studies, especially in Chinese and Japanese cohorts, have characterized genomic abnormalities in OSCC and have explored potential targets and associated pathways [123-125, 127, 128, 136, 139, 141, 143, 155-157, 213, 239, 296]. Exploration of gene mutations and their correlation with OSCC development, prognosis, and progression has led to identification of driver genes and potential biomarkers. The analysis by Song et al., [125] of 17 OSCC samples by whole genome sequencing and 71 OSCC samples by whole exome sequencing identified *FAM135B* as a prognostic marker for OSCC in China. They reported that *FAM135B* expression promoted malignancy in oesophageal squamous cell carcinoma cells. In another study, Du et al., [131] found *AJUBA* to be frequently disrupted with truncating and frameshift mutations. Moreover, they found higher *AJUBA* expression in OSCC tumour tissues compared to normal samples, however, lower levels in *AJUBA*-mutated tumour samples than in *AJUBA*-wild-type tumour samples in OSCC. Lin et al., [124] identified *ZNF750* deletions in 3.4% of OSCC tumours, with lower mRNA levels observed in oesophageal tumours compared with normal tissue. Furthermore, depletion of *ZNF750* promoted cell proliferation in OSCC. A recent study

by Cui et al., [136] identified *NFE2L2* as a tumour suppressor in OSCC, and that *NFE2L2* mutations could impair its tumour-suppressive function or even confer oncogenic activities. Immunohistochemistry analysis showed lower expression of *NFE2L2* in *NFE2L2* mutant tumours compared to normal tissues. Knockdown of wild-type *NFE2L2* increased cell proliferation, while knockout of mutant *NFE2L2* decreased cell proliferation in OSCC cell lines. Further analysis showed that the *NFE2L2* mutations (p.R34Q and p.E79K) hindered the tumour-suppressive activity of *NFE2L2* and exerted an oncogenic role [136].

Given that it is unlikely for a single gene to independently cause cancer or manifest clinical phenotypes [323], the accumulation and combination of mutations in driver genes may hold more significance than their individual occurrence in cancer development and prognosis [326]. Therefore, the importance of analysing driver gene mutations should be adjusted to focus on exploring the crosstalk of genes within pathways and investigating their biological functions and role in cancer development as a combination of genes rather than single genes [128, 246]. Previous studies have found significantly disrupted oncogenic pathways in OSCC, including those involved in cell cycle and apoptosis regulation, *KEAP1-NFE2L2* pathway, "histone modification pathway", NOTCH signalling, Hippo pathway, PI3K/AKT pathway, DNA-repair and Wnt pathway [77, 122-134, 136-143, 156]. Among these, cell cycle regulators constituted the most frequently disrupted pathway in oesophageal squamous cell carcinoma, via genetic mutations, deletions and amplifications, and/or altered expression [128, 131, 136, 138, 140, 142, 156, 246]. In addition to loss-of-function *TP53* mutations and amplification of Cyclin D1 (*CCND1*), frequent truncating mutations (nonsense mutations or frameshift indels), and deletion of *CDKN2A* have been reported in OSCC. Alterations in *CDKN2A* have been linked to defects in the G1/S transition of the cell cycle and associated with survival and senescence [167]. Another gene that has been shown to play an important role in regulating cell senescence is *NFE2L2* [327]. Apart from its role in regulating protective responses against reactive oxygen species (ROS), delaying cell senescence and preventing age-related diseases [214, 327, 328], recent evidence indicates *NFE2L2* involvement in cell cycle pathways through it is association with and activation by p21, leading to enhanced reactive oxygen species protection and resistance to chemotherapeutic agents [329], indicating a crosstalk between p21-realated pathway and *KEAP1-NFE2L2* signalling pathway.

Defects in DNA damage repair (DDR) pathways are implicated in genomic instability and oncogenesis [330, 331]. Previous studies have evaluated genetic variations revealing a

significant association between DNA repair pathways and OSCC [330]. A recent study by Li and colleagues [156], integrated OSCC sequencing data from 1930 patients across 33 studies found somatic mutations in DNA repair pathway genes in more than 27% of OSCC samples, including genes such as *BRCA2*, *TDG*, *FANCM*, *RIF*, *ATM*, *POLE*, and *CHEK2*. Furthermore, tumours with one or more somatic mutations in these genes had significantly elevated mutational loads and greater contributions from mutation signature SBS2 and SBS13, which are associated with APOBEC related processes [156].

Our analysis of 31 patients by whole-genome sequencing and 67 patients by whole exome sequencing analysis found mutations in genes such as *TP53*, *CDKN2A* (encoding proteins p14ARF and p16INK4a), *NFE2L2*, *NOTCH1* and *ZNF750*, or aberrant pathways, such as cell cycle *NFE2L2-KEAP* pathway, cell cycle and NOTCH pathways. These pathways may drive the development of OSCC. Despite recent efforts to characterise genomic alterations in OSCC [22, 77, 122-125, 127, 128, 130-133, 135-137, 139, 140, 142, 143, 154-157, 239], the number of clinically relevant biomarkers for this disease are still limited [136]. Additionally, there is a need for more genomic studies on OSCC within African populations to enhance our understanding of the tumorigenic processes and develop better diagnostic and prognostic tools and therapeutic targets in Africa [16, 19, 31]. This part of the study aimed to characterise the extent of alteration in selected genes by investigating the expression profile of genes involved in cell cycle and *NFE2L2-KEAP* pathways; *CDKN2A* products: p14ARF and p16INK4a and *NFE2L2*, respectively, in two sets of samples. The first set of samples included nine matched tumour-normal pairs that were sequenced using WGS, which were found to have mutations in *CDKN2A* and *NFE2L2* (Chapter 2), as well as a larger sample size of 79 matched OSCC tumour-normal pairs, the second set of samples. The objective was to determine gene expression patterns as potential biomarkers for OSCC risk and/or predictors of patients' outcome. The nine sets of samples were used to examine possible trends in tumour compared to wild-type, normal adjacent tissue, and subsequently, to validate gene expression patterns in a larger sample size of 79 matched tumour-normal pairs. In addition, several DNA damage repair genes, including *TOPBP1*, *ERCC6* and *C20orf196* were found to be frequently mutated in our cohort. Considering that we found mutation signatures associated with defects in DNA repair including mutation signatures SBS10b, SBS6 and SBS15, we examined the expression of *TOPBP1*, *ERCC6* and *C20orf196* in the two sets of samples.

3.2.Results

3.2.1. Epidemiological characteristics of the OSCC patients

Tumour, adjacent normal tissue, and demographic information were collected from patients with OSCC. Epidemiological data and lifestyle risk factors, including age, gender, tumour differentiation, histopathology of tumours, smoking status, and survival status of patients, are listed in Table 3.1. During the follow-up period, 46 patients were reported deceased, while 10 patients were alive, and 23 patients could not be contacted on their last follow up.

Table 3.1 Characteristics of patients with OSCC enrolled in this cohort.

	(n=79)
Site	**GSH**
Clinical factors	
Gender	
Female	36
Male	43
Age (years)	
Median	59
Min-Max	28-82
Tumour differentiation	
Poor	10
Moderate	54
Well	6
No info	9
Smoking history	
Yes	25
Ex-smoker	35
No	19
Ethnicity	
Black	47
Mixed ancestry	32
Survival status at last follow-up	
Alive	10
Deceased	46
Lost to follow up	23
Histology	
OSCC	79

GSH - Groote Schuur Hospital, n - number of cases, OSCC - oesophageal squamous cell carcinoma.

3.2.2 Cell cycle regulators: *p14ARF* and *p16INK4a* mRNA levels in OSCC

The *CDKN2A* gene locus codes for two distinct tumour suppressor proteins; *p16INK4a* and *p14ARF* (alternative reading frame) [164, 332], which are generated by alternative mRNA splicing by using alternate reading frames [163]. These two transcripts share common exons 2 and 3 but have alternatively spliced first exons. Exon 1β codes for p14ARF, located 15 kb upstream of exon 1α, which encodes p16INK4a (Figure 3.1). These two transcripts are translated into two distinct proteins with two overlapping ORFs (open reading frames) starting with different first exons. As a result, they share no amino acid sequence identity and have distinct functions [164, 169, 178]. The p16INK4a protein is a 156-amino acid polypeptide [333], which induces a G1 cell cycle arrest by inhibiting the phosphorylation of the retinoblastoma protein (Rb) by the cyclin-dependent kinases, CDK4 and CDK6. In contrast, p14ARF is a polypeptide of 132 amino acids [169], which activates a p53 response by promoting degradation of MDM2 and induces cell cycle arrest in both G1 and G2/M [163, 167, 168, 179]. Alterations in the *CDKN2A* locus which functionally impair both p14ARF and p16INK4a expression have the potential to impair both Rb and p53 pathways [169]. Furthermore, in many cancers, both p16INK4a and p14ARF have been reported to be inactivated [167, 169, 170, 179, 185].

Given the frequencies of *CDKN2A* mutations in our cohort—35% (11 out of 31) in the WGS cohort and 10% (7 out of 67) in the WES cohort—and considering that *CDKN2A* mutations can affect both *p14ARF* and *p16INK4a*, our initial analysis focused on evaluating the overall mRNA levels of *p14ARF* and *p16INK4a* (reflecting the combined expression levels of *p14ARF* and *p16INK4a*). This analysis was performed on nine matched tumour-normal pairs to gain insight into how *CDKN2A* mutations influence the overall expression from this gene locus, using reverse-transcription quantitative PCR (RT-qPCR). These pairs included tumour samples with *CDKN2A* mutations identified through WGS analysis (see Chapter 2). Primers for detecting the overall *p14ARF-p16INK4a* mRNA levels were designed in exon 3, targeting a common region (Figure 3.1), as previously described by Nakashima et al., [334]. Analysis of the RT-qPCR data was conducted using the comparative threshold cycle (CT) method [335], to calculate the expression fold changes between samples.

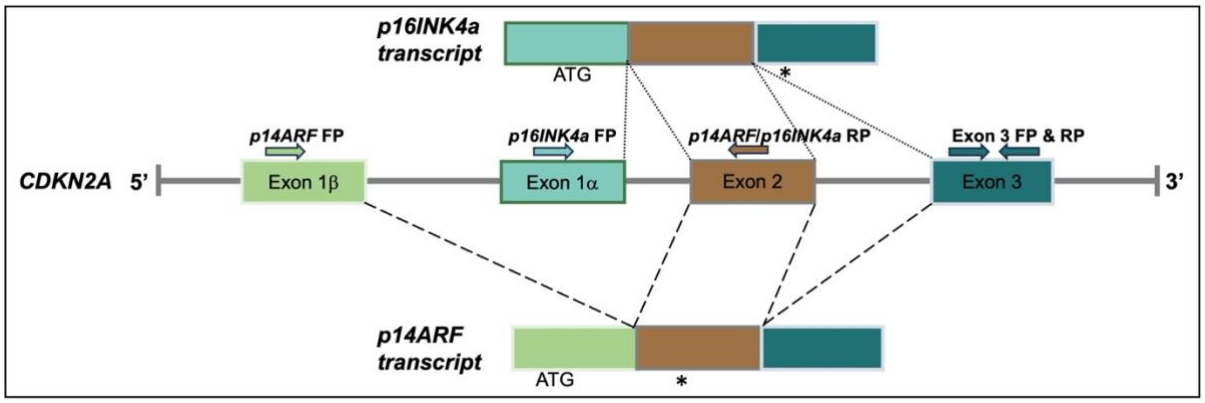


Figure 3.1 Schematic representation of the genomic structure of the *CDKN2A* locus, *p14ARF* and *p16INK4a* transcripts and primer design.

CDKN2A gene produces two proteins, p161NK4a and p14ARF. Although *p14ARF* and *p16INK4a* share exon 2 and exon 3, they differ in their first exons. *p14ARF* includes exon 1β, while *p16INK4a* includes exon 1α. The long-dashed lines below the genomic structure indicate the transcribed parts of mature *p14ARF* mRNA, while the dotted lines above the genomic structure indicate the transcribed parts of mature *p16INK4a* mRNA. The asterisks (*) indicate stop codons. For *p14ARF*, the stop codon is in exon 2, whereas for *p16INK4a*, it is in exon 3. Arrows above each exon show the locations where primers are designed. For *p16INK4a* and *p14ARF*, two forward primers were designed; one forward primer from exon 1β for *p14ARF* amplification and the other forward primer designed from exon 1α for *p16INK4a* amplification with a common reverse primer in exon 2. Primers to examine overall *p14ARF-p16INK4a* mRNA levels were designed in exon 3, targeting a common region. Sequences for *p16INK4a* and *p14ARF* primers were described previously by Burri et al., [336], while exon 3 primers were previously described by Nakashima et al., [334]. Figure was modified from Fontana et al., [169]. FP- forward primer, RP- reverse primer, ATG - start codon (encode for the amino acid methionine).

Analysis of the overall *p14ARF-p16INK4a* mRNA levels revealed variability between tumour samples and their corresponding normal tissues (Figure 3.2). Elevated mRNA levels were observed in 3 out of 9 tumour samples, including those with *CDKN2A* mutations: sample PD39449, which has a 40 bp frameshift insertion, and case PD50651 with a nonsense mutation. Notably, sample PD39452, despite lacking *CDKN2A* mutations, also exhibited elevated combined *p14ARF-p16INK4a* mRNA levels. This elevation in sample PD39452, could be attributed to various factors affecting gene expression, such as epigenetic modifications, transcription factors, copy number variations, and cellular signalling pathways [179, 189]. In contrast, 3 out of 9 tumour samples displayed significantly reduced overall *p14ARF-p16INK4a* mRNA levels compared to their adjacent normal tissues (Figure 3.2). All three of these samples did not have *CDKN2A* mutations.

Unfortunately, further analysis couldn't be conducted on some samples found with *CDKN2A* mutations (listed in Chapter 2, Table 2.8) due to insufficient RNA samples. Therefore, mRNA levels in some samples with *CDKN2A* mutations, such as PD39451 and PD39455 with the [p.R58*] mutation, and sample PD39453 with the [p.D108N] mutation, could not be evaluated.

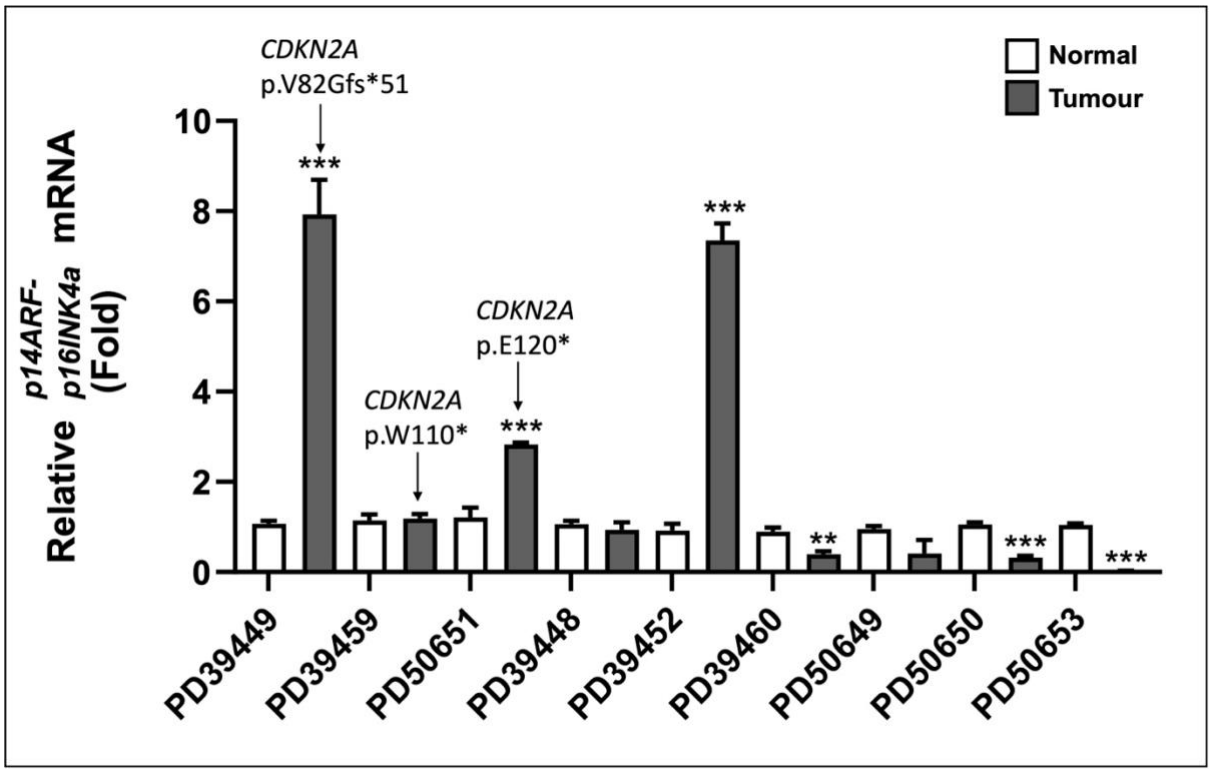


Figure 3.2 Overall *p14ARF-p16INK4a* mRNA levels in OSCC samples.

The overall *p14ARF-p16INK4a* mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed variable mRNA levels in tumours compared to adjacent normal tissues. (n=9). 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The primers pairs used to detect the overall *p14ARF-p16INK4a* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. RT-qPCR primers were designed in exon 3 of *CDKN2A* gene which were described previously by Nakashima et al., [334]. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Considering the observed variability in overall *p14ARF-p16INK4a* mRNA levels in tumours compared to adjacent normal tissues, as well as the differences between *CDKN2A*-mutated tumour samples and *CDKN2A* wild-type tumour samples, it is noteworthy that mutations in *CDKN2A* exon 2 have the potential to affect the expression of both *p16INK4a* and *p14ARF* genes [170]. In the case of *p16INK4a*, truncating mutations in PD39449 and PD50651 samples were present within the key ankyrin repeat domains (ankyrin repeat 3 and ankyrin repeat 4) of the p16INK4a protein. Previous studies have indicated that mutations in these domains lead to the expression of truncated p16INK4a protein, lacking CDK binding activity [189]. On the other hand, tumour-associated mutations in exon 2 of *p14ARF* have been shown to functionally impair p14ARF activity by reducing its ability to stabilize p53 [337]. To further investigate the expression patterns of these genes in OSCC, our subsequent analysis focused on assessing the

individual mRNA levels of *p14ARF* and *p16INK4a* in a larger cohort of 79 matched tumour-normal pairs, a separate set of samples, distinct from those analysed using WGS and WES, using RT-qPCR.

For this analysis, primers were designed to specifically detect either *p14ARF* or *p16INK4a* mRNA levels, as previously described by Burri et al., [336]. A 210-bp fragment with a forward primer designed based on the sequence of exon 1α of the *p16INK4a* locus was used to detect *p16INK4a* transcript, while a 207-bp fragment with a forward primer designed in exon 1β of *p14ARF* was used detect *p14ARF* transcript. Both utilized a common reverse primer in exon 2 (Figure 3.1). The RT-qPCR analysis was validated by melting curve analysis, confirming the specific amplification of *p14ARF* and p*16INK4a* transcripts (data not shown).

Our qPCR analysis revealed variable levels of *p16INK4*a and *p14ARF* mRNA in tumours compared with corresponding adjacent normal tissues (Figure 3.3B and Figure 3.4B). *p14ARF* mRNA levels were significantly lower in 38 out of 79 (48%) tumours (Figure 3.3A), whereas *p16INK4a* mRNA levels were significantly lower in 48 out of 79 (61%) tumour tissue samples (Figure 3.4A). Furthermore, 26 out of 79 (33%) tumour samples had extremely low to undetectable mRNA levels for both *p16INK4*a and *p14ARF* when compared to adjacent normal controls (Figure 3.3A and Figure 3.4A). Notably, some tumours displayed significantly elevated *p16INK4*a and *p14ARF* mRNA levels. p14ARF mRNA levels were significantly elevated in 20 out of 79 (25%) tumours compared with corresponding normal adjacent tissues, while p16INK4a mRNA levels were significantly elevated in 13 out of 79 (16%) tumour samples when compared to their normal adjacent tissues (Figure 3.3A and Figure 3.4A).

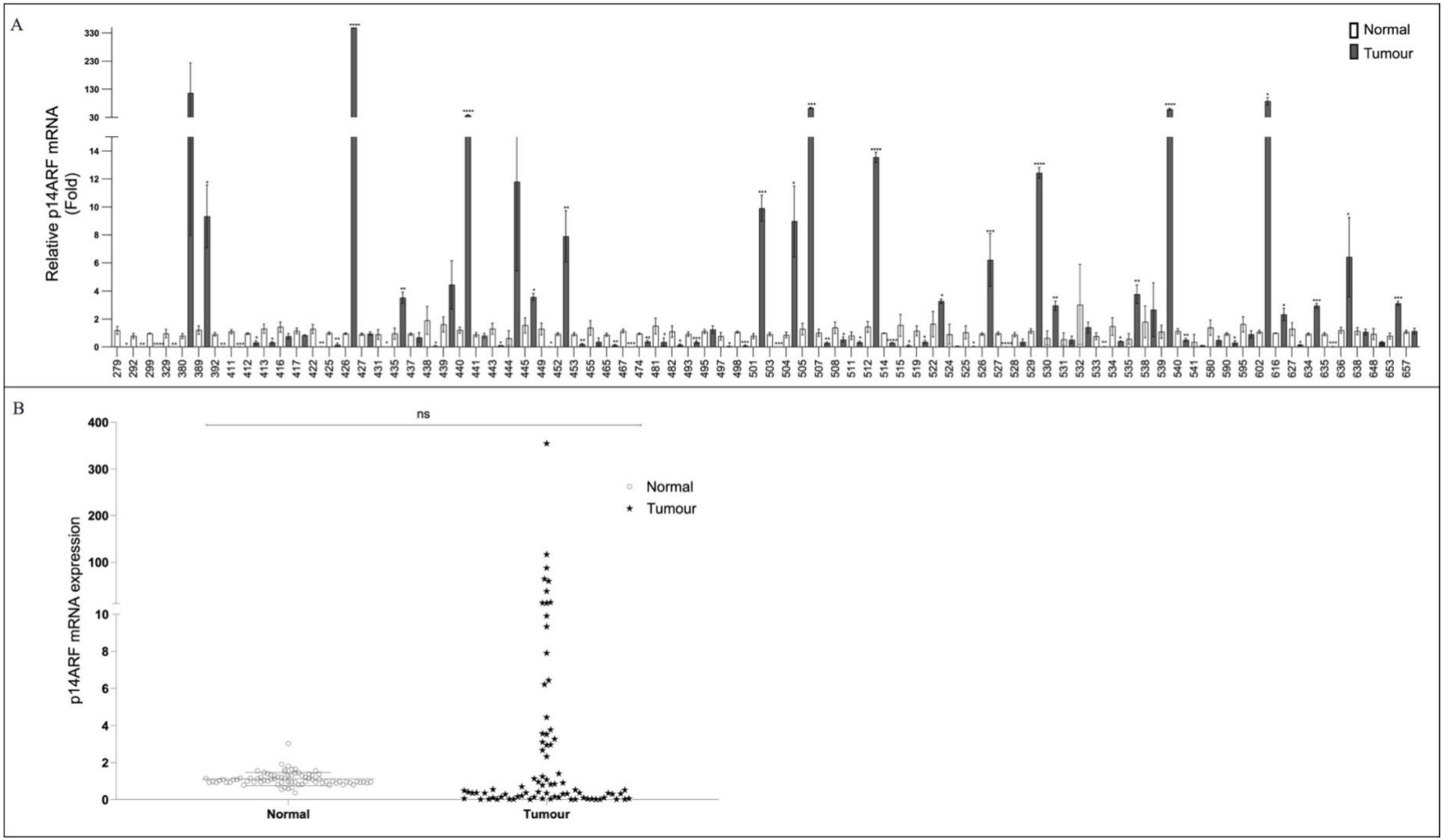


Figure 3.3 Relative *p14ARF* mRNA levels in OSCC samples.

p14ARF mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed variable expression of *p14ARF* in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *p14ARF* specific primers pairs used to detect *p14ARF* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. **A**) RT-qPCR analysis of *p14ARF* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *p14ARF* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3

replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

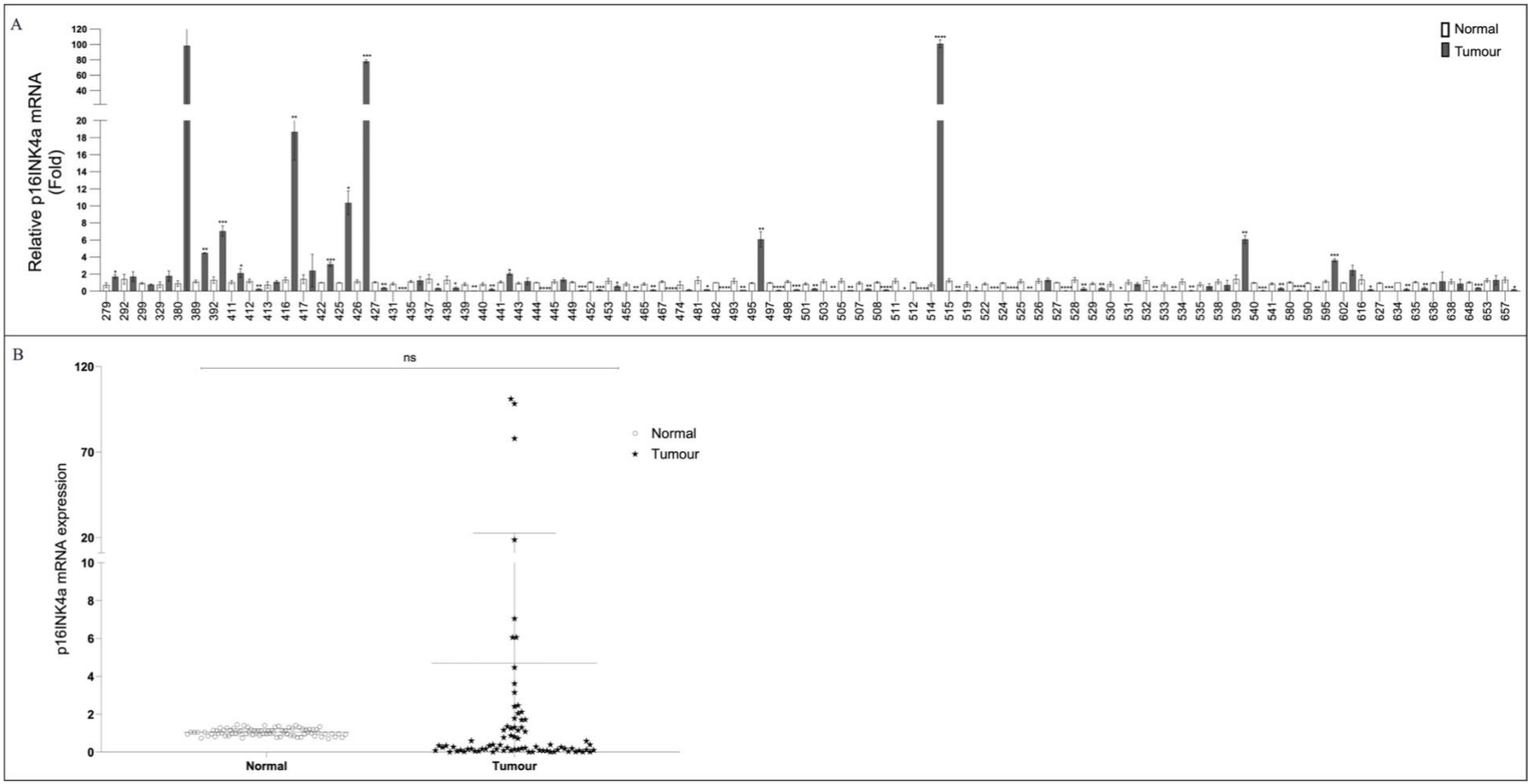


Figure 3.4 Relative *p16INK4a* mRNA levels in OSCC samples.

p16INK4a mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed variable expression of *p16INK4a* in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *p16INK4a* specific primers pairs used to detect *p16INK4a* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. **A**) RT-qPCR analysis of *p16INK4a* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *p16INK4a* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation

(SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

In summary, our analysis showed variable overall *p14ARF-p16INK4a* mRNA levels in OSCC tumour tissues as compared with their adjacent normal tissues. The combined *p14ARF-p16INK4a* mRNA levels were significantly elevated in tumour samples, including two *CDKN2A*-mutated tumour samples, and significantly low in three samples containing wild-type *CDKN2A*. Furthermore, individual *p16INK4*a and *p14ARF* showed variable mRNA levels in OSCC tumour tissues. Notably, *p16INK4a* and *p14ARF* mRNA levels were significantly lower in 61% and 48% of tumour samples, respectively, compared to their adjacent normal tissues. Additionally, our analysis identified significantly elevated levels of *p16INK4a* and *p14ARF* in 16% and 25% of tumours, respectively, compared to their normal adjacent tissues. This aligns with previous findings showing reduced or absent *p16INK4a* and/or *p14ARF* expression in OSCC samples, particularly in Asian populations.[174-176, 192]. De Almeida Simao and colleagues [192] reported variable expression of *p14ARF* and *p16INK4a*, with suppressed or lower *p14ARF* and *p16INK4a* mRNA levels in 58.8% and 64.7% of the OSCC samples analysed, respectively. Similarly, Xing et al., [175] observed lower expression of *p16INK4a* and *p14ARF* mRNAs in 12 of 18 (67%) Chinese OSCC samples, with 4 of the 18 (22%) samples maintaining elevated mRNA levels for both these genes. The variable expression of *p16INK4*a and *p14ARF* in tumours compared to normal tissues may indicate inter-tumoral and intra-tumoral heterogeneity within OSCC [284], highlighting the complexity of OSCC within this patient population. Furthermore, the differential expression of *p14ARF* and *p16INK4a* in cancer can be influenced by various factors, including genetic alterations, epigenetic modifications, and interactions with other signalling pathways [170, 174, 175, 197, 338]. The differential expression of *p16INK4a* and *p14ARF* observed in some OSCC patients may be attributable to truncating mutations and oncogenic missense mutations identified in *CDKN2A*. These mutations have the potential to influence the stability of mRNA for these genes. This altered expression pattern of *p16INK4*a and *p14ARF* appear to be a common occurrence in OSCC, suggesting that p*16INK4*a and *p14ARF* may serve as putative driver genes for OSCC.

3.2.3 *NFE2L2-KEAP* pathway: *NFE2L2* gene expression in OSCC patients

The *KEAP1-NFE2L2* pathway is crucial for controlling cellular stress by regulating the expression of genes responsible for antioxidant response and detoxification, thus restoring cellular homeostasis. Frequent mutations in the key regulators of the *KEAP1-NFE2L2* pathway including *NFE2L2*, *KEAP1*, and *CUL3*, have been identified in various cancers, including OSCC [128, 131, 132, 136, 139]. In our cohort, the *KEAP1-NFE2L2* pathway was deregulated

with mutations in *NFE2L2*. *NFE2L2*, also known as *NRF2*, serve as a transcriptional activator that regulates the expression of target genes such as NAD(P)H quinine oxidoreductase-1 (*NQO1*) and peroxiredoxins (*PRDX*) in response to oxidative stress [214, 216]. Normally, low levels of *NFE2L2* are maintained through proteasomal degradation through the KEAP1/CUL-dependent proteasomal pathway under non-stressed conditions [216]. Activation of *NFE2L2* enhances cellular tolerance to oxidative stress, providing growth advantages to cells [339]. In our cohort, the majority of *NFE2L2* variants occurred in exon 2, resulting in amino acid changes within the DLG or ETGE motifs. Mutations in these motifs have been shown to disrupt binding to the KEAP1 dimer, inhibiting KEAP1-mediated degradation of NFE2L2 [216]. Studies have demonstrated that certain *NFE2L2* mutants, such as p.R34Q, promote tumour growth in OSCC, and *NFE2L2* mutations are significantly associated with poor prognosis in OSCC [136].

Previous studies have shown the oncogenic potential of *NFE2L2* mutations that disrupt the KEAP1-binding site in various squamous cell carcinomas, including OSCC, particularly in regions such as Japan and China [136, 216, 221]. However, there is a lack of information regarding *NFE2L2* alterations and gene expression patterns associated with OSCC in South Africa. To address this gap, we investigated the *NFE2L2* mRNA levels in a sample of nine OSCC samples, which included tumour samples identified with *NFE2L2* mutations by WGS. In addition, we analysed the mRNA levels of *NFE2L2* in a larger sample size of 79 matched tumour-normal pairs, distinct from those analysed using WGS and WES.

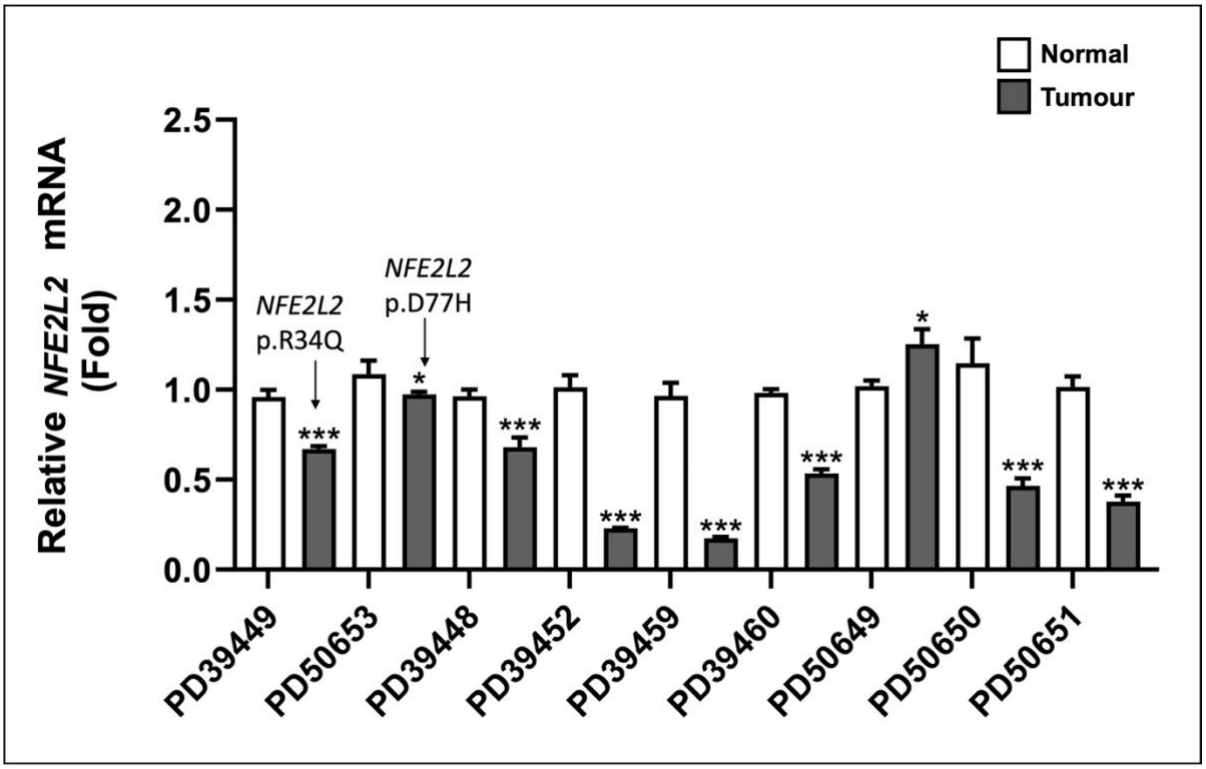


Figure 3.5 Relative *NFE2L2* mRNA levels in OSCC samples.

NFE2L2 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed lower *NFE2L2* mRNA levels in tumours compared to adjacent normal tissues. (n=9). 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *NFE2L2* specific primers pairs used to detect *NFE2L2* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Among the nine tumour samples, eight tumour samples had significantly lower *NFE2L2* mRNA levels compared to their corresponding normal adjacent tissues (Figure 3.5). This includes case PD39449 with c.101G>A mutation [p.R34Q] and PD50653 with c.229 G>C mutation [p.D77H].

Having observed *NFE2L2* expression pattern in OSCC tumours, including those with *NFE2L2* mutations, in the initial nine paired normal and tumour samples (included tumour samples identified with *NFE2L2* mutations by WGS), we subsequentially examined *NFE2L2* mRNA levels in a larger sample size of 79 matched tumour-normal pairs (distinct from those analysed using WGS and WES). This was done to gain a comprehensive understanding of *NFE2L2* expression patterns associated with OSCC in South African population.

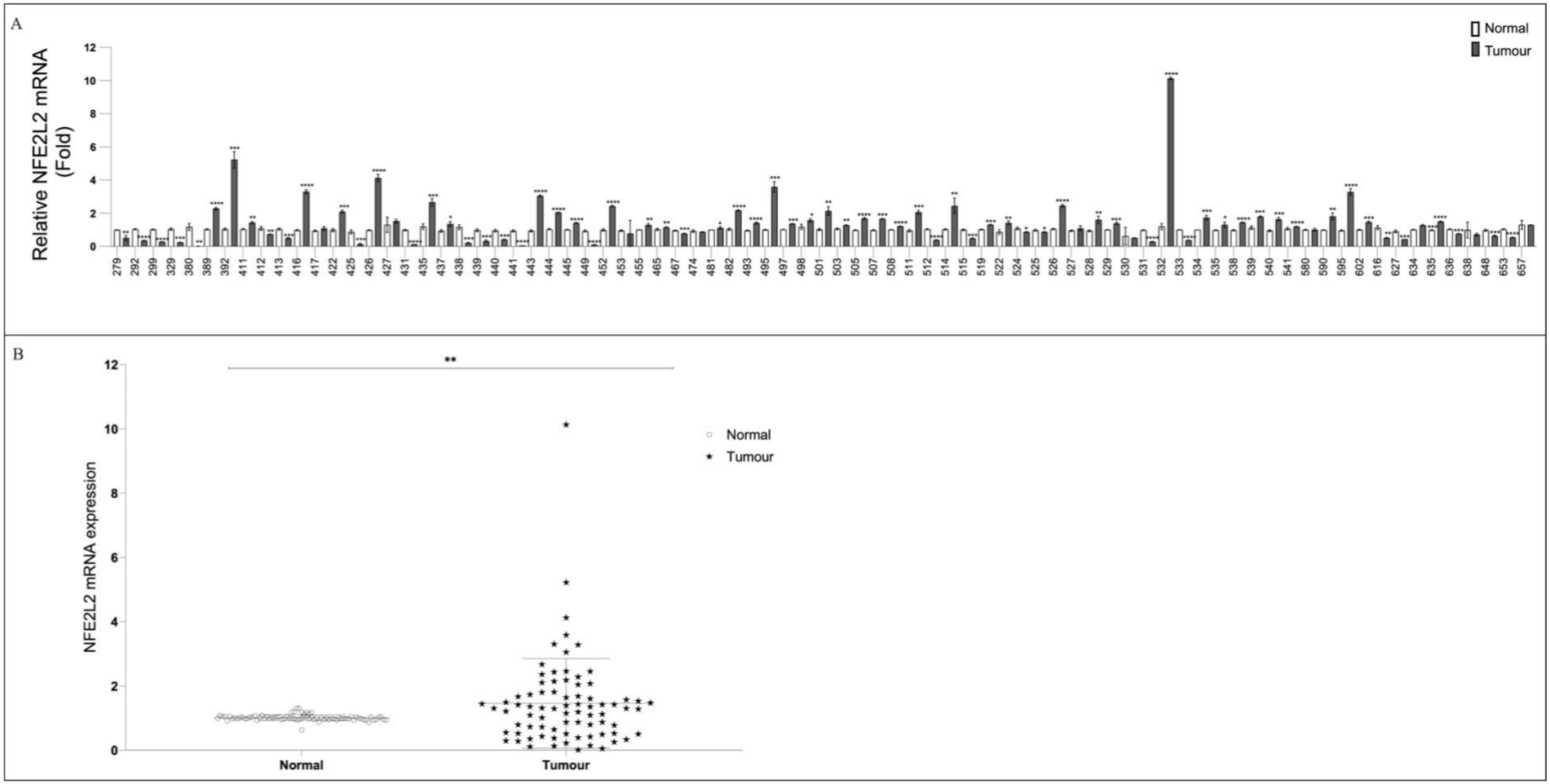


Figure 3.6 Relative *NFE2L2* mRNA levels in OSCC samples.

NFE2L2 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed variable expression of *NFE2L2* in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *NFE2L2* specific primers pairs used to detect *NFE2L2* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12 **A**) RT-qPCR analysis of *NFE2L2* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *NFE2L2* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

Our analysis showed variable expression of *NFE2L2* in OSCC tumour tissues compared to their corresponding adjacent normal tissues, with a significant difference observed in *NFE2L2* mRNA levels between tumour and normal samples ($p < 0.01$) (Figure 3.6B). *NFE2L2* mRNA levels were significantly elevated in 44 out of 79 (56%) tumours, while 26 of 79 (33%) tumours showed significantly lower *NFE2L2* mRNA levels in OSCC tumour tissues when compared with their adjacent normal tissues (Figure 3.6A). In the nine samples analysed, including those with *NFE2L2* mutations (PD39449 [p.R34Q] and PD50653 [p.D77H]), *NFE2L2* mRNA levels were significantly lower compared to their adjacent normal tissues. This finding aligns with a recent study by Cui et al., [136], which also reported decreased expression of *NFE2L2* in *NFE2L2* mutant samples compared to wild-type samples. Additionally, their knockout studies of mutant *NFE2L2* reduced cell proliferation in OSCC cell lines [136]. The contrast between the smaller panel of nine samples and the larger panel of 79 samples highlights the potential impact of sample size on observed outcomes. While the smaller panel showed a clear reduction in *NFE2L2* mRNA levels possibly linked to mutations, the larger panel revealed a broader range of expression levels, suggesting that *NFE2L2* amplification could be influencing results in a subset of cases. Therefore, the low *NFE2L2* expression observed in 33% of OSCC patients may be associated with missense mutations within the DLG/ETGE motifs of *NFE2L2*, whereas the elevated *NFE2L2* expression in 56% of patients may be attributed to *NFE2L2* amplification [136, 221, 340].

In summary, the variable expression of *NFE2L2* in OSCC suggests potential dysregulation of this gene in OSCC. Our analysis found significantly lower *NFE2L2* mRNA levels in *NFE2L2*-mutated tumour samples compared to their adjacent normal tissues. Additionally, the majority (56%) of tumours exhibited significantly elevated *NFE2L2* mRNA levels relative to normal tissues, indicating a possible association between increased *NFE2L2* expression and OSCC development or progression in South African patients. In contrast, the role of decreased *NFE2L2* expression in OSCC pathogenesis remains not well understood, despite 33% of tumours showing significantly lower *NFE2L2* mRNA levels compared to adjacent normal tissues.

3.2.4 The DNA damage repair pathways

The DNA damage repair pathways are activated to repair damaged DNA and maintain genome integrity in response to different endogenous and exogenous stresses [341]. Defects in these pathways can lead to genomic instability, predisposing to tumorigenesis [330, 331]. Amplification and over expression of genes involved in DNA damage repair have been linked to poorer overall survival in OSCC patients [331]. In our cohort, we found frequent mutations in several DNA repair genes including topoisomerase (DNA) II binding protein 1 (*TOPBP1*), excision repair cross-complementing group 6 (*ERCC6*) and chromosome 20 open reading frame 196 (*C20orf196*) (see Chapter 2). Additionally, we found mutation signatures associated with defects in DNA repair including mutation signatures SBS10b, SBS6 and SBS15. Given these findings, we investigated the mRNA levels of *TOPBP1*, *ERCC6* and *C20orf196* in OSCC.

3.2.4.1 *TOPBP1* expression

Initially, we examined the *TOPBP1* mRNA levels in nine matched tumour/normal OSCC samples from patients with identified *TOPBP1* mutations by WGS. The results showed elevated *TOPBP1* mRNA levels in tumours compared with their corresponding normal adjacent tissue (Figure 3.7), where 8 out of 9 tumour samples had significantly higher *TOPBP1* mRNA levels. This includes sample PD39449, which had the c.1429T>A mutation [p.F477I] (Figure 3.7).

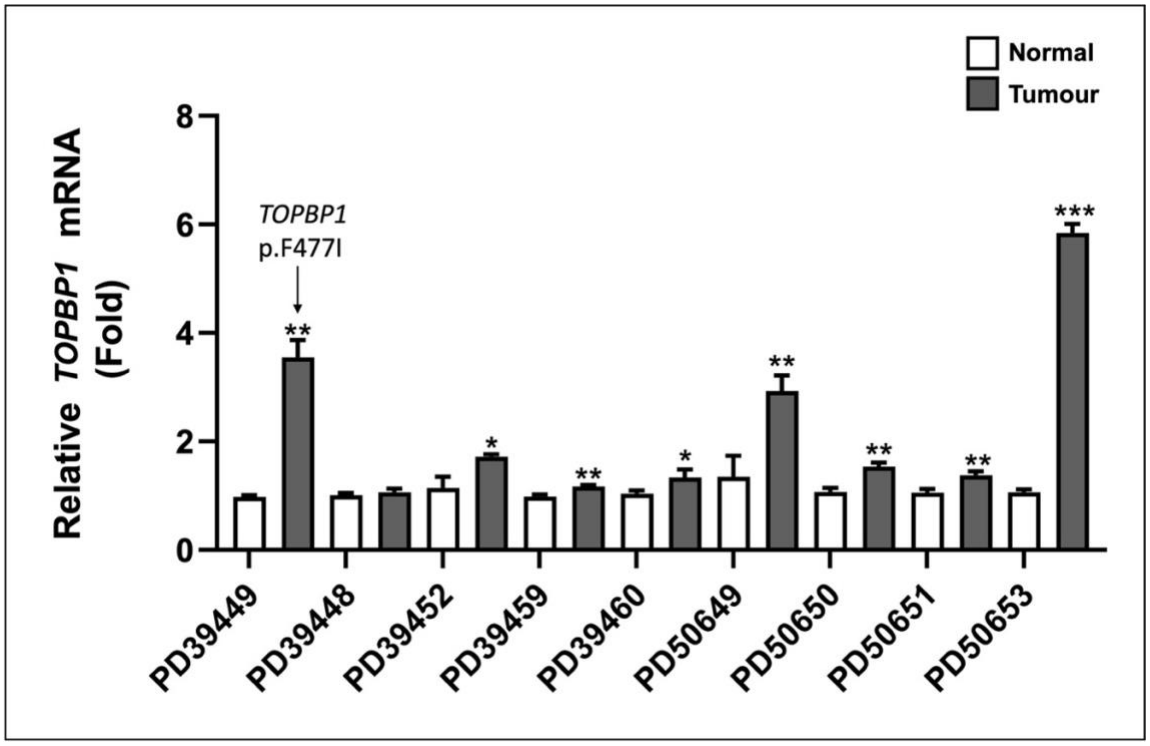


Figure 3.7 Relative *TOPBP1* mRNA levels in OSCC samples.

*TOPBP1*mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed elevated *TOPBP1* mRNA levels in tumours compared to adjacent normal tissues. (n=9). 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *TOPBP1* specific primers pairs used to detect *TOPBP1* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

To assess whether the observed elevated *TOPBP1* mRNA levels is a consistent trend in OSCC, as previously reported in OSCC [342, 343], we expanded our investigation of *TOPBP1* mRNA levels in a larger sample size of 79 matched tumour-normal pairs (distinct from those analysed using WGS and WES). Our findings demonstrated that *TOPBP1* mRNA levels were significantly higher in tumours compared to adjacent normal tissues ($p < 0.0001$) (Figure 3.8B). Moreover, *TOPBP1* mRNA levels were significantly elevated in 61 out of 79 (77%) tumours, whereas only four samples (5%) showed significantly reduced *TOPBP1* mRNA levels in tumour samples compared to normal adjacent tissue (Figure 3.8A).

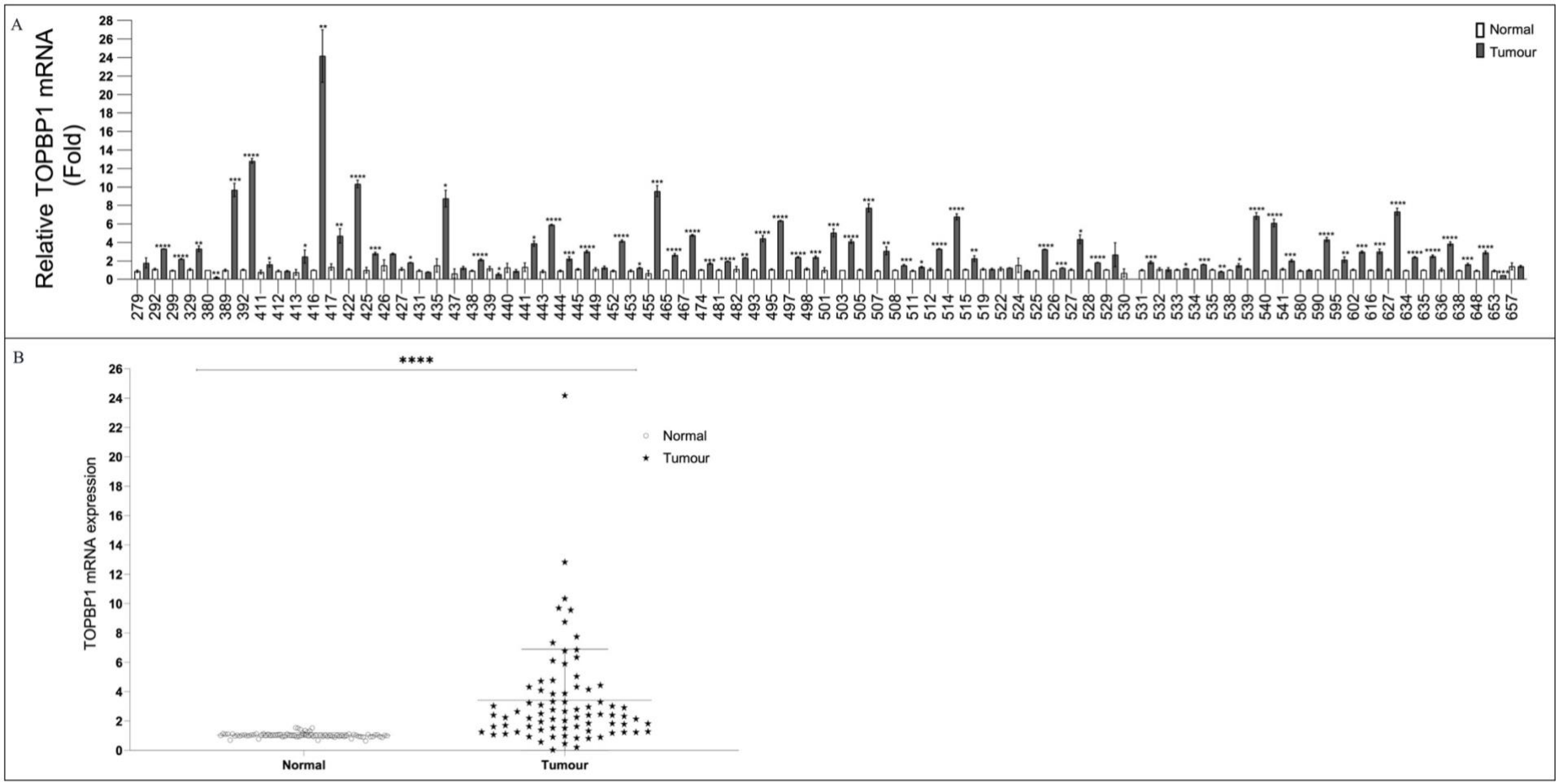


Figure 3.8 Relative *TOPBP1* mRNA levels in OSCC samples.

TOPBP1 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed higher *TOPBP1* mRNA levels in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *TOPBP1* specific primers pairs used to detect *TOPBP1* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. **A**) RT-qPCR analysis of *TOPBP1* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *TOPBP1* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

Previous studies have highlighted the crucial role of *TOPBP1* in initiating DNA replication and activating DNA damage checkpoint signalling as a scaffolding protein [344]. In our study, we observed *TOPBP*1 elevated in tumour samples, a feature that has previously been associated with poor survival and suggestive of its potential oncogene role in promoting OSCC progression [343]. Our finding implies that *TOPBP*1 upregulation may play a significant role in the development of OSCC.

3.2.4.2 *ERCC6* expression

Excision repair cross-complementing group 6 (*ERCC6*) also known as Cockayne syndrome group B (CSB), plays a role in transcription and nucleotide excision repair (NER), particularly in removing bulky adducts and repairing DNA damage induced by UV irradiation [345].

In our analysis, we examined *ERCC6* mRNA levels in nine OSCC samples, which included tumour samples with identified *ERCC6* mutations by WGS. Interestingly, we observed lower *ERCC6* mRNA levels in 6 out of 9 tumours compared to their normal adjacent tissue. This includes two tumour samples with *ERCC6* mutations: PD39448 and PD39460 with [p.S1085C] and [p.E130*] mutations, respectively (Figure 3.9).

However, when we examined *ERCC6* mRNA levels in a larger sample size of 79 matched tumour-normal pairs (distinct from those analysed using WGS and WES), *ERCC6* mRNA levels were significantly elevated in 48 of 79 (61%) tumours, while 6 of 79 (8%) tumours exhibited significantly lower *ERCC6* mRNA levels in OSCC tumour tissues compared with their adjacent normal tissues (Figure 3.10).

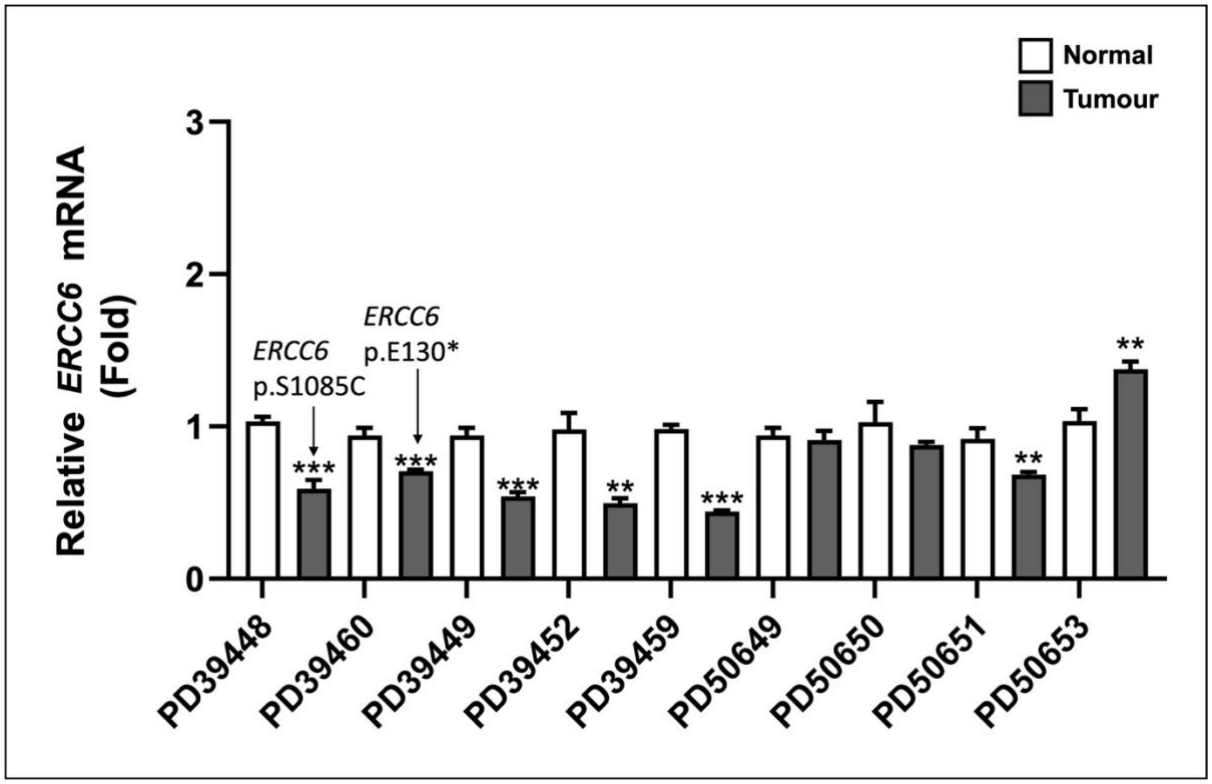


Figure 3.9 Relative *ERCC6* mRNA levels in OSCC samples.

ERCC6 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed lower *ERCC6* mRNA levels in six tumours compared to adjacent normal tissues and significantly higher in five tumours compared to adjacent normal tissues. (n=9). 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *ERCC6* specific primers pairs used to detect *ERCC6* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Despite *ERCC6* showing significant high mRNA levels in the majority of tumour samples compared to adjacent normal tissues in our larger cohort, two of OSCC samples with mutations in *ERCC6* showed significantly lower *ERCC6* mRNA levels compared to adjacent normal tissues. Previous studies have reported mutations in *ERCC6* in OSCC, with more amplification events than deletions in genes associated with non-homologous end joining (NHEJ), including *ERCC6* [331]. Given the relatively low frequency of *ERCC6* mutations in our cohort (6% in WGS cohort and none in WES cohort), the low mRNA levels of *ERCC6* observed in 8% of OSCC patients may be attributed to these mutations. In contrast, we speculate that the elevated mRNA levels of *ERCC6* observed in 61% of OSCC patients may potentially be attributed to *ERCC6* amplification.

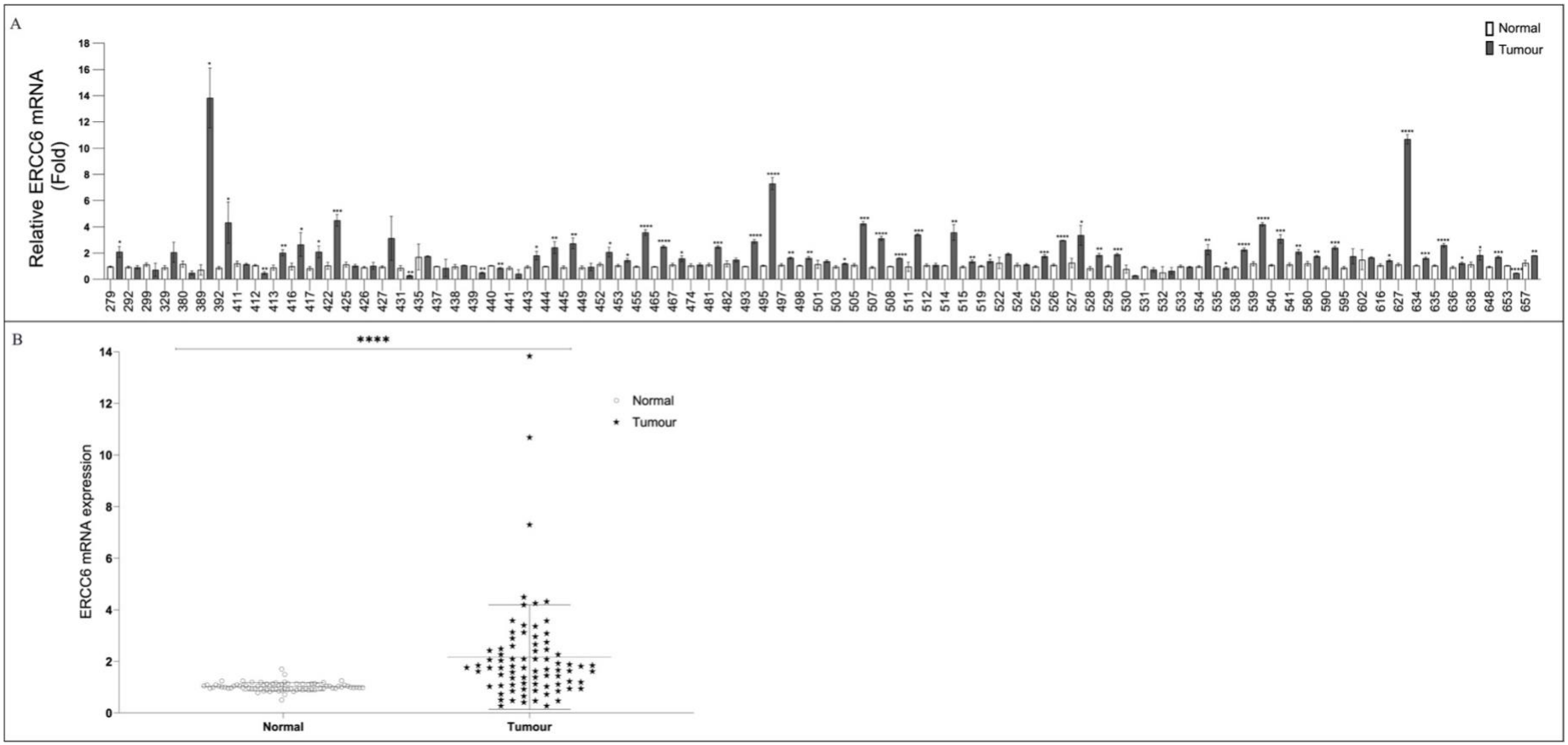


Figure 3.10 Relative *ERCC6* mRNA levels in OSCC samples.

ERCC6 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed elevated *ERCC6* mRNA levels in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *ERCC6* specific primers pairs used to detect *ERCC6* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. **A**) RT-qPCR analysis of *ERCC6* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *ERCC6* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

3.2.4.3 *C20orf196* expression

C20orf196, also known as SHLD1, forms a complex with FAM35A to promote NHEJ DNA repair [346, 347]. While previous studies have explored the role of *C20orf196* in DNA damage repair, its involvement in cancer remained poorly understood, and none have investigated *C20orf196* expression in OSCC samples.

In our study, we examined the mRNA levels of *C20orf196* in nine OSCC samples, including tumour samples with identified *C20orf1961* mutations by WGS. Among these samples, *C20orf196* mRNA levels were significantly elevated in two tumour samples, including sample PD39459 with the c.211T>A [p.S71T] mutation, and significantly low in two of the nine tumour samples (Figure 3.11).

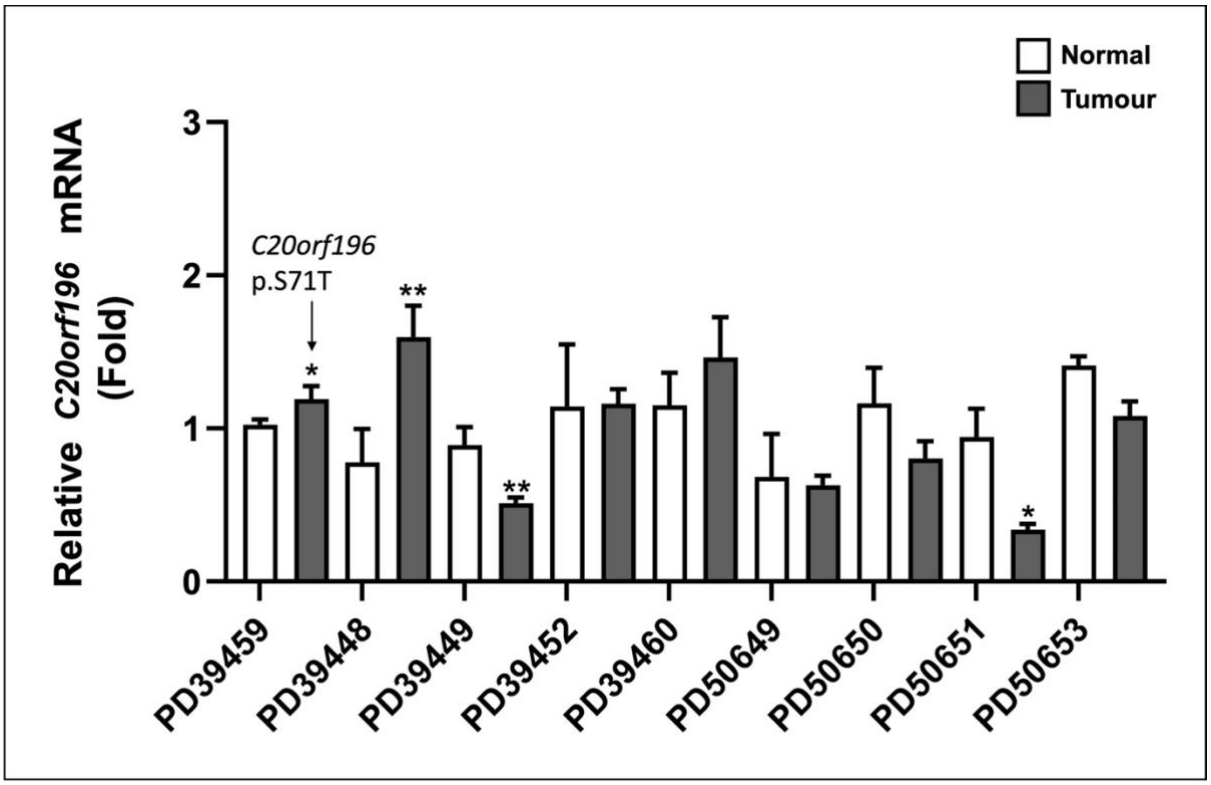


Figure 3.11 Relative *C20orf196* mRNA levels in OSCC samples.

C20orf196 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed elevated *C20orf196* mRNA levels in tumours compared to adjacent normal tissues. (n=9). 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *C20orf196* specific primers pairs used to detect *C20orf196* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

When analysing *C20orf196* mRNA levels in a larger sample size of 79 matched tumour-normal pairs, we observed variable mRNA levels in tumours compared to corresponding normal tissues ($p < 0.01$) (Figure 3.12B). *C20orf196* mRNA levels were significantly elevated in 28 out of 79 (35%) tumour samples and significantly decreased in 23 out of 79 (29%) tumour samples ($p < 0.01$) (Figure 3.12A). *C20orf196* mRNA levels exhibit significant variability in OSCC tumours compared to normal tissues. This variability includes both significant elevations and reductions in mRNA levels, the relevance of these observations are unclear.

The findings reported above yield several key conclusions regarding the role of DNA damage repair genes in OSCC. Firstly, the observed elevation in *TOPBP1* and *ERCC6* mRNA levels in OSCC suggests their potential involvement in OSCC development. Secondly, the differential expression of *TOPBP1,* and *ERCC6,* alongside the mutation frequencies detected in these genes, indicate the existence of multiple processes influencing gene expression in addition to the oncogenic missense mutations found in *TOPBP1*, *C20orf196*, and *ERCC6.* Lastly, these alterations in *TOPBP1*, *C20orf196* and *ERCC6* could contribute to malignant transformation and the growth of cancerous cells [171, 348, 349].

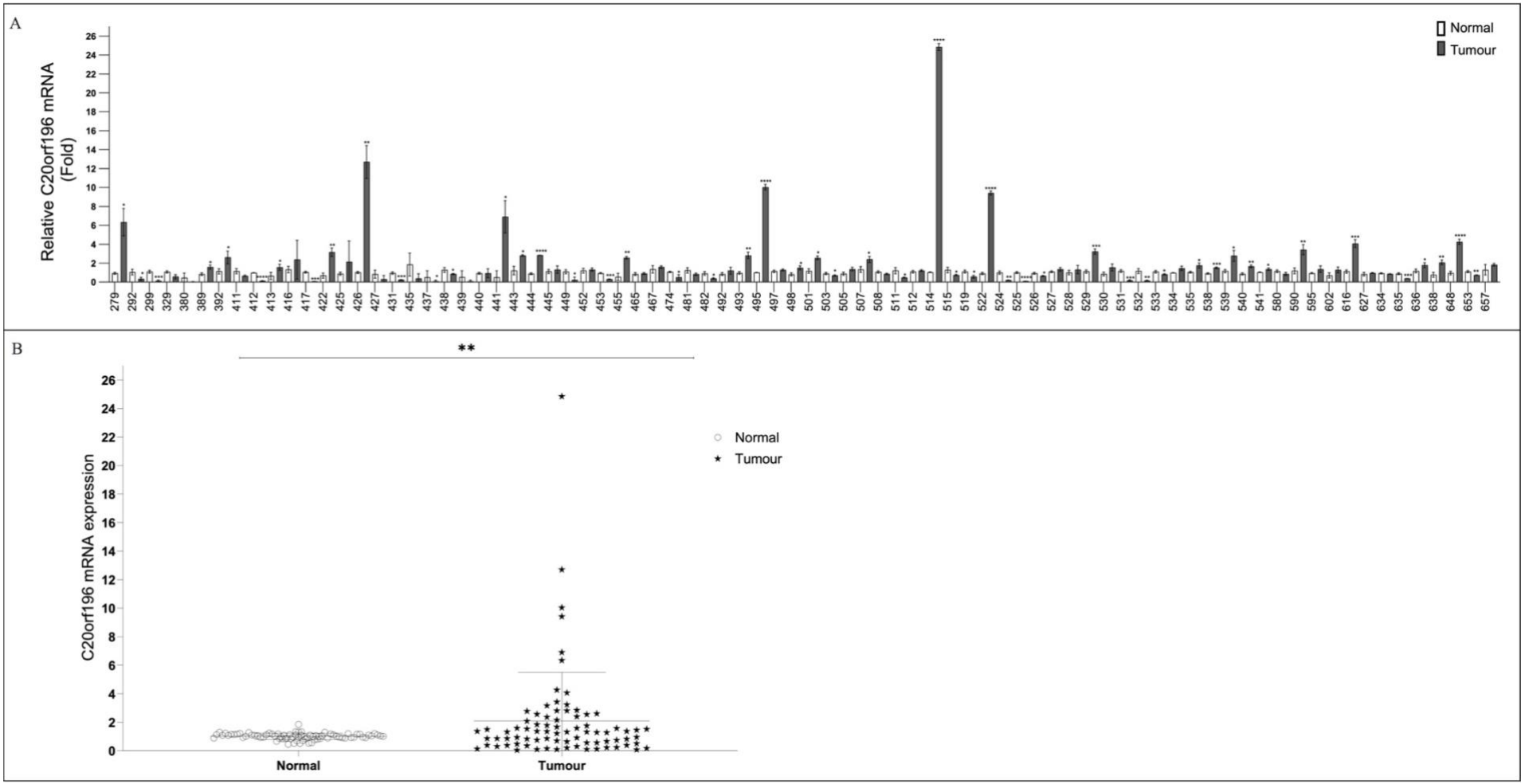


Figure 3.12 Relative *C20orf196* mRNA levels in OSCC samples.

C20orf196 mRNA levels in matched OSCC tumour-normal pairs analysed by quantitative real-time PCR showed variable *C20orf196* expression in OSCC tumour tissues as compared with their adjacent normal tissues. 200ng of total RNA extracted from matched OSCC tumour-normal pairs was used as template for cDNA synthesis as described in Materials and Methods in section 6.2.8.1. The *C20orf196* specific primers pairs used to detect *C20orf196* mRNA levels are indicated in Materials and Methods in section 6.2.8.2, Table 6.12. *GAPDH* was used for normalization. **A**) RT-qPCR analysis of *C20orf196* mRNA levels between tumour tissues and their adjacent normal tissues. (n=79). **B**) Scatter plots of *C20orf196* mRNA levels in cancer tissues and matched adjacent normal tissues from 79 cases of OSCC patients. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in tumour versus normal sample was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

3.2.5 Correlation of gene expression with clinicopathological data

Having observed significantly differential gene expression between OSCC tumour and normal adjacent tissue, we next investigated whether gene expression patterns could distinguish, and possibly give biological insight into, the clinical heterogeneity of the tumours among OSCC patients. We investigated whether there are any correlations between *p14ARF*, *p16INK4a*, *NFE2L2*, *TOPBP1*, *ERCC6* and *C20orf196* mRNA levels and clinical parameters reported in the study, including tumour differentiation and patient survival.

We investigated whether there is an association between gene expression levels (low/high) and tumour differentiation. To explore this relationship, we employed Chi-square analysis. A contingency table was constructed to show the frequencies of low/high mRNA levels and tumour differentiation categories, from which chi-square values and corresponding p-values were derived. This analysis encompassed all tumours with known differentiation status and significant mRNA levels classified as either low or high. Tumours lacking known differentiation status were excluded from the study. Additionally, to investigate the impact of the expression of selected genes on the overall survival of OSCC patients, we performed the survival analysis using the Kaplan–Meier (KM) estimator [350], with the p-value determined using the Log-rank (mantel-cox) test [351]. The Log-rank test was used to determine whether the difference in survival times between two groups (patients with high or low mRNA levels) is statistically different or not [350], specifically focusing on patients reported to have died. A p-value ≤ 0.05 was considered significant.

3.2.5.1 Correlation of *p14ARF* and *p16INK4a* mRNA levels with clinicopathological data

Tumours with either low (n=30) or high (n=14) *p14ARF* mRNA levels were more frequently found to be moderately differentiated tumours compared to those that were poor or well differentiated. However, our Chi-square test analysis found no significant association between *p14ARF* mRNA levels and tumour differentiation ($p = 0.1713$) (Figure 3.13A).

We acknowledge that the sample size for poorly and well-differentiated tumours in our cohort was quite small, with only 10 out of 79 and 6 out of 79, respectively). This limited representation of poorly and well-differentiated tumours may affect the reliability of our findings regarding the association between mRNA levels and tumour differentiation in OSCC

patients. Further studies with larger and diverse sample size are needed to validate and fully understand this relationship.

Next, we evaluated the association between *p14ARF* mRNA levels and overall survival in patients with OSCC. There was no statistically significant correlation observed between *p14ARF* mRNA levels and overall survival. The median overall survival times for OSCC patients with either low or high *p14ARF* mRNA levels were quite similar (11.1 weeks or 6.7 weeks, respectively), with no significant difference found (p = 0.1397) (Figure 3.13B).

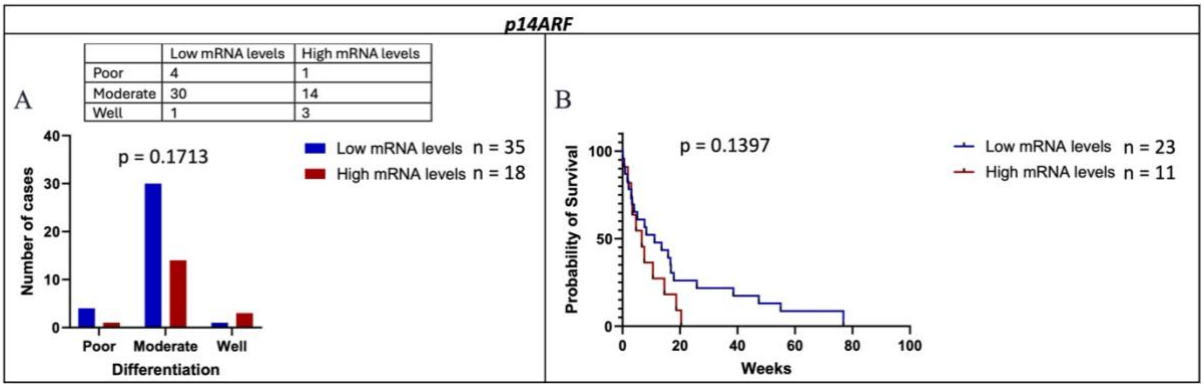


Figure 3.13 Association of *p14ARF* mRNA levels with tumour differentiation and survival of patients with OSCC.

Comparison of *p14ARF* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. While no significant associations were found between *p14ARF* mRNA levels and tumour differentiation (p = 0.1713), it is noteworthy that a high number of tumours (n=30) with low *p14ARF* mRNA levels were moderately differentiated compared to tumours with high *p14ARF* mRNA levels or tumours that were poorly or well-differentiated. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed using Kaplan-Meier curves for the correlation of overall survival and *p14ARF* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. There was no significant difference was observed in overall survival among OSCC patients with low or high *p14ARF* mRNA levels. The median overall survival was found to be 11.1 weeks for patients with low mRNA levels and 6.7 weeks for those with high mRNA levels, with a p-value of 0.1397.

The majority of tumours (n = 35) with low *p16INK4a* mRNA levels were more frequently found to be moderately differentiated tumours compared to those with high *p16INK4a* mRNA levels or tumours that were poorly or well-differentiated (Figure 3.14A). This trend is somewhat expected, given that a significant proportion of tumour samples in the cohort were moderately differentiated (54 out of 79) and most tumour samples expressed lower levels of *p16INK4a* mRNA (48 out of 79). The analysis did not reveal a significant association between *p16INK4a* mRNA levels and OSCC tumour differentiation (p = 0.3464) (Figure 3.14A).

Moreover, we investigated the relationship between *p16INK4a* mRNA levels and the overall survival in patients with OSCC. The median overall survival in OSCC patients with low or high *p16INK4a* mRNA levels was 7.6 weeks or 16.9 weeks, respectively (p = 0.7292) (Figure 3.14B). These findings suggest that while tumours with low *p16INK4a* mRNA levels tend to exhibit moderate differentiation and potentially lower overall survival rates, the observed trends did not achieve statistical significance in this study. Further studies with a larger sample size are necessary to explore the prognostic implications of *p16INK4a* mRNA levels in OSCC patients more comprehensively.

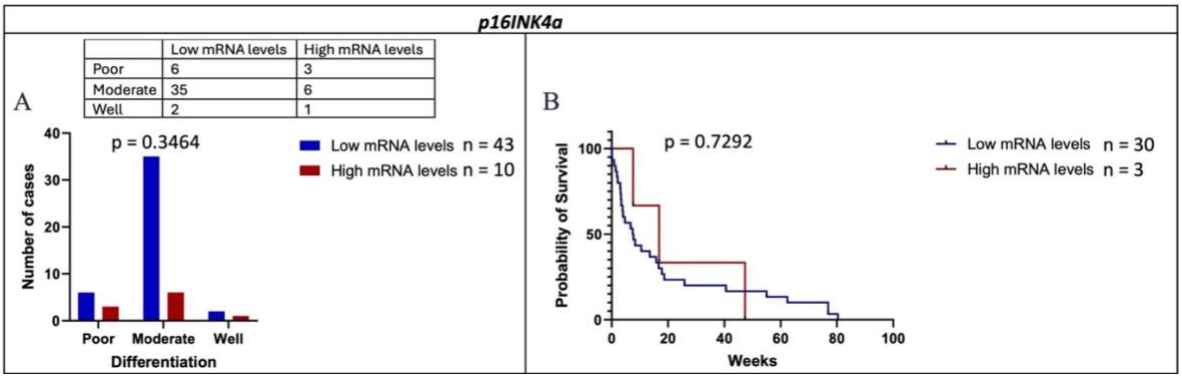


Figure 3.14 Association of *p16INK4a* mRNA levels with tumour differentiation and survival of patients with OSCC.

Comparison of *p16INK4a* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. While no significant association was found between *p16INK4a* mRNA levels and tumour differentiation, p = 0.3464, a high number of tumours (n=35) with low *p16INK4a* mRNA levels were moderately differentiated compared to those with high *p16INK4a* mRNA levels or tumours that were poorly or well-differentiated. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed using Kaplan-Meier curves for the correlation of overall survival and *p16INK4a* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. Decreased *p16INK4a* mRNA levels in tumours displayed a trend of a reduced overall survival rate in patients with OSCC, although this trend did not reach statistical significance. The median overall survival was found to be 7.6 weeks for patients with low mRNA levels and 16.9 weeks for those with high mRNA levels, with a p-value of 0.7292.

3.2.5.2 Correlation of *NFE2L2* mRNA levels with clinicopathological data

NFE2L2 mRNA levels were not associated with tumour differentiation status (p = 0.7416), or overall survival rate in patients (p = 0.7117) (Figure 3.15). The median overall survival was 7.4 weeks for individuals with low *NFE2L2* mRNA levels compared to 9.4 weeks for those with high *NFE2L2* mRNA levels (Figure 3.15B).

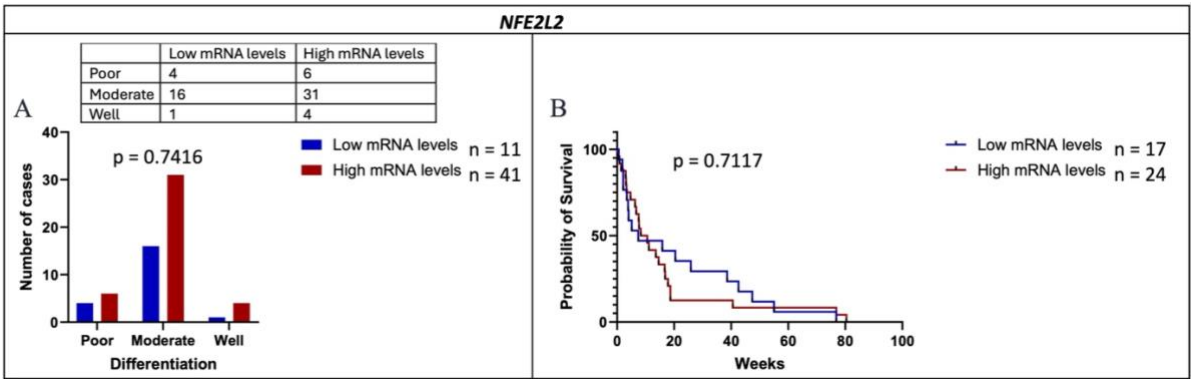


Figure 3.15 Association of *NFE2L2* mRNA levels with tumour differentiation and survival of patients with OSCC.

Comparison of *NFE2L2* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. There were no significant associations between *NFE2L2* mRNA levels and tumour differentiation, p = 0.7416. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed Kaplan-Meier curves for the correlation of overall survival and *NFE2L2* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. There was no significant difference regarding overall survival among OSCC patients with low or high *NFE2L2* mRNA levels. The median overall survival was found to be 7.4 weeks for patients with low mRNA levels and 9.4 weeks for those with high mRNA levels, with a p-value of 0.7117.

3.2.5.3 Correlation of *TOPBP1, ERCC6* and *C20orf196* mRNA levels with clinicopathological data

Subsequently, we explored the relationship between mRNA levels of DNA damage repair genes (*TOPBP1*, *ERCC6* and *C20orf196*) and clinicopathological features i.e. tumour differentiation, patients' survival status in patients with OSCC.

Although a significant proportion of tumours (n=43) with high *TOPBP1* mRNA levels were more frequently moderately differentiated compared to those with low *TOPBP1* mRNA levels or tumours that were poorly or well-differentiated, the analysis found no statistically significant association between *TOPBP1* expression and tumour differentiation status (p = 0.4768) (Figure 3.16A). This observation is expected, given a significant number of tumours in the cohort were moderately differentiated (54 out of 79), whereas the sample sizes for poorly and well-differentiated tumours were smaller, with only 10 out of 79 and 6 out of 79, respectively. Additionally, the majority of tumours exhibited higher levels of *TOPBP1* mRNA expression (77%).

Furthermore, the analysis of *TOPBP1* mRNA levels in relation to overall survival in OSCC patients did not reveal a statistically significant correlation. The median survival rates were 14.6 weeks for individuals with low *TOPBP1* mRNA levels versus 9.4 weeks for those with high *TOPBP1* mRNA levels, with a p-value of 0.7576 (Figure 3.16B).

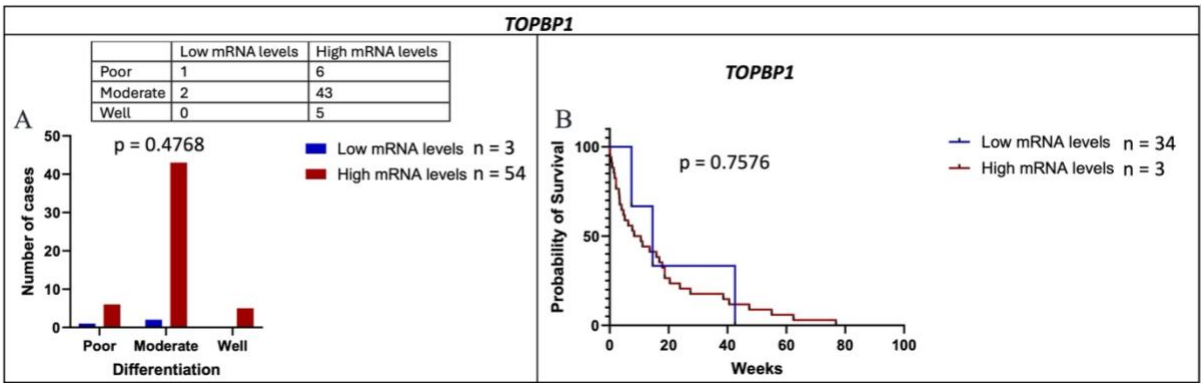


Figure 3.16 Association of *TOPBP1* mRNA levels with tumour differentiation and survival of patients with OSCC.

Comparison of *TOPBP1* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. There were no significant associations between *TOPBP1* mRNA levels and tumour differentiation, p = 0.4768. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed using Kaplan-Meier curves for the correlation of overall survival and *TOPBP1* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. *TOPBP1*mRNA levels was not significantly associated with overall survival rate in patients with OSCC. The median overall survival was found to be 14.6 weeks for patients with low mRNA levels and 9.4 weeks for those with high mRNA levels, with a p-value of 0.7576.

ERCC6 mRNA levels did not show a significant association with tumour differentiation (p = 0.2858) (Figure 3.17A). Furthermore, the analysis of *ERCC6* mRNA levels in relation to overall survival in OSCC patients did not reveal a statistically significant correlation. The median survival duration times were 11.00 weeks for individuals with low mRNA levels and 11.1 weeks for those with high mRNA levels, (p = 0.8259) (Figure 3.17B).

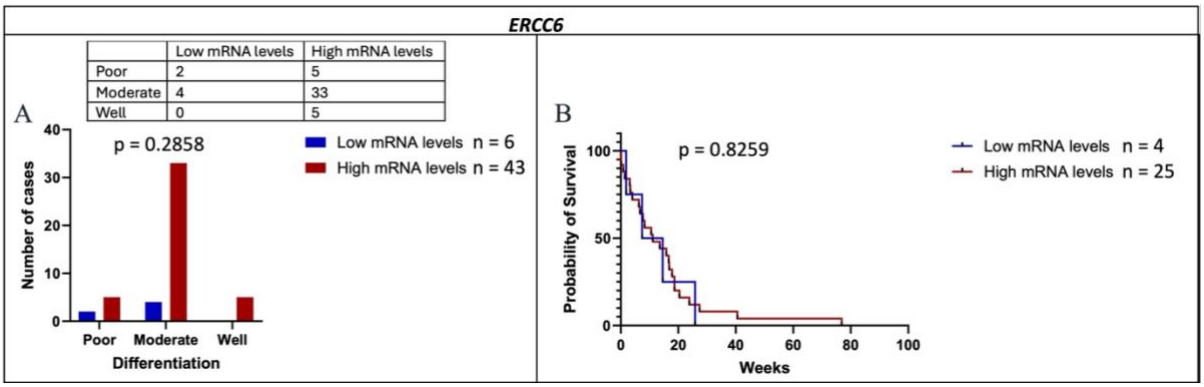


Figure 3.17 Association of *ERCC6* mRNA levels with tumour differentiation and survival of patients.

Comparison of *ERCC6* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. There were no significant associations between *ERCC6* mRNA levels and tumour differentiation, p = 0.2858. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed using Kaplan-Meier curves for the correlation of overall survival and *ERCC6* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. *ERCC6* mRNA levels was not significantly associated with overall survival rate in patients with OSCC. The median overall survival was found to be 11.0 weeks for patients with low mRNA levels and 11.1 weeks for those with high mRNA levels, with a p-value of 0.8259.

Altered *C20orf196* mRNA levels did not show a statistically significant association with tumour differentiation (p = 0.0560) (Figure 3.18A). Typically, a p-value of 0.05 or less is considered statistically significant. Given that our p-value of 0.0560 is greater than 0.05, our results suggest that altered *C20orf196* mRNA levels may not strongly correlate with tumour differentiation.

Additionally, there was no significant difference observed in overall survival between cases with low or high *C20orf196* mRNA levels. The median overall survival for OSCC patients with low *C20orf196* mRNA levels was 17.3 weeks, while for those with high *C20orf196* mRNA levels it was 9.4 weeks (p = 0.9725) (Figure 3.18B).

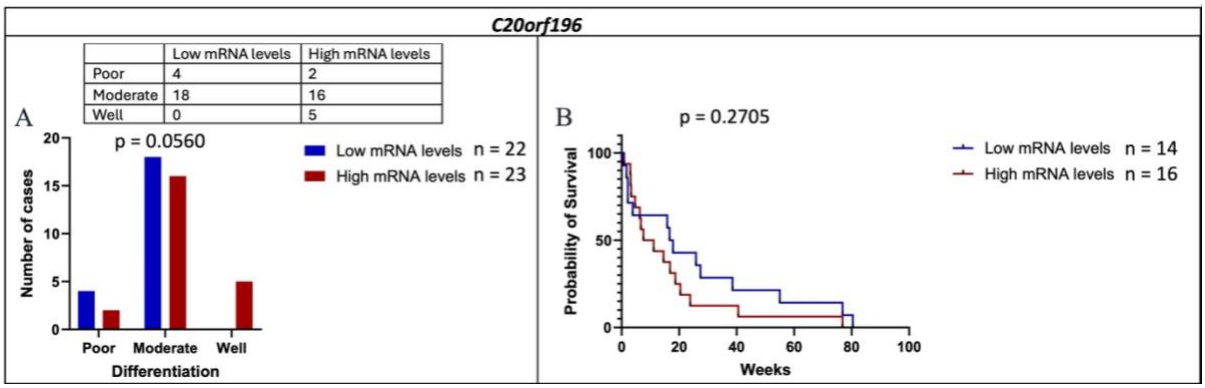


Figure 3.18 Association of *C20orf196* mRNA levels with tumour differentiation and survival of patients with OSCC.

Comparison of *C20orf196* mRNA levels in patients with OSCC with different **A**) Tumour differentiation (Using the Chi-squared test ($\chi 2$ test). The distribution of cases across different tumour grades (poor, moderate, and well) was analysed. Altered *C20orf196* expression was not associated with tumour differentiation, p = 0.0560. **B**) Overall survival rates (log-rank test). Overall survival rate was assessed using Kaplan-Meier curves for the correlation of overall survival and *C20orf196* mRNA levels, with expression classified as low and high mRNA levels. The log-rank test was employed to identify significant differences in survival rates. The analysis included only patients reported to have died. The data shows the time of survival (in weeks) among the patients. There was no significant difference between overall survival *C20orf196* mRNA levels in OSCC patients, the median overall survival was found to be 17.3 weeks for patients with low mRNA levels and 9.4 weeks for those with high mRNA levels, with a p-value of 0.2705.

The findings suggest that while there are trends in mRNA expression levels of *p14ARF*, *p16INK4a*, *TOPBP1*, *ERCC6*, and *C20orf196* in relation to tumour differentiation and survival, these genes may not independently serve as biomarkers for prognostication in OSCC. Furthermore, the limited representation of poorly and well-differentiated tumours in the cohort might have affected the ability to detect significant associations between mRNA levels and tumour differentiation. While the analysis provides valuable insights into gene expression patterns in OSCC, further studies should aim to expand sample sizes and possibly incorporate other biomarkers or clinical parameters to better understand the clinical implications and potential of *p14ARF*, *p16INK4a*, *TOPBP1*, *ERCC6*, and *C20orf196* as potential biomarkers in this context.

3.3 Discussion

Despite previous extensive research aimed at characterising of oesophageal cancer, much of the intricate underlying biology of OSCC remains poorly understood, especially in African OSCC patients. One of the major aims of this study was to identify genes potentially involved in the development of OSCC, which could be used to track disease progression or as therapeutic targets. Considering the mutated genes, driver genes, and affected signalling pathways identified in our cohort (chapter 2), this chapter examined mRNA levels of selected genes involved in cell cycle, *NFE2L2-KEAP* pathway and DNA damage repair pathway. Initially, mRNA levels were assessed in nine OSCC samples identified with *CDKN2A*, *NFE2L2*, *TOPBP1*, *ERCC6* and *C20orf196* mutations by WGS and WES. Subsequently, mRNA levels patterns were investigated in a larger cohort of 79 matched tumour-normal pairs by RT-qPCR. The smaller sample set was used to examine any expression trends in mutated tumour samples, followed by validation in a larger cohort. Notable expression trends were observed in tumour samples compared to adjacent normal tissues within our larger cohort. The differential expression observed in our cohort, alongside the mutation frequencies detected in these genes, suggests the presence of multiple regulatory processes influencing gene expression, operating alongside the identified truncating mutations and oncogenic missense mutations in our genes of interest. Additionally, an association between mRNA levels, tumour differentiation and overall survival rate in OSCC patients was investigated, to identify expression patterns as potential molecular markers of OSCC risk and, or prognosis prediction. Six genes: *p14ARF*, *p16INK4a*, *NFE2L2*, *TOPBP1*, *ERCC6* and *C20orf196* were investigated for gene expression patterns in OSCC. Our analysis revealed significant differential gene expression at mRNA level in OSCC. However, our analysis of overall survival showed no correlation between gene mRNA levels and survival rates.

Gene expression analysis in this study was conducted solely at the mRNA level, without parallel analysis at the protein level. While mRNA levels provide valuable insights into gene transcriptional activity and potential dysregulation [352-354], protein expression levels are also crucial for validating these findings and providing a direct link to functional activity within cells. The correlation between mRNA levels and the corresponding protein levels is an important consideration and has been extensively discussed in literature [354-359]. Emerging evidence suggests that while mRNA expression patterns are very useful, they are often insufficient for fully describing biological systems quantitatively [354]. This discrepancy is commonly attributed to additional levels of regulation between transcripts and their

corresponding protein products [352, 355, 357, 360]. These includes posttranscriptional mechanisms that control protein translation rates [361, 362], the half-lives of specific proteins or mRNAs [363, 364], delays in protein synthesis, and protein transport [355, 365]. Nonetheless, several studies have reported significant correlations between the mRNA and protein levels across different genes, tissues and cell lines [352, 355, 356, 358, 359]. Furthermore, genes with differentially expressed mRNA often exhibit higher correlations between mRNA and protein levels [356].

The *CDKN2A* genetic locus at human chromosome band 9p21 encodes two partially overlapping transcripts; *p14ARF* and *p16INK4a*, through alternative exon, with both being involved in the regulation of G1 to S phase progression [164-166, 169, 170, 196]. Given the prevalence of somatic mutations in *CDKN2A* gene in our cohort (chapter 2), we hypothesized that the alterations in *CDKN2A* alter *p14ARF* and *p16INK4a* mRNA levels in OSCC, potentially driving cellular transformation in the oesophageal epithelium. Our analysis revealed variable levels of *p16INK4*a and *p14ARF* mRNA in tumours compared to adjacent normal tissue. *p16INK4*a and *p14ARF* levels were significantly lower in 61% and 48% of tumour samples, respectively. Moreover, 33% of tumour samples exhibited extremely low to undetectable levels of both transcripts compared to adjacent normal controls. The differential expression of *p14ARF* and *p16INK4a* in cancer can be influenced by various factors, including genetic alterations, epigenetic modifications, and interactions with other signalling pathways [170, 174, 175, 197, 338]. In addition, the variable expression of *p16INK4a* and *p14ARF* observed in some OSCC patients may be attributable to truncating mutations and oncogenic missense mutations identified in *CDKN2A*. These mutations have the potential to influence the stability of mRNA for these genes.

Differential expression pattern of *p16INK4*a and *p14ARF* appear to be a common occurrence in OSCC, suggesting that p*16INK4*a and *p14ARF* may serve as putative driver genes for OSCC. This aligns with previous findings showing reduced or absent *p16INK4a* and/or *p14ARF* expression in OSCC samples, particularly in Asian populations.[174-176, 192]. De Almeida Simao and colleagues [192] reported variable expression of *p14ARF* and *p16INK4a*, with lower *p14ARF* and *p16INK4a* mRNA levels in 58.8% and 64.7% of the OSCC samples analysed, respectively. Similarly, Xing et al., [175] observed lower *p16INK4a* and *p14ARF* mRNAs in 12 of 18 (67%) Chinese OSCC samples, with 4 of the 18 (22%) samples maintaining elevated mRNA levels for both these genes. In terms of protein expression, 36 of 53 (67.9%) samples

from a study in Germany did not express or showed decreased expression of p16INK4a, as determined by immunohistochemistry [192]. Similarly, in French study, 15 out of 33 (46%) showed reduced expression of p16INK4a protein [193]. In Japanese samples, the proportion was notably higher, with 38 out of 42 (90.5%) lacking or exhibiting decreased levels of p16INK4a protein [194]. This consistent pattern across different populations underscores the potential significance of *p14ARF* and *p16INK4a* dysregulation in OSCC. Low expression of *p14ARF* and *p16INK4a* disrupts the regulation of cell cycle, increase susceptibility to oncogenic stimuli, thereby promoting tumour progression and influencing therapeutic responses in various cancer types [168, 170, 186, 195-199]. In addition, low or absence of p16INK4A protein adversely affects the initial treatment response of other cancer such as oral carcinoma [192]. Overall, our results suggest that *p16INK4a* and *p14ARF* transcript expression is decreased in OSCC by over 60 - 50%, respectively, similar to reports by others. Furthermore, reduced or absent *p14ARF p16INK4a* expression has been associated with more aggressive tumour behaviour and poorer prognosis in various cancers including OSCC [170, 184, 198, 199, 366-368]. Subsequently, we evaluated whether expression of *p16INK4a* and *p14ARF* were associated with tumour differentiation and the overall survival outcomes in OSCC patients, our findings revealed no statistically significance association between *p14ARF* and *p16INK4a* altered expression and tumour differentiation, nor overall survival rate in patients with OSCC.

Furthermore, combined *p16INK4*a and *p14ARF* mRNA levels were significantly elevated in tumour samples, including two samples identified with *CDKN2A* mutations. Additionally, we observed significantly elevated *p16INK4*a and *p14ARF* in 16% and 25% of tumours, respectively, compared to their normal adjacent tissues. Although the overexpression of *p16INK4a* in these cases cannot be attributed by the truncating mutation found in our cohort, previous studies have shown that the expression of both p14ARF and p16INK4a tends to increase in highly proliferative cells such as those affected by oncogenic stimuli, provided the genes are intact [316, 317]. In addition, overexpression of *p16INK4a* has also been observed in human papillomavirus (HPV)-related cancers including cervical cancer [197, 369, 370]. In our study, elevated levels of *p14ARF* mRNA were more frequently observed than those for *p16INK4a* in tumours compared to corresponding adjacent normal tissue, suggesting the presence of abnormal growth stimuli in these tissues, possibly triggering an antiproliferation response in these cells [175]. While our study and previous literature [175, 192, 193] align on mRNA expression profiles of selected genes, we acknowledge that mRNA levels may not

always correspond with protein activity. Therefore, further investigation into protein activity is crucial for understanding tumours dynamics and validating these findings.

Similar analysis were done for *NFE2L2*, revealing varied expression patterns of *NFE2L2* in OSCC tumour tissues compared to corresponding adjacent normal tissues. *NFE2L2* mRNA levels were significantly elevated in 56% tumours and significantly lower in 33% tumours. These differences in *NFE2L2* expression in OSCC could be attributed to different processes including genetic mutations or amplifications, thereby activating antioxidant response pathways in cancer cells, as previously reported [136, 221, 340, 371]. Several studies consistently found higher nuclear NFE2L2 protein expression in OSCC tumours, often accompanied by *NFE2L2* amplification and mutations within the DLG/ETGE motifs [136, 221, 340]. Jiang and colleagues [340] reported high expression of nuclear NFE2L2 protein associated with *NFE2L2* alterations, which have been consistently associated with poor prognosis in patients with OSCC [273, 340]. Wang et al., [371] reported fluctuating expression levels of nuclear NFE2L2 protein among patients were, indicating interindividual heterogeneity of tumours [371]. In contrast, Cui et al., [136] reported lower *NFE2L2* expression in mutant *NFE2L2* samples compared to wild-type samples. Similarly, in our previous chapters (Chapter 2) we reported frequent *NFE2L2* alterations clustered in either the DLG or ETGE motifs of NFE2L2. Our analysis of *NFE2L2* mRNA levels in nine tumour samples, including *NFE2L2*-mutated tumour samples (PD39449 [p.R34Q] and PD50653 [p.D77H]) releveled significantly lower *NFE2L2* mRNA levels compared to their normal adjacent tissues. This discrepancy with the larger sample set, which showed elevated *NFE2L2* mRNA levels in 56% of tumours, may reflect sampling error due to the smaller sample size. Previous studies have indicated that elevated levels of *NFE2L2* in OSCC has been correlated with poorer prognosis [221, 340, 371]. However, in our analysis, *NFE2L2* mRNA levels were not significantly associated with overall survival rate in patients with OSCC.

The mRNA levels and prognostic value of DNA repair genes: *TOPBP1, ERCC6* and *C20orf196* in OSCC was analysed, our analysis of *TOPBP1* mRNA levels in OSCC tumours revealed overexpression, including in a tumour with a *TOPBP1* mutation (PD39449 [p.F477I]), consistent with previous research in OSCC [342, 343]. Overexpression of *TOPBP1* has been shown to induce transformation with a TP53 mutation in non-tumorigenic breast epithelial cells [372]. Furthermore, immunohistochemical analyses conducted on biopsy samples from breast cancer patients revealed an upregulation of TOPBP1 in breast cancer tissues. This

overexpression of TOPBP1 was correlated with higher-stage tumours and diminished survival rates among patients [373-375]. Further reports suggest *TOPBP1* overexpression is associated with poor survival in patients with OSCC and may function as oncogene to promote the progression of OSCC [343]. Similarly, the DNA damage response pathways gene, *ERCC6* showed significant high mRNA levels in OSCC compare to adjacent normal tissues. However, tumour samples with *ERCC6* (PD39448 and PD39460 with [p.S1085C] and [p.E130*] mutations, respectively), showed decreased *ERCC6* mRNA levels when compared to normal adjacent tissue. Previous studies have reported mutations in *ERCC6* in OSCC, with more amplification events than deletions in NHEJ associated genes including *ERCC6* [331], potentially explaining the *ERCC6* expression patterns observed. Despites associations of *ERCC6* variants with increased risk of other cancers such as bladder cancer [376] and breast cancer [377], our results did not show statistically significant association between *ERCC6* expression and tumour differentiation or overall survival in OSCC patients. We also observed variable *C20orf196* expression in tumours, with 35% tumour samples had significantly elevated mRNA levels, including mutated tumour sample (PD39459 with the [p.S71T] mutation), and 29% tumour samples showing significantly low expression compared with corresponding normal tissues. Previous studies suggest depletion of C20orf196 protein impairs both NHEJ and class switch recombination (CSR) DNA repair mechanisms, while promoting homologous recombination [378]. Furthermore, altered *C20orf196* expression patterns did not significantly correlate with tumour differentiation or overall survival in patients with OSCC. Considering the mutation frequencies in DNA repair genes (*TOPBP1*, *C20orf196* and *ERCC6*) in our cohort and the significant altered expression in these genes, we suspect that there are multiple mechanisms likely to influence gene expression in these genes alongside identified missense.

In summary our study reveals altered expression patterns of key regulators involved in cell cycle control (*p14ARF* and *p16INK4a*), the *KEAP1-NFE2L2* pathway (*NFE2L2*), and DNA damage response pathways (*TOPBP1*, *C20orf196,* and *ERCC6)* in OSCC. There was no consistent pattern of *p14ARF*, *p16INK4a* or *C20orf196* expression in tumour compared to the adjacent normal tissue. However, the observed elevation in *TOPBP1* and *ERCC6* mRNA levels in OSCC suggests their potential involvement in OSCC development. The observed differential expression of *p16INK4*a and *p14ARF* in OSCC patients could be a consequence of the truncating mutations and oncogenic missense mutations found *CDKN2A* in addition to other factors that might influence gene expression such as epigenetic modifications, transcription

factors and cellular signalling. *NFE2L2* mutations might be associated with lower *NFE2L2* mRNA levels. However, we may need to assess the impact of the protein activity of the selected genes on the prognosis/survival of patients to better understand the role of these genes in OSCC.

Chapter 4
Investigation of p14ARF and p16INK4a role in OSCC

4.1 Introduction

The search for new target genes for cancer therapy holds promise for advancing cancer treatment and improving patient outcomes. Ultimately the task is to identify genes that may play a role in OSCC and the identification of specific clinical biomarkers and therapeutic targets [137]. OSCC is a heterogeneous disease with unclear molecular classifications, poor clinical outcomes, no reliable prognostic biomarkers and limited targeted therapies in OSCC patients [19, 136]. Defects in molecular and genetic mechanisms associated with OSCC in African patients are not well defined due, in part, to a lack of epidemiological and genetic studies [19, 137]. Hence, there is a need to identify cancer driver genes and to better understand the molecular biology for African OSCC and identify prognostic biomarkers specific to OSCC in African patients.

In chapter 2, we reported that *CDKN2A (p14ARF* and *p16INK4a*) mutations occurred in 35% and 10% (in WGS cohort and WES cohort, respectively) of the samples analysed in this study, suggesting that these two genes may be involved in the development of OSCC. Our pathway enrichment analysis (Chapter 2, sections 2.2.2.3 and 2.2.3.3) showed that cell-cycle regulating genes, including *p14ARF* and *p16INK4a* were involved in many of the significantly affected pathways in OSCC. In addition, our gene expression analysis (Chapter 3) using RT-qPCR in 79 OSCC tumours and their adjacent normal tissues has shown fluctuating levels of *p14ARF* and *p16INK4a* mRNA in OSCC tumours compared to their normal adjacent tissues. Interestingly, *p16INK4a* mRNA levels were significantly reduced or even absent in over 60% of the OSCC tumours, while *p14ARF* mRNA levels were significantly lower in 48% tumours, suggesting a potential dysregulation or suppression of the *p14ARF* and *p16INK4a* genes in OSCC.

Both *p14ARF* and *p16INK4a* play important roles in regulating cell growth and apoptosis through the p53 and RB pathways, respectively [166, 170]. p14ARF binds and promote the degradation of Mouse Double Minute-2 (MDM2) leading to stabilization and accumulation of p53, resulting in the expression of many apoptosis inducers and cell cycle inhibitory genes such

as p21 [167, 168]. On the other hand, *p16INK4a* is a member of the INK4 family of inhibitors of the cyclin-dependent kinase 4 (inhibitor of *CDK4*) [169, 333]. In response to specific signals, p16INK4a binds and inhibits the cyclin D-CDK4/CDK6 complex activity required for G1 to S cell cycle progression [379], thus inhibiting the phosphorylation of Rb by cyclin D-CDK4/CDK6 complex. In this way, Rb is maintained in the hypo-phosphorylated state (its growth-suppressive state) [380, 381] (Figure 4.1).

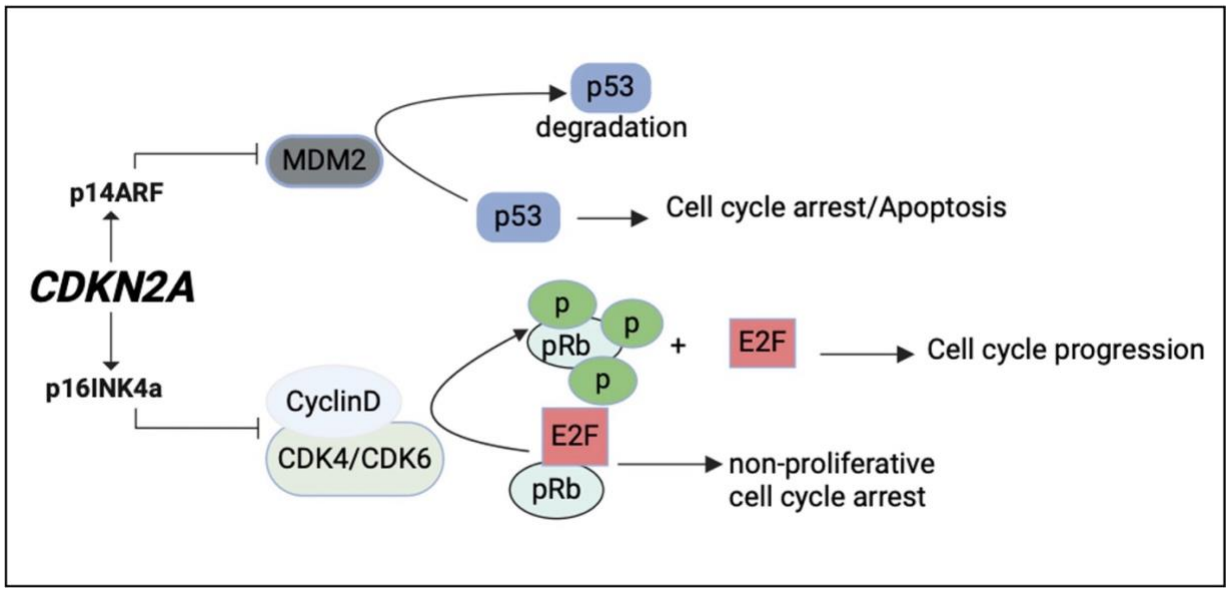


Figure 4.1 The role of *p14ARF* and *p16INK4a* in cells.

Pathways regulated by p161NK4a and p14ARF. p14ARF protein stabilizes p53 by promoting the degradation of MDM2. p16INK4a protein binds to the cyclinD-cyclin dependent kinase 4/6 (cyclinD-CDK4/6) complex and inhibits the phosphorylation of retinoblastoma protein, thus maintaining the Rb in its growth-suppressive mode and inhibits the activation of the transcription factor, E2F1, which induces cells to move from the G1 phase to S phase in the cell cycle. Figure was modified from Zhao et al., [189] and was created with BioRender.com.

The role of *p16INK4a* has been more extensively explored than *p14ARF* and plays an important role in carcinogenesis, particularly in cervical cancer, familial melanoma, pancreatic cancer-melanoma syndrome and head and neck squamous cell carcinoma [185-188]. In contrast, the role of *p14ARF* in human cancers is less well established [170]. The well-defined function of p14ARF is its interaction with and inhibition of MDM2, which target p53 for degradation thus potentiating p53 activity [163, 167]. In addition, *p14ARF* has a link between the two-tumour suppressor pathways; the Rb and the p53 pathways. It has been demonstrated that p14ARF triggers a p53-dependent checkpoint arrest when the Rb pathway is compromised. In particular, when inactivated, Rb causes the release of E2F transcription factors, which in turn induce increased *p14ARF* transcription (reviewed in [169]).

Previous studies have observed low expression levels of *p16INK4a* and/or *p14ARF* mRNA in OSCC [174, 175, 192]. De Almeida Simao and collaborators [192] reported absent or reduced *p14ARF* and *p16INK4a* mRNA levels in 58.8% and 64.7% of the OSCC samples analysed, respectively. In addition, Xing and others [175] have detected lower levels of *p16INK4a* and *p14ARF* mRNAs in 12 of 18 (67%) Chinese OSCC samples while 4 of the 18 (22%) samples maintained an elevated level of mRNA levels for both of these genes. Although these studies analysed the mRNA levels of *p14ARF*, *p16INK4a* in OSCC, there has been relatively few functional studies concerning the inactivation of *p16INK4a* and *p14ARF* genes in OSCC. In this study, we investigated the individual roles of *p14ARF* and *p16INK4a* in human oesophageal squamous cell carcinoma cells. Small interfering RNA (siRNA) targeting either *p14ARF* or *p16INK4a* were transfected into KYSE30 cells to knockdown the two mRNA transcripts and investigate the effects of the knockdown of *p14ARF* and *p16INK4a* on genes involved in cell cycle regulation, apoptosis, NFE2L2-KEAP pathway. Additionally, we conducted an in-silico analysis to investigate the impact of missense variants in *p14ARF* and *p16INK4a* on their respective protein structures.

4.2 Results

4.2.1 Screening for *CDKN2A* mutations in OSCC cell lines

Oesophageal squamous cell carcinoma cell lines were used to investigate the effects of *p14ARF* and *p16INK4a* knockdown. The presence exon 2 *CDKN2A* gene mutations that were detected in the WGS and WES in OSCC patients (Chapter 2) were screened for in the KYSE30, KYSE150, KYSE180, KYSE450, WHCO1, WHCO5, WHCO6 and EPC2 cell lines. Mutations detected in exon 2 of the *CDKN2A* gene in OSCC patients were examined in these cell lines are shown in table 4.1.

Table 4.1 List of *CDKN2A* exon 2 mutations screened in the oesophageal cell lines based on WGS and WES data in OSCC patients.

Gene	Pos (GRCh37)	Mutation (*p14ARF*)	Mutation (*P16INK4a)*	Exon
CDKN2A	g.21971186G>A	c.215C>T [p.P72L]	c.172C>T [p.R58*]	2
	g.21971155G>A	Silent	c.203C>T [p.A68V]	2
	g.21971120G>A	c.281C>T [p.P94L]	c.238C>T [p.R80*]	2
	g.21971113A>ACGGGTCGGGTGAGAGTGGCGGGGTCGGCGCAGTTGGGCTC	c.287_288insGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCG [p.A97Sfs*77]	c.244_245insGAGCCCAACTGCGCCGACCCCGCCACTCTCACCCGACCCG [p.V82fs*51]	2
	g.21971108C>T	c.293G>A [p.R98Q]	c.250G>A [p.D84N]	2
	g.21971036C>G	c.365G>C [p.R122P]	c.322G>C [p.D108H]	2
	g.21971036C>T	c.365G>A [p.R122Q]	c.322G>A [p.D108N]	2
	g.21971036C>A	c.365G>T [p.R122L]	c.322G>T [p.D108Y]	2
	g.21971029C>T	Silent	c.329G>A [p.W110*]	2
	g.21971001C>A	Outside coding region	c.358G>T [p.E120*]	2
	g.21970969A>G	Outside coding region	c.389T>C [p.L130P]	2

To screen for the mutations in exon 2 of the *CDKN2A* gene, a 410-bp PCR product (450-bp PCR product, if the 40-bp insertion is present) spanning intron 1–exon 2–intron 2 were amplified (amplifying the entire exome 2) by PCR using cell line genomic DNA (Figure 4.2A). PCR products were fractionated in 1% agarose gels and visualized by staining with Novel Juice (DNA staining reagent) to confirm successful amplification (Figure 4.2B). The 410-bp fragment of *CDKN2A* was successfully amplified in only EPC2, KYSE30, KYSE180, and WHCO5 cell lines, whereas fragments of *CDKN2A* were not amplified in KYSE150, KYSE450, WHCO1 and WHCO6, suggesting the loss of exon 2 (Figure 4.2B), these results were further explored and validated when we examined *p14ARF* and *p16INK4a* mRNA levels in the same cell lines (Section 4.2.2). As a control, all of the cell line DNAs successfully amplified other genes including *NFE2L2*, *TOPBP1*, *ERCC6* and *C20orf196* (data not shown), furthermore, the same *CDKN2A* primers were used to validate *CDKN2A* mutations and successfully amplified patients' DNA (Chapter 2). Samples that were successfully amplified were then subjected to DNA sequencing to examine the presence of the exon 2 mutations.

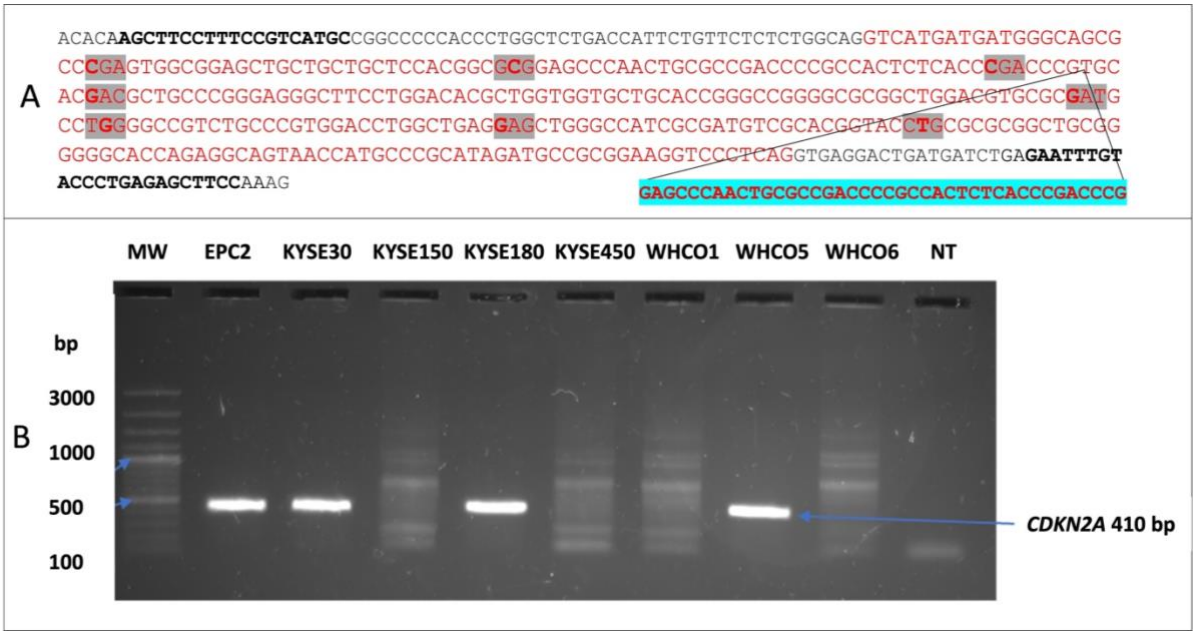


Figure 4.2 Primer design and PCR amplification of *CDKN2A* exon 2 mutations.

A) Sequence of the PCR amplified region encompassing *CDKN2A* exon 2 (red) flanked by *CDKN2A* intron 1 and 2 (black). Two amplicon sizes were expected: a wild-type amplicon of 410 bp and a mutant amplicon of 450 bp (with the 40-bp insertion). The 40-bp insertion is highlighted in blue. The eight codons mutated by single base substitutions are highlighted in grey. The mutated base is in bold. (Mutations are displayed in p16INK4a protein). The specific primer pairs used for *CDKN2A* mutational analysis are indicated in Materials and Methods Table 6.4. The positions of the two primers used for PCR amplification are shown in bold. **B**) Agarose gel analysis of PCR products of the indicated cell lines. Genomic DNA was extracted from the eight cell lines used for analysis (seven OSCC cell lines (KYSE30, KYSE150, KYSE180, KYSE450, WHCO1, WHCO5 and WHCO6 and a human telomerase immortalized normal oesophageal cell line; EPC2) as described in Material and Methods section 6.2.3.3. 50ng of genomic DNA was used in the PCR mixture with 5% DMSO. PCR reaction was initially denatured at 94°C for 10 min. Amplification consisted of 1 min each at 94°C, 58°C and 72°C for 35 cycles. PCR products were separated on a 1% agarose gel for 55 minutes. Of the eight cell lines used for the analysis, only EPC2, KYSE30, KYSE180, and WHCO5 cell lines were successfully amplified. The successfully amplified PCR products were subjected to DNA sequencing to examine the presence of *CDKN2A* exon 2 mutations.

Sequence analysis of the PCR products showed that the exon 2 mutations detected in patients were not present in EPC2, KYSE180 and WHCO5 cell lines, and no additional mutations were detected within the region of interest either (data not shown). In the KYSE30 cell line, we detected one (c.358G>T [p.E120*]) of the eleven *CDKN2A* exon 2 mutations, previously detected in our patients (Table 4.1). No additional mutations were detected in the region of interest (Figure 4.3).

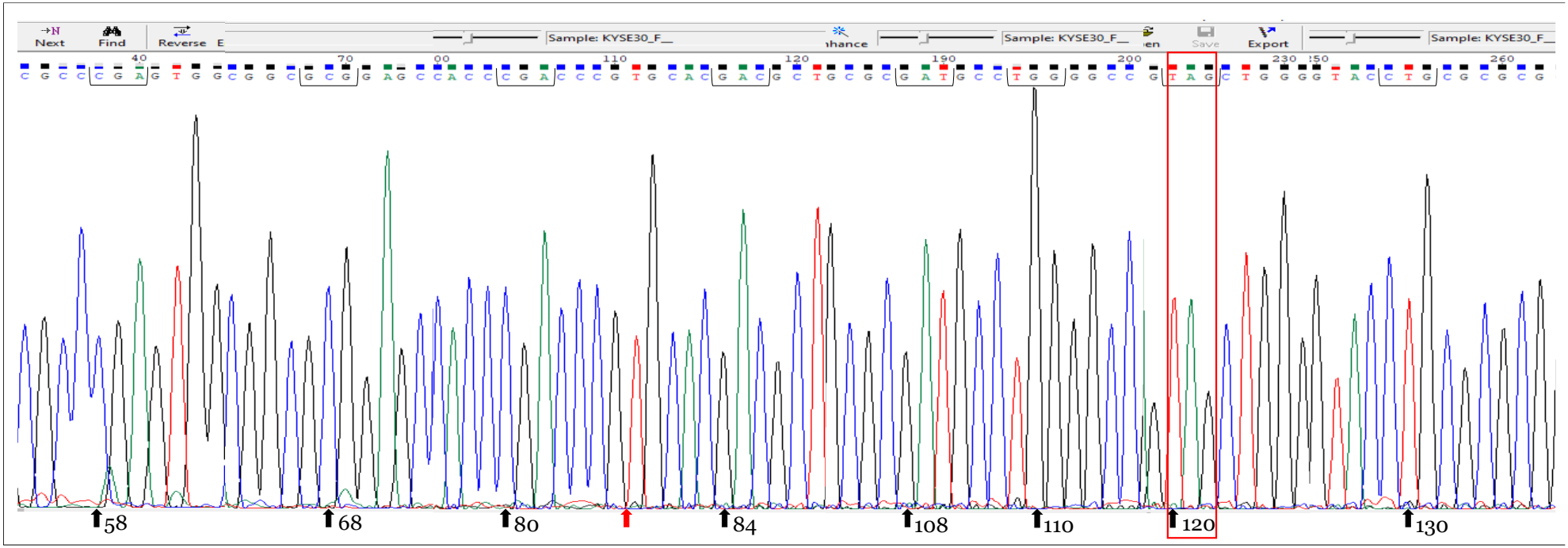


Figure 4.3 Mutations of *CDKN2A* exon 2 analysis in cell lines using PCR-Sanger sequencing.

Sanger sequencing chromatograms showing the mutated codons in exon 2 of *CDKN2A* gene (mutations are displayed in p16INK4a protein) previously observed in chapter 2. The position of nucleotide substitution is indicated by a black arrow. The codons mutated by single base substitutions are indicated with a bracket. (The red arrow indicates the 40-bp insertion position. Of the eleven mutation that were reported in (Table 4.1), we detected c.358G>T [p.E120*] mutation in the KYSE30 cell line, highlighted by a red box, changing glutamic acid into a stop codon at position 120 [p.E120*]. No additional mutations were detected in the region of interest.

4.2.2 mRNA levels of *p14ARF* and *p16INK4a* in OSCC cell lines

We subsequently assessed *p14ARF* and *p16INK4a* mRNA levels in seven OSCC cell lines (two containing wild-type exon 2 of *CDKN2A* (KYSE180 and WHCO5), based on sequencing analysis, four possibly lacking exon 2 of *CDKN2A*, as indicated by the absence of PCR products in Figure 4.2, based on PCR analysis (KYSE150, KYSE450, WHCO1 and WHCO6), and one containing a truncation mutation [p.E120*] (KYSE30)) and the control; non-tumorigenic EPC2 cell line (containing wild-type exon 2 of *CDKN2A*).

The *p14ARF* and *p16INK4a* mRNA levels in the cell lines were assessed using RT-qPCR analysis with primers specifically designed to detect either *p14ARF* or *p16INK4a* (as described in Figure 3.2). RT-qPCR amplification generated a single product by melting curve analysis (data not shown). RT-qPCR products were fractionated in 1% agarose gels to confirm successful amplification of targeted regions of *p14ARF* and *p16INK4a* in the cell lines analysed (Figure 4.4A).

The results showed differential expression patterns of *p14ARF* and *p16INK4a* (Figure 4.4). *p16INK4a* mRNA levels being significantly lower in all of the OSCC cell lines except KYSE30 cells when compared to the control cells (Figure 4.4B). In contrast, *p14ARF* mRNA levels were undetectable in five OSCC cell lines (KYSE150, KYSE180, KYSE450, WHCO1 and WHCO6), but was significantly higher in KYSE30 and WHCO5 cells (Figure 4.4B).

Of note, the g.21971001C>A variant in *CDKN2A*, which results in the c.358G>T [p.E120*] mutation in *p16INK4a* and is present in KYSE30, occurs outside the p14ARF coding region. This variant may explain the observed differences in the *p14ARF* and *p16INK4a* mRNA levels in KYSE30 cells (Figure 4.4). These results also suggest that the truncating mutation [p.E120*] in KYSE30 had no significant effects on the *p16INK4a* mRNA levels. This mutation could be located beyond the mRNA stability or transcriptional regulation regions. Premature stop codons can lead to nonsense-mediated mRNA decay, which typically results in mRNA degradation [382]. However, if the mutation occurs in a region where nonsense-mediated mRNA decay does not occur, or if the mRNA remains stable enough to sustain detectable levels, this could explain why *p16INK4a* mRNA levels are not significantly affected in KYSE30 cells.

No *p14ARF* mRNA levels and very low *p16INK4a* were detected in KYSE150, KYSE450, WHCO1 and WHCO6 cell lines, consistent with the absence of PCR products for these cell lines (Figure 4.2 and Figure 4.4A). These results support the hypothesis that KYSE150, KYSE450, WHCO1 and WHCO6 cells possibly lack exon 2 of *CDKN2A*. In contrast, we observed a significant difference in *p14ARF* and *p16INK4a* mRNA levels in WHCO5 cell line - *p14ARF* mRNA levels were significantly higher in WHCO5 cells, as opposed to the significantly lower *p16INK4a* mRNA levels in WCHO5 cells (Figure 4.4B). Overall, these findings

indicate differential expression of *p14ARF*, and *p16INK4a* mRNA levels in cell lines studied. Given that p16INK4A and p14ARF are transcribed from distinct promoters, this may contribute to the observed differences in mRNA expression levels of these genes. Additionally, the complex interplay of factors such as genetic alterations, post-transcriptional regulation, epigenetic modifications, and interactions with signalling pathways likely influences the variability in mRNA levels among these cell lines [170, 174, 175, 197, 338].

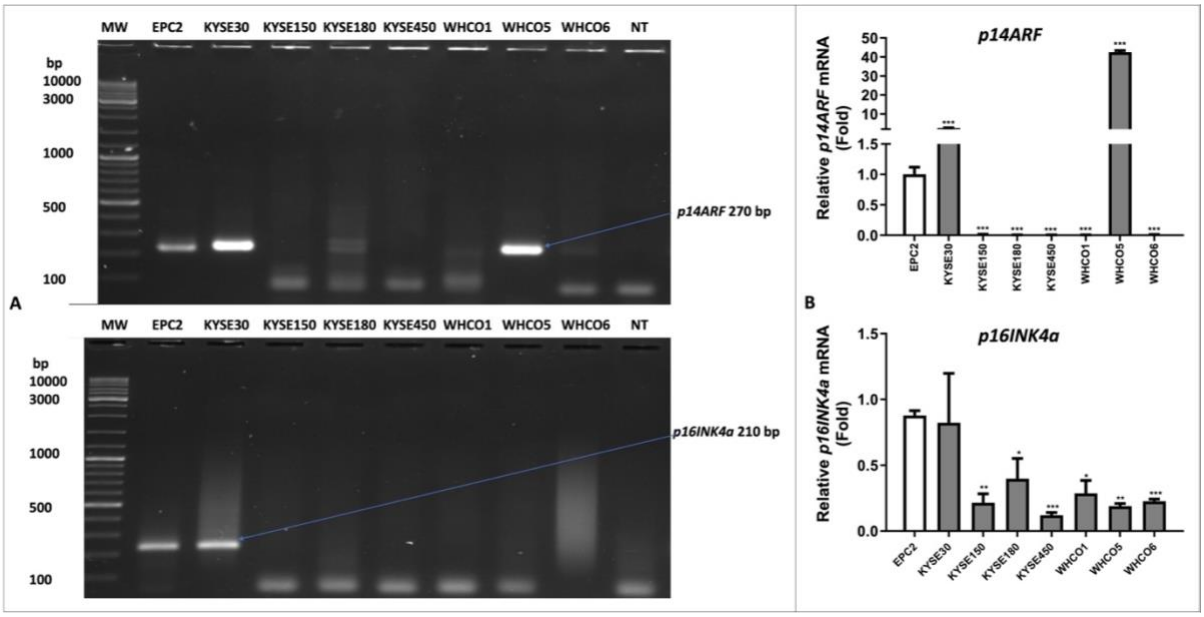


Figure 4.4 Analysis of *p14ARF* and *p16INK4a* mRNA levels in OSCC cell lines.

Analysis of *p14ARF* and *p16INK4a* mRNA levels in seven OSCC cell lines (KYSE30, KYSE150, KYSE180, KYSE450, WHCO1, WHCO5 and WHCO6) and an immortalized normal oesophageal squamous cell line (EPC2). **A**) A representative agarose gel electrophoresis shows the RT-qPCR analysis product, which was done to confirm whether our amplification is specific. **B**) The *p14AR*F and *p16INK4a* mRNA levels in seven OSCC cell lines showed lower *p14ARF* and *p16INK4a* mRNA levels in OSCC cell lines compared to the control cells. 1μg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. *p14ARF* and *p16INK4a* specific primers (Table 6.11) were used to detect *p14ARF* and *p16INK4a* mRNA levels. RT-qPCR primers were designed as described previously by Burri et al., [336], a 207-bp fragment with a forward primer in exon 1β for *p14ARF* amplification and a 210-bp fragment with a forward primer in exon 1α for *p16INK4a* amplification and a common reverse primer in exon 2. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

4.2.3 siRNA mediated *p14ARF* and *p16INK4a* knockdown

Given the observed *p14ARF* and *p16INK4a* mRNA levels in KYSE30 (Figure 4.4), and the fact that the truncating mutation in *p16INK4a* [p.E120*] showed no significant effects on *p16INK4a* mRNA levels, while in *p14ARF* it occurred outside the *p14ARF* coding region, KYSE30 cells were used as a model to investigate the effects of siRNA mediated knockdown of *p14ARF* and *p16INK4a* in oesophageal squamous cell carcinoma.

siRNA duplexes were designed based on the sequence of exon 1α of the *p16INK4a* locus, to target *p16INK4a* without affecting the partially overlapping *p14ARF* transcript encoded by exon 1β or designed in exon 1β of

p14ARF to target *p14ARF* transcripts without affecting *p16INK4a* expression [383]. *p14ARF* and *p16INK4a* siRNA duplexes were custom designed by us based on published data (Table 4.2) [195, 383]. The most effective siRNAs had symmetric two-nucleotide (UU or TT) 3´overhangs, which help to form RNA-induced silencing complexes (RISCs) with antisense and sense strands in equal ratio. The ideal target mRNA should contain an AA(N19)TT motif where N indicate the target sequence, and if no AA(N19)TT motif was found, NA(N19)TT or NA(N21) motifs could be used [384]. For our genes, no suitable matches could be found to the AA(N19)TT format, but we managed to find siRNA sequences corresponding to AA(N19) and NA(N19) for *p16INK4a* and *p14ARF*, respectively, with converted two nucleotides in 3'ends of the sense siRNA for *p16INK4a*, while for *p14ARF*, both sense and antisense siRNAs two nucleotides in 3'ends were converted to TT (Table 4.2).

Table 4.2 *p14ARF* and *p16INK4a* siRNA sequences.

Gene		siRNA sequence	Exon	Pattern
p14ARF	*mRNA sequence (5'-3')*	GAGGGTTTTCGTGGTTCACATCC	1β	NA(N19)
	siRNA sequence (Sense) (5'-3')	GGGUUUUCGUGGUUCACAU[dT][dT]		
	siRNA sequence (Antisense) (5'-3')	AUGUGAACCACGAAAACCC[dT][dT]		
p16INK4a	*mRNA sequence (5'-3')*	AACGCACCGAATAGTTACGGTCG	1α	AA(N19)
	siRNA sequence (Sense) (5'-3')	CGCACCGAAUAGUUACGGU[dT][dT]		
	siRNA sequence (Antisense) (5'-3')	ACCGUAACUAUUCGGUGCG[dT][dT]		

To confirm the effectiveness of the siRNA duplexes in reducing expression of the respective genes, KYSE30 cells were transiently transfected with 120 pmol of siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Effects of the siRNAs on *p14ARF* or *p16INK4a* mRNA levels were detected by RT-qPCR analysis. To confirm that the knockdown of either gene did not affect the expression of the partially overlapping second product of the *CDKN2A* locus, we also examined *p14ARF* or *p16INK4a* mRNA levels in cells transfected with either *p16INK4a*-directed or *p14ARF*-directed siRNA, respectively.

p14ARF and *p16INK4a* knockdown was validated at 24 and 48 hours post transfection. *p14ARF* mRNA levels were significantly reduced by 60% and 80% at 24 hours and 48 hours post transfection, respectively, and *p16INK4a* mRNA levels were significantly reduced by 60% both at 24 hours and 48 hours post transfection, compared to the untreated cells (Figure 4.5). Further, no significant effect was seen in *p16INK4*a expression following transfection of *p14ARF*-directed siRNA or in *p14ARF* expression following transfection of *p16INK4a*-directed siRNA (Figure 4.5).

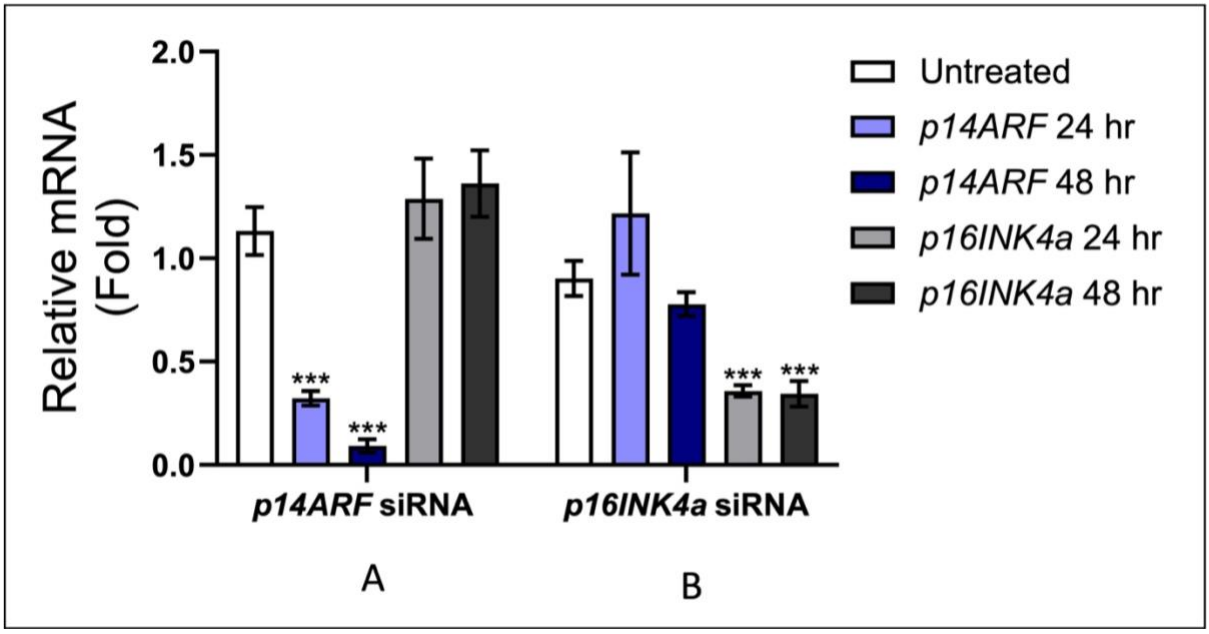


Figure 4.5 siRNA mediated *p14ARF* and *p16INK4a* knockdown in KYSE30 cells.

KYSE30 cells were transiently transfected with 120 pmol siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 24 and 48 hours post transfection to confirm knockdown of *p14ARF* and *p16INK4a*. **A**) Cells transfected with *p14ARF* siRNA. **B**) Cells transfected with *p16INK4a* siRNA. 1μg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. *p14ARF* and *p16INK4a* specific primers (Table 6.11) were used to detect *p14ARF* and *p16INK4a* mRNA levels. The siRNA significantly reduced *p14ARF* by 60% and 80% at 24 hours and 48 hours post transfection, respectively, and *p16INK4a* mRNA levels by 60% both at 24 hours and 48 hours post transfection compared to the untreated cells. The knockdown of either of the two genes did not significantly affect the mRNA levels of the other gene. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$. Three biological repeats were done.

Since the siRNA duplexes were proven to be effective tools for reducing *p14ARF* and *p16INK4a* mRNA levels, we next determined the biological implications of their inhibition on cell cycle regulators (especially genes involved in p14ARF/p53 and p16INK4a/Rb pathway), apoptotic and anti-apoptotic genes, *NFE2L2-KEAP* pathway regulators; *NFE2L2*, cell cycle and colony formation. In all further siRNA transfections, total RNA extracts were collected 48 hours after transfection.

4.2.3.1 The effects of *p14ARF* and *p16INK4a* knockdown on expression of cell cycle regulators

We examined the effects of *p14ARF* and *p16INK4a* knockdown on the expression of cell cycle regulators involved in the p14ARF/p53 and p16INK4a/Rb pathways.

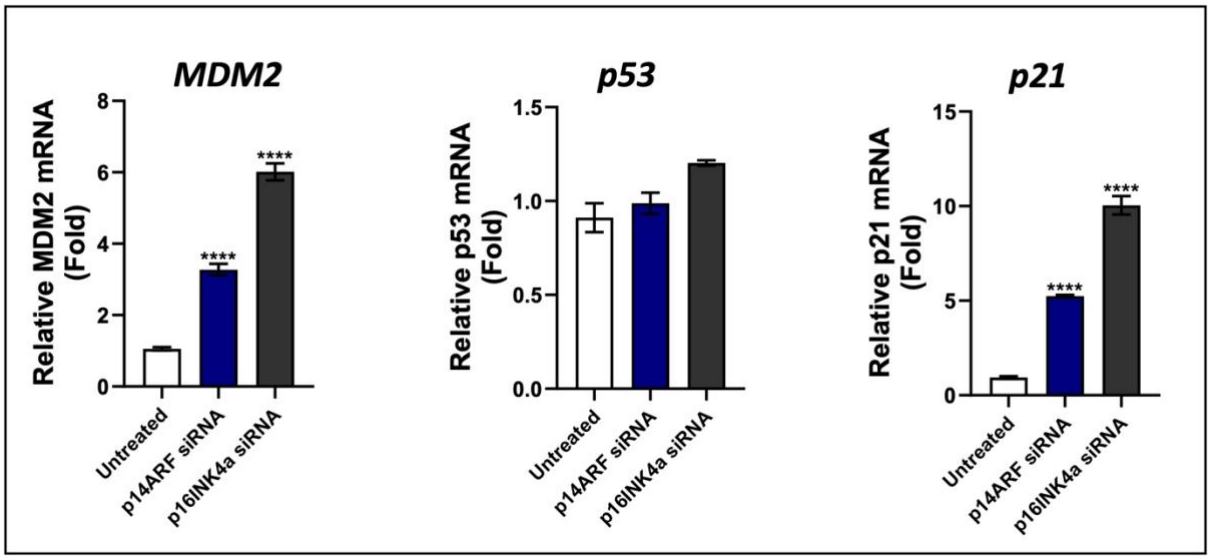


Figure 4.6 *p14ARF* and *p16INK4a* knockdown on p14ARF/p53 pathway regulators.

The effects of *p14ARF* and *p16INK4a* knockdown on the expression of p14ARF/p53 pathway regulators genes. KYSE30 cells were transiently transfected with 120 pmol siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 48 hours post transfection to examine the effects of *p14ARF* and *p16INK4a* knockdown on p14ARF/p53 pathway regulators; *MDM2*, *p53* and *p21* by RT-qPCR analysis. 1µg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. Primers used for RT-qPCR are listed in materials and methods (Table 6.11). siRNA-mediated depletion of *p14ARF* and *p16INK4a* mRNA in KYSE30 cells upregulated *MDM2* and *p21* expression compared to the untreated cells. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

The siRNA-mediated depletion of *p14ARF* and *p16INK4a* mRNA at 48 hours post transfection observed in KYSE30 cells in Figure 4.5 had no significant effect on p53 mRNA levels (Figure 4.6). However, mRNA levels of *MDM2* and *p21* were significantly elevated (Figure 4.6). While for p16INK4a/Rb pathway regulators, the knockdown of both *p14ARF* and *p16INK4a* mRNA levels downregulated *CCND1* mRNA levels, while *Rb* mRNA levels were significantly elevated in *p16INK4a* knockdown cells *(*Figure 4.7).

These results suggest that upregulation of *MDM2* may promote the rapid degradation of p53 and inhibit the ability of p53 to trans-activate target genes in KYSE30 cells. In contrast, upregulation of *p21* and *Rb*, as well as downregulation of *CCND1* may inhibit the activity of CDK-cyclin complexes, such as CDK4/6-CCND1. This inhibition prevents the phosphorylation of Rb protein, a key regulator of the G1 to S phase transition [168, 385].

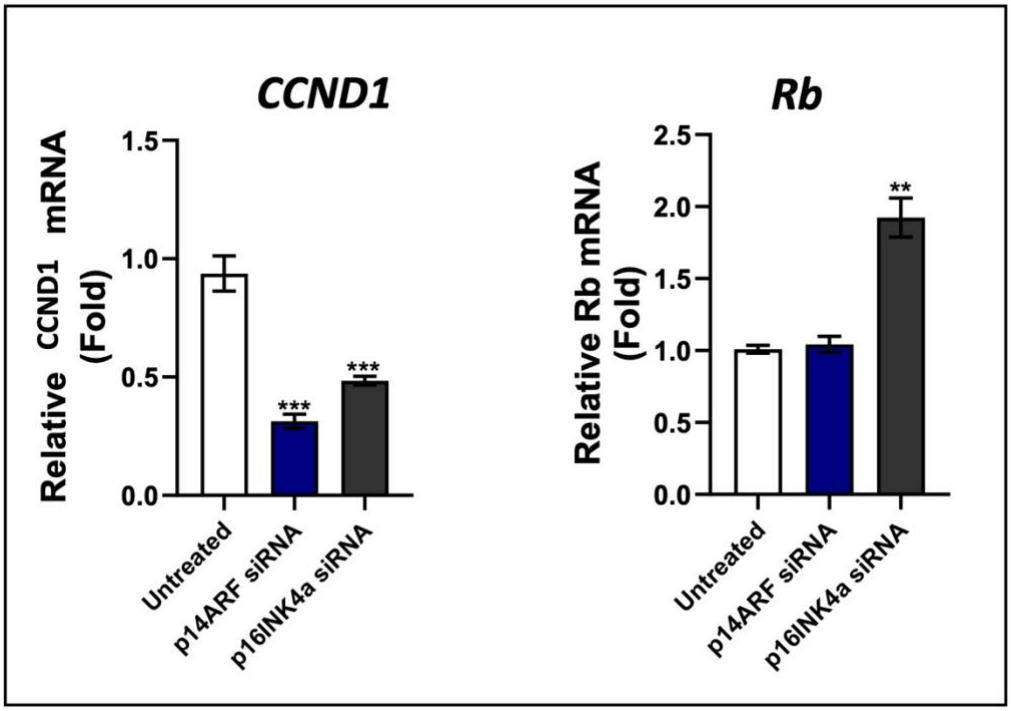


Figure 4.7 *p14ARF* and *p16INK4a* knockdown on p16INK4a/Rb pathway regulators.

The effects of *p14ARF* and *p16INK4a* knockdown on the expression of p16INK4a/Rb pathway regulators genes. KYSE30 cells were transiently transfected with 120 pmol of siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 48 hours post transfection to examine the effects of *p14ARF* and *p16INK4a* knockdown on p14ARF/p53 pathway regulators; *CCND1* and *Rb1* by RT-qPCR analysis. 1μg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. Primers used for RT-qPCR are listed in materials and methods (Table 6.11). siRNA-mediated depletion of p14ARF and p16INK4a mRNA in KYSE30 cells downregulated *CCND1* expression compared to the untreated cells, while *p16INK4a* knockdown upregulated *Rb* mRNA levels. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

4.2.3.2 The effects of *p14ARF* and *p16INK4a* knockdown on apoptotic and anti-apoptotic genes

Hanahan and Weinberg specified that one hallmark of cancer cells is their ability evade apoptosis [171, 348, 349]. In Chapter 2, we reported that *p14ARF* and *p16INK4a* play significant roles in "cellular responses to stimuli" pathways, particularly in response to oncogene and oxidative stress-induced senescence. Both p*14ARF* and *p16INK4a* trigger cell death or cell cycle arrest in response to oncogenic stress. Recent reports established that p14ARF induces cell death in a p53-independent manner (reviewed in [169]), as well as through the BAX protein independent of the p53 pathway [386, 387]. p14ARF expression induces of the pro-apoptotic multidomain protein BAX/BAK, leading to cell death via a Caspase-3-dependent pathway [386, 388]. However, the pro-apoptotic activity of BAX\BAK is primarily counteracted by the anti-apoptotic BCL2-related proteins such as BCL-XL, MCL1 and BCL2 [387], which block stress-induced apoptosis by preventing the release of cytochrome c into the cytoplasm of cells and subsequently inhibiting the activation of the caspase cascade (reviewed in [389]). The induction of mitochondrial apoptosis by p14ARF is facilitated by down-regulating the expression of anti-apoptotic proteins such MCL1 and BCL-XL [386]. Similarly, p16INK4a induces cell death by upregulating pro-apoptotic PUMA expression, activating Caspase-8 and repressing anti-apoptotic proteins such as MCL1 and BCL2 in p53-deficient leukaemia cells [390]. Currently, no studies have

examined the effects silencing of *p16INK4a* and *p14ARF* expression on apoptotic and anti-apoptotic genes in oesophageal squamous cell carcinoma. Therefore, we evaluated whether the knockdown of *p14ARF* and *p16INK4a* in KYSE30 cells modifies the expression of intrinsic apoptosis pathway members including pro-apoptotic genes (*BAX*, *Caspase-3*, and *Caspase-9*) and anti-apoptotic genes (*BCL-XL* and *BCL2*). Additionally, we also assessed the mRNA levels of *Caspase-3* and *Caspase-9*, as the effects of *p14ARF* and *p16INK4a* knockdown on the expression of these genes were not previously known.

The RT-qPCR analysis of apoptotic genes showed elevated *Caspase-3* and *Caspase-9* mRNA levels in *p14ARF* and *p16INK4a* knockdown cells. However, knocking down *p14ARF* and *p16INK4a* but did not alter *BAX* mRNA levels (Figure 4.8).

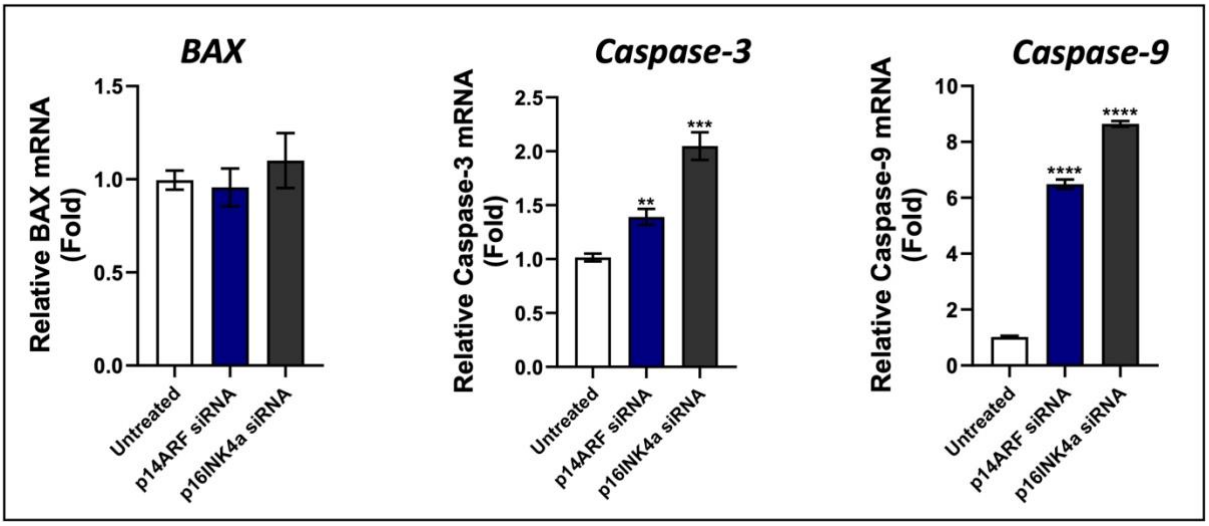


Figure 4.8 *p14ARF* and *p16INK4a* knockdown on apoptotic genes.

The effects of *p14ARF* and *p16INK4a* knockdown on the expression of apoptotic genes. KYSE30 cells were transiently transfected with 120 pmol of siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 48 hours post transfection to examine the effects of *p14ARF* and *p16INK4a* knockdown on apoptotic genes; *BAX*, *Caspase-3* and *Caspase-9* by RT-qPCR analysis. 1µg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. Primers used for RT-qPCR are listed in materials and methods (Table 6.11). siRNA-mediated depletion of *p14ARF* and *p16INK4a* mRNA in KYSE30 cells upregulated *Caspase-3* and *Caspase-9* mRNA levels compared to the untreated cells. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

Previous studies have demonstrated that over-expression of p14ARF and p16INK4a downregulates the expression of anti-apoptotic genes such as MCL1, BCL-XL and BCL2 in p53 protein-deficient cells, facilitating mitochondrial apoptosis induced by p14ARF or p16INK4a via BAK/BAX proteins in various cancer cell types [386, 388, 390]. Therefore, we investigated whether the silencing of *p14ARF* and *p16INK4a* alters the mRNA levels of anti-apoptotic genes *BCL-XL* and *BCL2* in KYSE30 cells.

Our RT-qPCR analysis revealed a slight increase in *BCL-XL* mRNA levels in *p14ARF* knockdown cells, while the depletion of p*14ARF* had no significant effect on *BCL2* mRNA levels. On the other hand, *p16INK4a* depletion led to a significant increase of *BCL-XL* and *BCL2* mRNA levels in KYSE30 cells (Figure 4.9). These results suggest that both *p14ARF* and *p16INK4a* exert effects through changes in the mRNA levels of the anti-apoptotic genes, with their depletion resulting in upregulation of these genes, albeit to different extents. In contrast, over-expression of p14ARF or p16INK4a down-regulate anti-apoptotic BCL2-related proteins to facilitate the induction of the Caspase-dependent apoptosis [386, 390], consistent with the results observed in this study.

Although knockdown of *p14ARF* and *p16INK4a* elevated transcription of pro-apoptotic *Caspase-3* and *Caspase-9* within 48 hours, the pro-apoptotic activity of BAX/BAK is counteracted by the expression of anti-apoptotic BCL2-related proteins such as BCL-XL and BCL2, which inhibit Caspase activation, thereby blocking the Caspase-dependent pathway of apoptosis [391, 392]. Consequently, overexpression of *BCL-XL* and *BCL2* in the same cells could contribute to evasion of apoptosis in KYSE30 cells by inhibiting Caspase activation and thereafter attenuating p16INK4a and p14ARF-induced apoptosis.

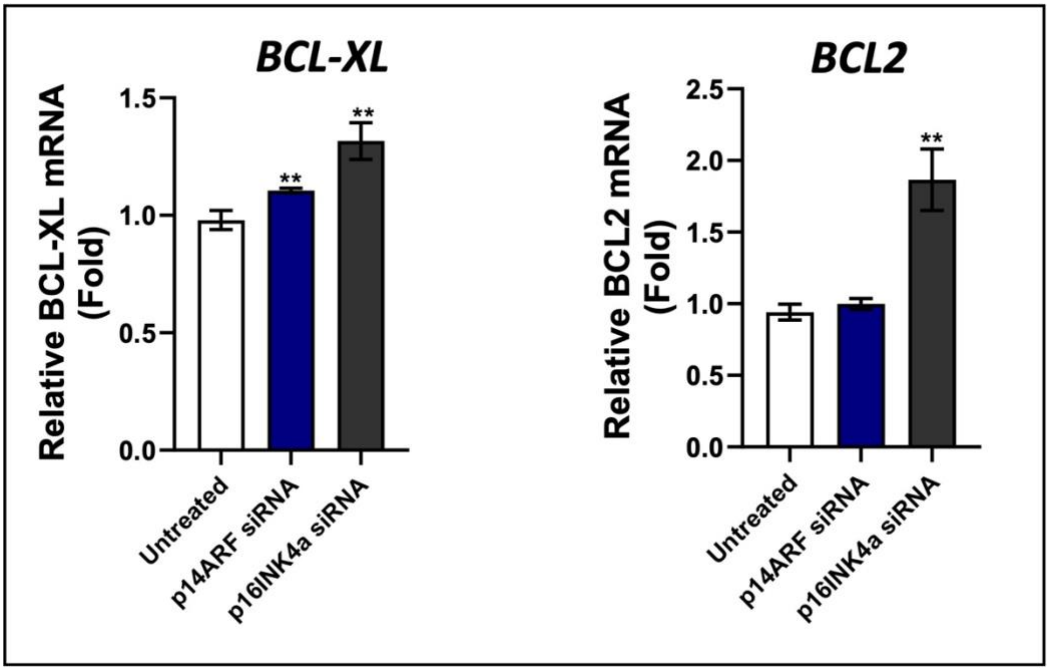


Figure 4.9 *p14ARF* and *p16INK4a* knockdown on anti-apoptotic genes.

The effects of *p14ARF* and *p16INK4a* knockdown on the expression of anti-apoptotic genes. KYSE30 cells were transiently transfected with 120 pmol of siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 48 hours post transfection to examine the effects of *p14ARF* and *p16INK4a* knockdown on anti-apoptotic genes; *BCL-XL* and *BCL2* by RT-qPCR analysis. 1µg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. Primers used for RT-qPCR are listed in materials and methods (Table 6.11). siRNA-mediated depletion of *p14ARF and p16INK4a* mRNA in KYSE30 cells upregulated *BCL-XL* mRNA levels compared to the untreated cells, while *p16INK4a* depletion significantly upregulated *BCL2* mRNA levels. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

4.2.3.3 The effects *of p14ARF* and *p16INK4a* knockdown on NFE2L2-KEAP pathway regulators

In addition to being identified as one of the driver genes in our cohort (Section 2.2.3.1), *NFE2L2* mRNA levels were elevated in over half of OSCC tissues compared to their normal adjacent tissues (Section 3.2.3), potentially leading to increased expression of *NFE2L2* target genes. p14ARF interacts with NFE2L2 and inhibits NFE2L2's ability to transcriptionally activate its target genes [393]. Moreover, NFE2L2 overexpression abrogates p14ARF's ability to induce p53-independent tumour growth suppression [393]. On the other hand, *p16IKN4a* is a target gene of *NFE2L2* and mutations or deletion of p16INK4a prevent its upregulation that would result from NFE2L2 activation [394]. In addition, OShimori and others reported *NFE2L2* involvement in cell cycle pathway, as it is associated with and stabilized by p21 resulting in enhanced oxidative stress reactive oxygen species protection and resistance to chemotherapeutic agents (258). Therefore, we aimed to examine the effects of siRNA-mediated knockdown of both *p14ARF* and *p16INK4a* on *NFE2L2* mRNA levels in KYSE30 cells.

We searched for the presence exon 2 *NFE2L2* gene mutations in our OSCC cell lines and the control, EPC2 cells seen in our patients by WGS and WES as shown in Table 4.3. None of exon 2 *NFE2L2* mutations were detected in any of the cell lines analysed, neither were there any additional mutations spanning the region of interest (data not shown).

Table 4.3 List of *NFE2L2* exon 2 mutations screened in oesophageal cell lines.

Gene	Mutation	Exon
NFE2L2	c.72G>C [p.W24C]	2
	c.85G>C [p.D29H]	2
	c.92G>C [p.G31A]	2
	c.91G>A [p.G31R]	2
	c.94_96delGTA [p.V32delV]	2
	c.101G>A [p.R34Q]	2
	c.112G>C [p.D38H]	2
	c.181_234del54 [p.Q61_E78delQLQKEQEKAFFAQLQLDE]	2
	c.229G>C [p.D77H]	2
	c.230A>G [p.D77G]	2
	c.235G>C [p.E79Q]	2
	c.241G>A [p.G81S]	2
	c.246A>T [p.E82D]	2
	c.1199C>T [p.S400F]	2
	c.1631G>C [p.G544A]	2

Analysis of *NFE2L2* mRNA levels in seven OSCC cell lines, and the control EPC2 cell line, by RT-qPCR showed significantly lower *NFE2L2* mRNA levels in all OSCC cell lines when compared to the control cells, although it is recognised that only one control cell was used in this study (Figure 4.10).

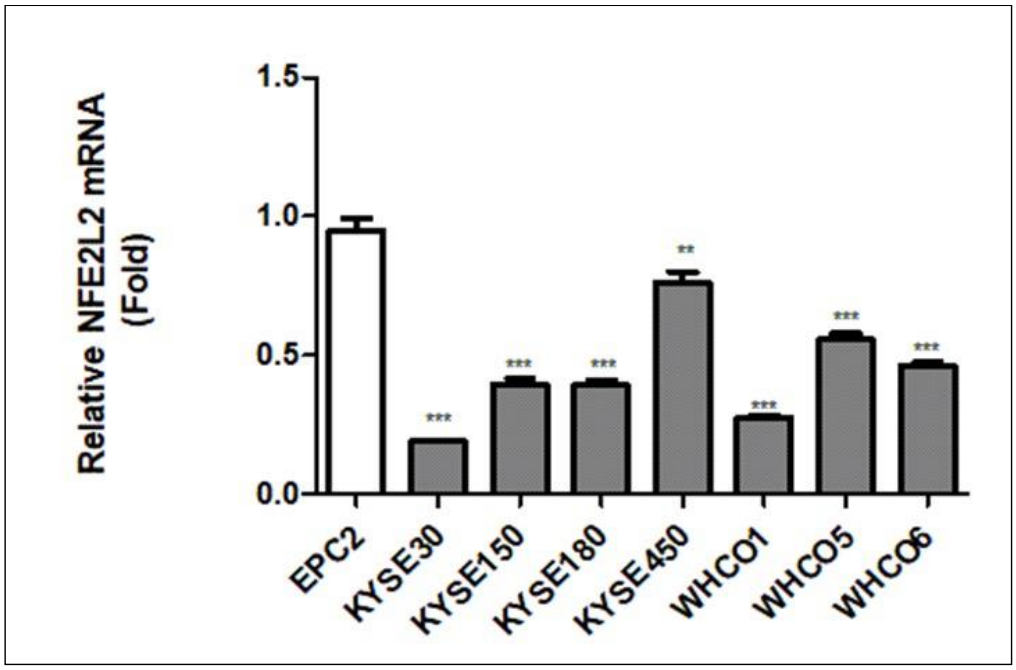


Figure 4.10 Analysis of *NFE2L2* mRNA levels in OSCC cell lines.

The *NFE2L2* levels in seven OSCC cell lines (KYSE30, KYSE150, KYSE180, KYSE450, WHCO1, WHCO5 and WHCO6) and an immortalized normal oesophageal squamous cell line (EPC2) were analysed by quantitative real-time PCR showed lower *NFE2L2* mRNA levels in OSCC cell lines compared to the normal control. 1µg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. *NFE2L2* specific primers (Table 6.11) were used to detect *NFE2L2* mRNA levels. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Previous studies have indicated that p14ARF has no obvious effect on NFE2L2 stability, but inhibits NFE2L2-mediated transcriptional activation of *NFE2L2* target genes [393]. However, siRNA-mediated depletion *p14ARF* and *p16INK4a* mRNA in KYSE30 cells resulted in significantly elevated *NFE2L2* mRNA levels (Figure 4.11), which could potentially lead to upregulation of NFE2L2 protein levels. The precise mechanism by which *p14ARF* and *p16INK4a* regulates *NFE2L2* requires further elucidation. It is noteworthy that overexpression of *NFE2L2* and activation of NFE2L2 is critical for tumour growth [393], and in various types of cancers, *NFE2L2* overexpression correlates with tumour progression, aggressiveness, and poor prognosis [395].

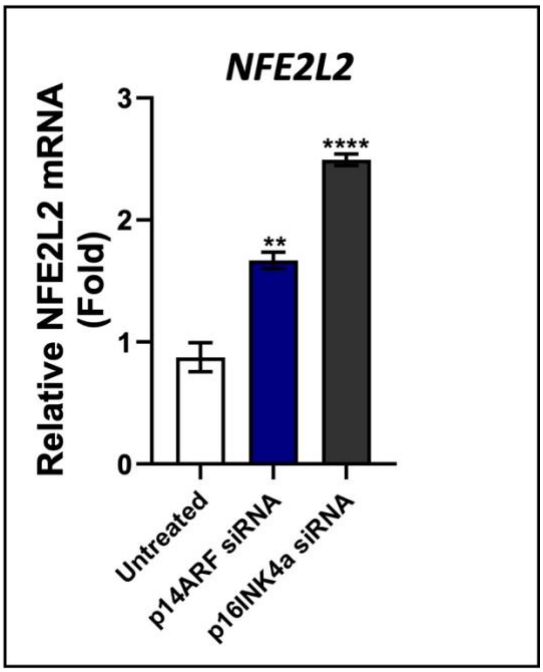


Figure 4.11 *p14ARF* and *p16INK4a* knockdown on *NFE2L2*.

The effects of *p14ARF* and *p16INK4a* knockdown on the expression of *NFE2L2*. KYSE30 cells were transiently transfected with 120 pmol of siRNA against either *p14ARF* or *p16INK4a* and the untreated cells were used as a negative control. Total RNA extracts were harvested from the cells 48 hours post transfection to examine the effects of *p14ARF* and *p16INK4a* knockdown on NFE2L2-KEAP pathway regulator; *NFE2L2* by RT-qPCR analysis. 1μg of total RNA was used as template for cDNA synthesis as described in Materials and Methods. Primers used for RT-qPCR are listed in materials and methods (Table 6.11). siRNA-mediated depletion of *p14ARF* and *p16INK4a* mRNA in KYSE30 cells significantly upregulated *NFE2L2* mRNA levels compared to the untreated cells. *GAPDH* was used for normalization. Error bars show standard deviation. Each bar represents the mean ± standard deviation (SD) of 3 replicates. Statistical analysis to determine significance difference of gene expression in OSCC cell lines and normal control was done using an unpaired t-test. ns $p > 0.05$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$, ****$p \leq 0.0001$.

4.2.4 *In silico* analysis of structural and functional consequences of *p14ARF* and *p16INK4a* missense variants

Besides detecting protein-truncating and indel mutations in *CDKN2A*, we have also found various missense variants in exon 2 of *CDKN2A* that potentially affect the functions of both p14ARF and p16INK4a (Table 4.4). However, the lethality and the functional implications of these missense mutations on the stability and function p14ARF and p16INK4a proteins remains poorly understood, especially at the structural level. To address this gap, we conducted *in silico* analysis to investigate how these missense mutations might affect the stability of these two proteins.

Table 4.4 List of *CDKN2A* exon 2 missense mutations in *p14RAF* and *p16INK4a*.

Pos (GRCh37)	Mutation (*p14ARF*)	Mutation (*P16INK4a)*
g.21971186G>A	c.215C>T [p.P72L]	Nonsense
g.21971155G>A	Silent	c.203C>T [p.A68V]
g.21971120G>A	c.281C>T [p.P94L]	Nonsense
g.21971108C>T	c.293G>A [p.R98Q]	c.250G>A [p.D84N]
g.21971036C>G	c.365G>C [p.R122P]	c.322G>C [p.D108H]
g.21971036C>T	c.365G>A [p.R122Q]	c.322G>A [p.D108N]
g.21971036C>A	c.365G>T [p.R122L]	c.322G>T [p.D108Y]
g.21970969A>G	Outside coding region	c.389T>C [p.L130P]

The first column denotes the genomic location of the substitution using the GRCh37 (Genome Reference Consortium Human Build 37). The second and third columns indicate the mutated codon, displaying both the alteration and the affected codon for p14ARF and p16INK4a, respectively. The analysis exclusively focused on examining the effects of missense mutations on the stability of these two proteins.

4.2.4.1 Structural analysis of p14ARF and p16INK4a missense mutations using UCSF Chimera tool

The wildtype Protein Data Bank (PDB) files for p14ARF (AF_AFQ8N726F1) and p16INK4a (1DC2) were downloaded from Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), (https://www.rcsb.org) [396]. Mutant PDB files for both p14ARF and p16INK4a were prepared similarly using AlphaFold2, together with ColabFold [397]. The protein sequences of both p14ARF (Uniprot ID - Q8N726) and p16INK4a (Uniprot ID - P42771) were downloaded as FASTA files from UniProt [398]. The wildtype amino acid was substituted with the mutant amino acid for all identified mutations in p14ARF and p16INK4a, predicting mutant PDB files for each mutation in p14ARF and p16INK4a. These sequences were used as input for colabfold_batch, a component of ColabFold [398], which implements protein folding with Alphafold2 models using MMseqs2 to generate the multiple sequence alignments (MSAs). The models were generated with the default parameters. For protein visualization and molecular structure analysis, we employed the UCSF Chimera tool [399].

4.2.4.1.1 p16INK4a

The p16INK4a protein consists of four relatively conserved ankyrin repeat motifs, with each repeat forming a helix-turn-helix motif (Figure 4.12) that align in an antiparallel manner, held together by hydrophobic interactions, creating helical bundles constituting the long arm of the ‘L’ shape. Each ankyrin repeat comprises approximately 31 to 34 residues, connected by three long loops that are exposed to solvent and fold back onto the helical region, forming β turns. Loop 1 spans from Ala-36 to Arg-46, loop 2 from Gly-67 to Arg-80, and loop 3 from Gly-101 to Leu-113 [400]. The axis of these loops is perpendicular to the helical axis, contributing to the formation of the L-shaped structure. Residues in the second and third ankyrin repeats, along with the

loop linking these repeats, are predominantly involved in interactions with CDK [401]. Whereas residues in the first and fourth ankyrin repeats, including the flexible N- and C-termini, primarily contribute to stabilizing the overall structure of p16INK4a, with a lesser involvement in its association with CDK4 [400, 401].

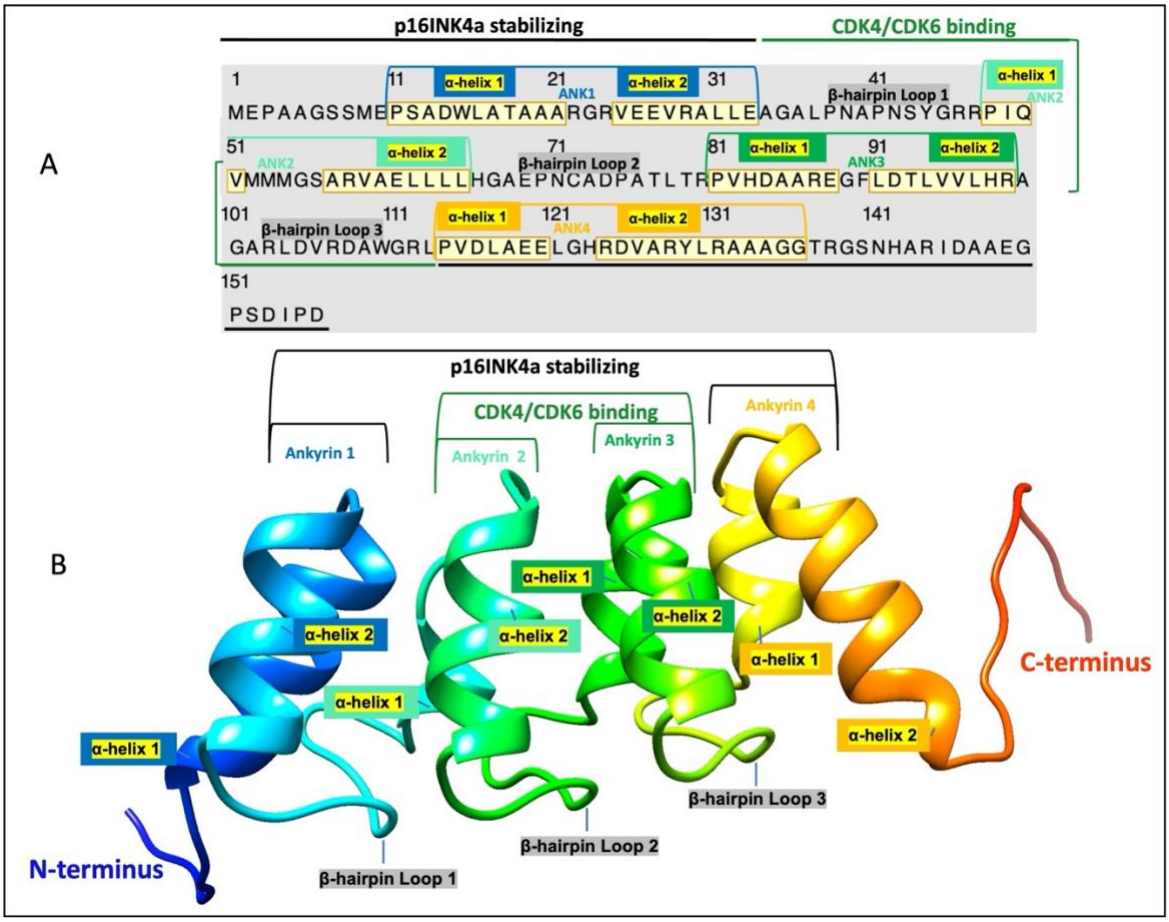


Figure 4.12 Sequence and domain structure of *p16INK4a*.

p16INK4a sequence and domain structure. **A**) The primary amino acid sequence of p16INK4a. In the sequence, regions of secondary structure (α-helices) are highlighted in yellow. **B**) The domain structure of p16INK4a. The p16INK4a contains four ankyrin repeats. The domains of p16INK4a are named as follow: ANK1 - Ankyrin repeat 1, ANK2 - Ankyrin repeat 2, ANK3 - Ankyrin repeat 3, and ANK4 - Ankyrin repeat 4. Each ankyrin repeat consists of a helix-turn-helix motif connected by three long loops. β-hairpin loop 1 (Ala-36 to Arg-46), β-hairpin loop 2 (Gly-67 to Arg-80), and β-hairpin loop 3 (Gly-101 to Leu-113). The secondary structure elements, connecting loops and the N-terminus and C-terminus are labelled. The wildtype PDB file for p16INK4a (1DC2) was downloaded from RCSB Protein Data Bank, (https://www.rcsb.org) [396]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

A network of hydrogen bonds connects the sidechains of multiple residues across the protein, accompanied by electrostatic interactions. Notably, most of the positively and negatively charged residues are positioned on the exterior of the molecule, while the majority of hydrophobic residues are predominantly packed within the interior (Figure 4.13A). The key hydrophobic residues at the core of the helix bundles play a crucial role in stabilizing the helix bundles structure [400]. Both these characteristics are important for the binding to CDK [400].

Binding and recognition of CDK are primarily facilitated by hydrogen-bond networks, involving several residues that are often mutated in cancer [401]. In our cohort, residues p.A68, p.D84, p.D108, and p.L130 have been identified with missense variants that are potentially damaging. These variants include p.A68V, p.D84N, p.D108H, p.D108N, p.D108Y, and p.L130P. The p.A68V variant is in the linking loop 2, between ankyrin repeat 2 and ankyrin repeat 3, while the p.D84N variant occurs within the ankyrin repeat 3. The p.D108H, p.D108N, and p.D108Y variants affect the same residue, located in the linking loop 3, between ankyrin repeat 3 and ankyrin repeat 4, and the p.L130P variant occurs within ankyrin repeat 4 (Figure 4.13B). Furthermore, all these residues form hydrogen bonds with both the backbone and the sidechains atoms of various residues within the protein (Figure 4.13B). These interactions are crucial for contributing to the folding, structural integrity, conformational stability, and overall function of p16INK4a [400].

A detailed comparison was done between the structures of the wild-type and mutant p16INK4a proteins, focusing on hydrogen bonds, electrostatic interactions, as well as contacts at these specific residues. The hydrogen bonds were limited to intra-model interactions, encompassing interactions within the molecule. Contacts comprised all types of direct interactions within the molecule, including both polar and nonpolar interactions, as well as favourable and unfavourable interactions. To detect contacts, negative cutoff values ranging from a distance of 0.0 to –1.0 Å with an allowance of 0.0 Å are generally reasonable; in our analysis, we utilized the default contact settings of –0.4 and 0.0 Å, respectively.

In both the wildtype and the p.A68V mutant p16INK4a proteins, a similar electrostatic interaction occurs between the same backbone atoms of the same residues (alanine 68/valine68 and leucine 63) (Figure 4.14A). However, while 13 contacts were found between wild-type A68 residue and surrounding atoms, there were 18 contacts in p.A68V mutant protein (Figure 4.14B). The larger side chain of the p.A68V mutant residue is more hydrophobic compared to the wild-type residue. This variation in size and hydrophobicity may disrupt the electrostatic interactions with the adjacent molecules [402]. Additionally, the introduction of the larger amino acid might cause bumps in protein structure, thereby increasing molecular interactions [402, 403], which could explain the observed increased contacts in the mutant protein, potentially impacting protein folding and leading to decreased stability of p16INK4a.

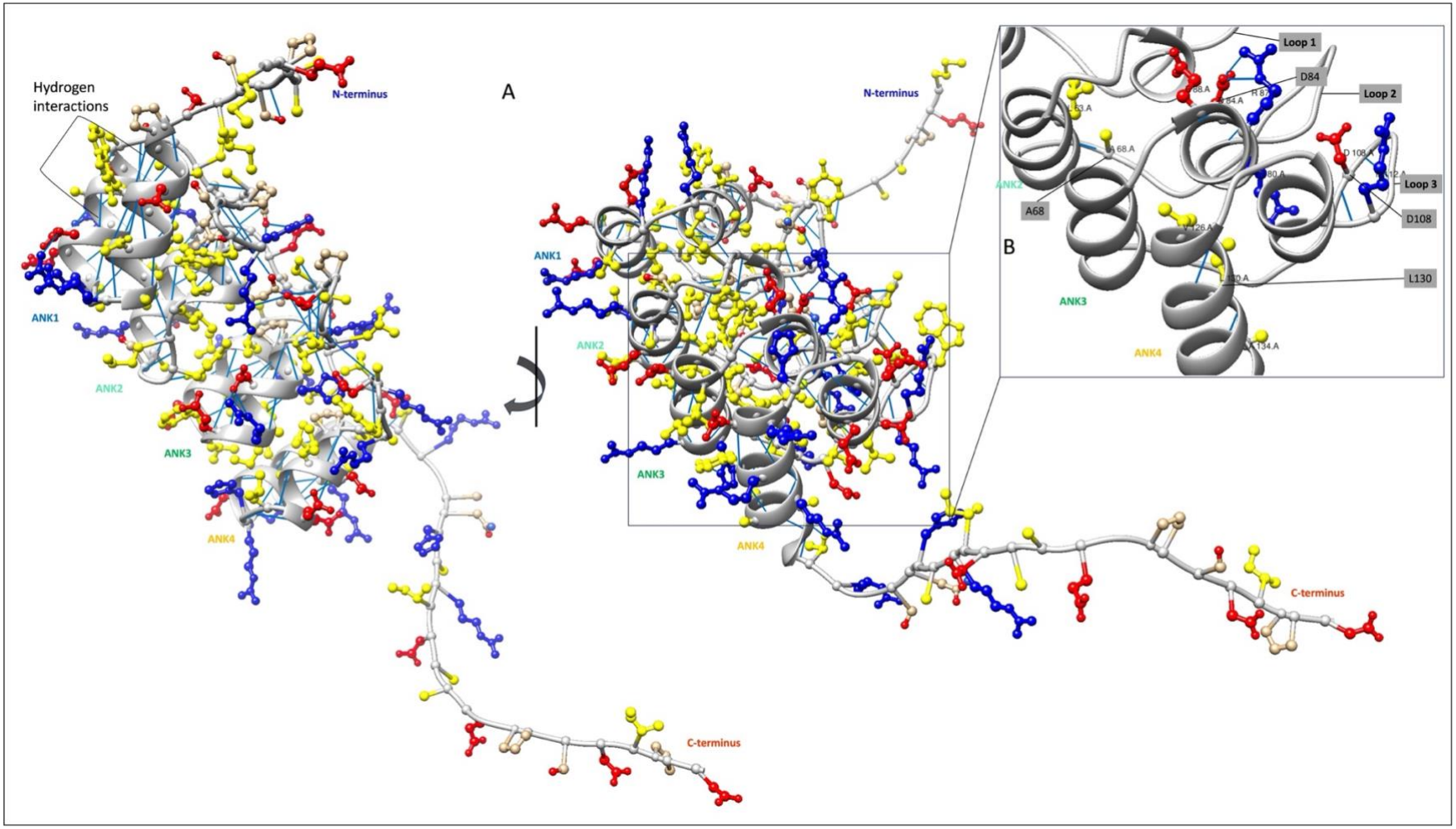


Figure 4.13 Molecular interaction pattern of wildtype p16INK4a.

The hydrogen bond interactions for wild-type p16INK4a protein. **A**) The hydrogen bond interactions within the p16INK4a protein, highlighting hydrophobicity and electrostatic interactions. Positively and negatively charged residues are marked in blue and red, respectively, while the hydrophobic residues are denoted in yellow. The backbone structure is depicted in grey. Notably, most of the charged residues are situated on the exterior of the molecule, while most of the hydrophobic residues are predominantly packed within the interior of the molecule. Hydrogen bonds are represented in blue and defined by solid lines, with the interacting protein residues represented in ball-and-stick format. **B**) Focus on specific residues (p.A68, p.D84, p.D108 and p.L130) identified with missense variants in exon 2 of p16INK4a. The highlighted hydrogen bond interactions formed between these residues and other within the molecule, crucial for folding and stability of p16INK4a. Both donor and acceptor atoms are shown. Residues of interest are labelled in grey. The wildtype PDB file for p16INK4a (1DC2) was downloaded from RCSB Protein Data Bank, (https://www.rcsb.org) [396]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

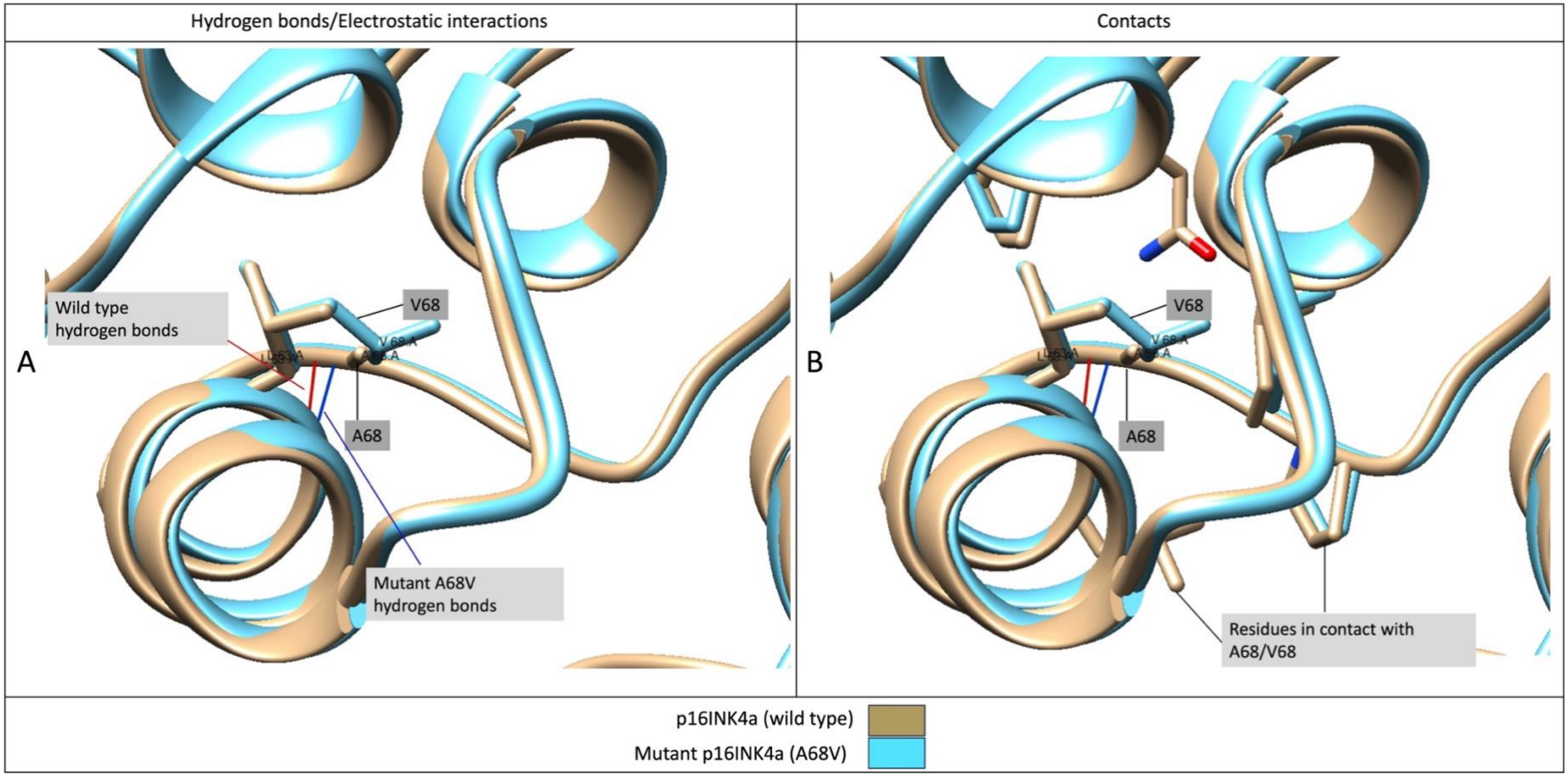


Figure 4.14 Structural analysis of the p.A68V variant on the structure of p16INK4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the A68 residue forms a salt bridge (indicated in red) between the backbone atoms of alanine 68 and leucine 63. Similarly, in the p.A68V variant, a comparable electrostatic interaction (indicated in blue) occurs between the same backbone atoms of the same residues (valine 68 and leucine 63). The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a proteins. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a.The introduction of the valine amino acid at position 68 leads to increased contacts in the mutant compared to the wild-type. Residues in contact within the wild-type are depicted in brown, while residues in contact within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

In the wild-type p16INK4a structure, at position 84 (aspartic acid 84), there are four salt bridges are observed: between aspartic acid 84 (donor) and arginine 80 (acceptor), two involving arginine 87 (donor) and aspartic acid 84 (acceptor) occurring between different atoms in the side chains of the two residues, and glutamic acid 88 (donor) and aspartic acid 84 (acceptor) (Figure 4.15A). However, the introduction of the uncharged asparagine amino acid at position 84 (D84N) results in the elimination of two of the four salt bridges found in the wild-type (specifically, two of the arginine 87 and aspartic acid 84 hydrogen bonds). Additionally, this mutation leads to the formation of a new salt bridge between the side chain atoms of asparagine 84 and the backbone atoms of arginine 80 in the mutant protein (Figure 4.15A). Additionally, there is a decrease in contacts in the mutant compared to the wild-type, from 18 in the wild-type to 14 contacts in the mutant protein (Figure 4.15B).

The p.D84N variant occurs within the ankyrin repeat 3; with residues in the second and third ankyrin repeats participating in hydrogen-bond interaction with CDK4 and CDK6 [401]. Wild-type aspartic acid 84 is a charged residue compared to the uncharged mutant asparagine 84, and most of these charged residues cluster on the surface of p16INK4a (Figure 4.13). These electrostatic interactions significantly contribute to the interaction between p16INK4a and CDK4 [400]. Alterations in the electrostatic interactions within these regions affect the protein's ability to interact with other molecules, such as cyclin-dependent kinases, thereby inhibiting the p16INK4a activity. Furthermore, structural changes in p16INK4a due to various mutants such as D84H, G101W, and H123Q have led to decreased or undetectable p16INK4a activity [400], potentially disrupting the regulation of cell cycle, increasing susceptibility to oncogenic stimuli, and promoting tumour progression [169, 401].

.

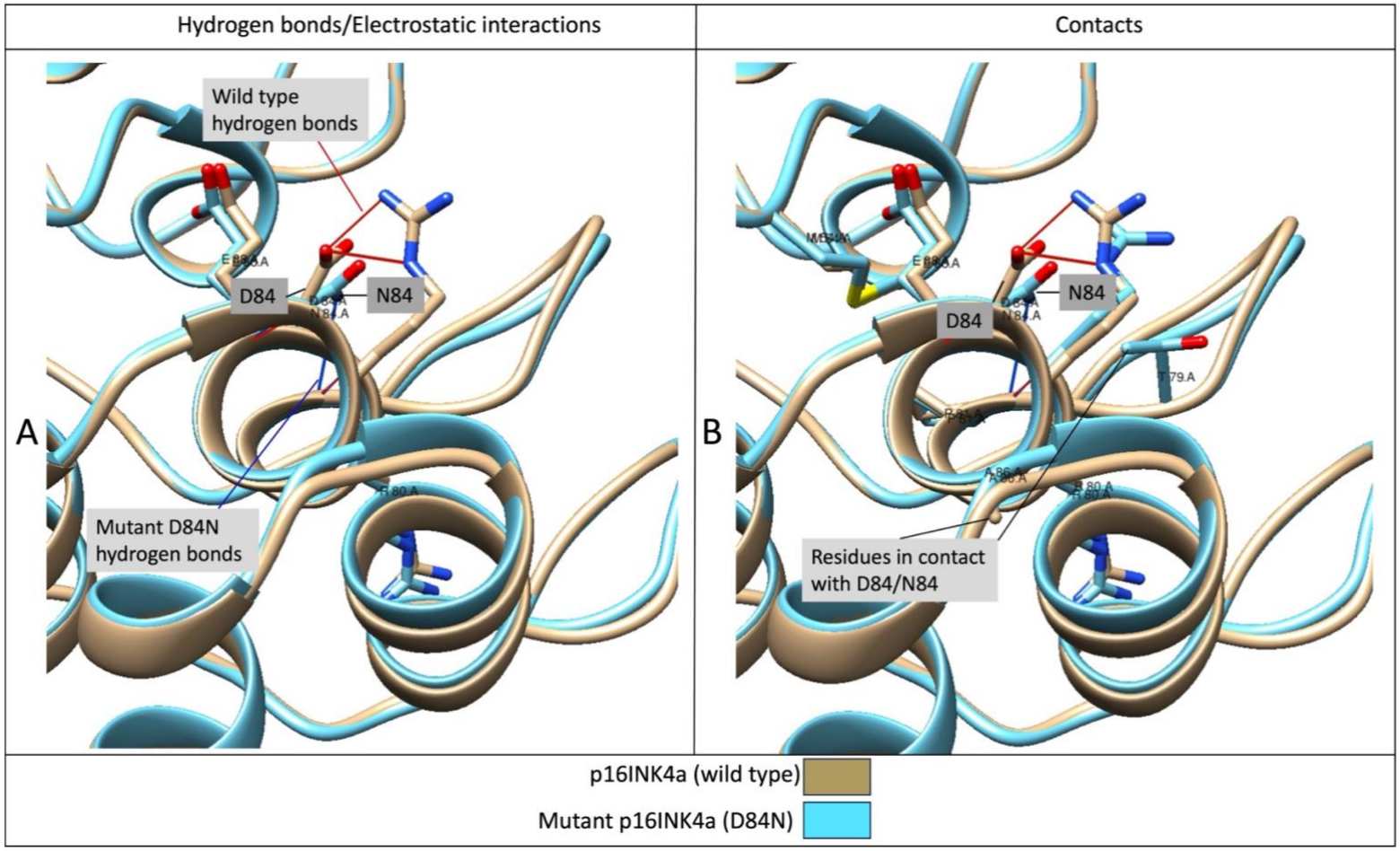


Figure 4.15 Structural analysis of the p.D84N variant on the structure of p16INK4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the D84 residue forms four salt bridges (indicated in red), two between the backbone atoms of aspartic acid 84 with arginine 80, and glutamic acid 88 backbone atoms, and two between the side chain atoms of aspartic acid 84 with side chain atoms of arginine 87. In contrast, the p.D84N variant led to the elimination of two of the four salt bridges found in the wild-type, and the formation of a new salt bridge between the side chain atoms of asparagine 84 and the backbone atoms of arginine 80. The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a protein. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a. The introduction of the asparagine amino acid at position 84 leads to decreased contacts in the mutant compared to the wild-type. Residues in contact within the wild-type are depicted in brown, while residues in contact within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

Two salt bridges are observed in the wild-type p16INK4a between the backbone atoms of aspartic acid 108 (D108) with arginine 112, as well as glycine 111 backbone atoms (Figure 4.16A). Similarly, in all three variants affecting this residue, namely p.D108H, p.D108N, and p.D108Y, comparable electrostatic interactions were observed between the backbone atoms of the respective residues (histidine, asparagine and tyrosine) with arginine 112, and glycine 111 backbone atoms (Figure 4.16A, Figure 4.17A, Figure 4.18A). However, the p.D108N variant leads to the formation of new salt bridges between the side chain atoms of asparagine 108 (N108) and the backbone atoms of arginine 112, and tryptophan 110, (Figure 4.17A), potentially contributing to the altered folding and conformational instability of p16INK4a. The p.D108Y variant replaced a negatively charged native residue, aspartic acid with a hydrophobic residue, tyrosine, (Figure 4.18A) which could disrupt electrostatic interactions with surrounding molecules and possibly compromise the hydrophilic function on the surface of p16INK4a [402, 403].

Twenty-two contacts were detected between atoms at D108 residue and neighbouring residues (Figure 4.16B). In contrast, in all three mutants, p.D108H, p.D108N, and p.D108Y, the observed contacts increased compared to those in the wild-type residues: 45 contacts, 25 contacts, and 36 contacts in p.D108H, p.D108N, and p.D108, respectively (Figure 4.16B, Figure 4.17B, Figure 4.18B). Furthermore, the alterations introduced larger side chain amino acids compared to the wild-type residue in all three mutants. This variation in size could lead to bumps in protein structure, thereby increasing molecular interactions [402, 403], which could explain the observed increased contacts in the mutant protein.

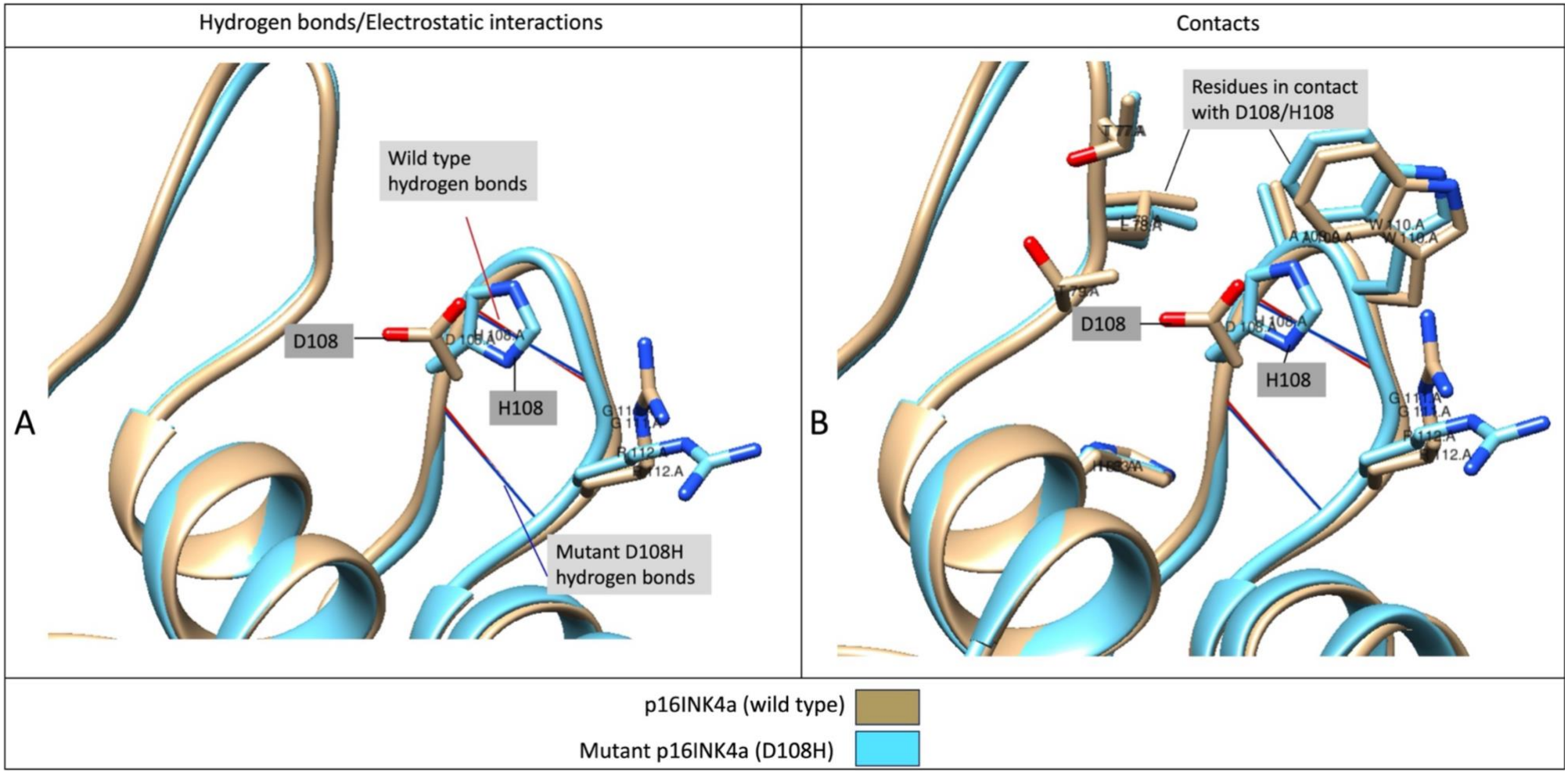


Figure 4.16 Structural analysis of the p.D108H variant on the structure of p16INK4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the D108 residue forms two salt bridges (indicated in red) between the backbone atoms of aspartic acid 108 with arginine 112, and glycine 111 backbone atoms. Similarly, in the p.D108H variant, same electrostatic interactions (indicated in blue) occur between the same backbone atoms of the same residues (histidine 108 with arginine 112, and glycine 111). The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a protein. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a. The introduction of the histidine amino acid at position 108 leads to increased contacts in the mutant compared to the wild-type. Residues in contacts within the wild-type are depicted in brown, while residues in contacts within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

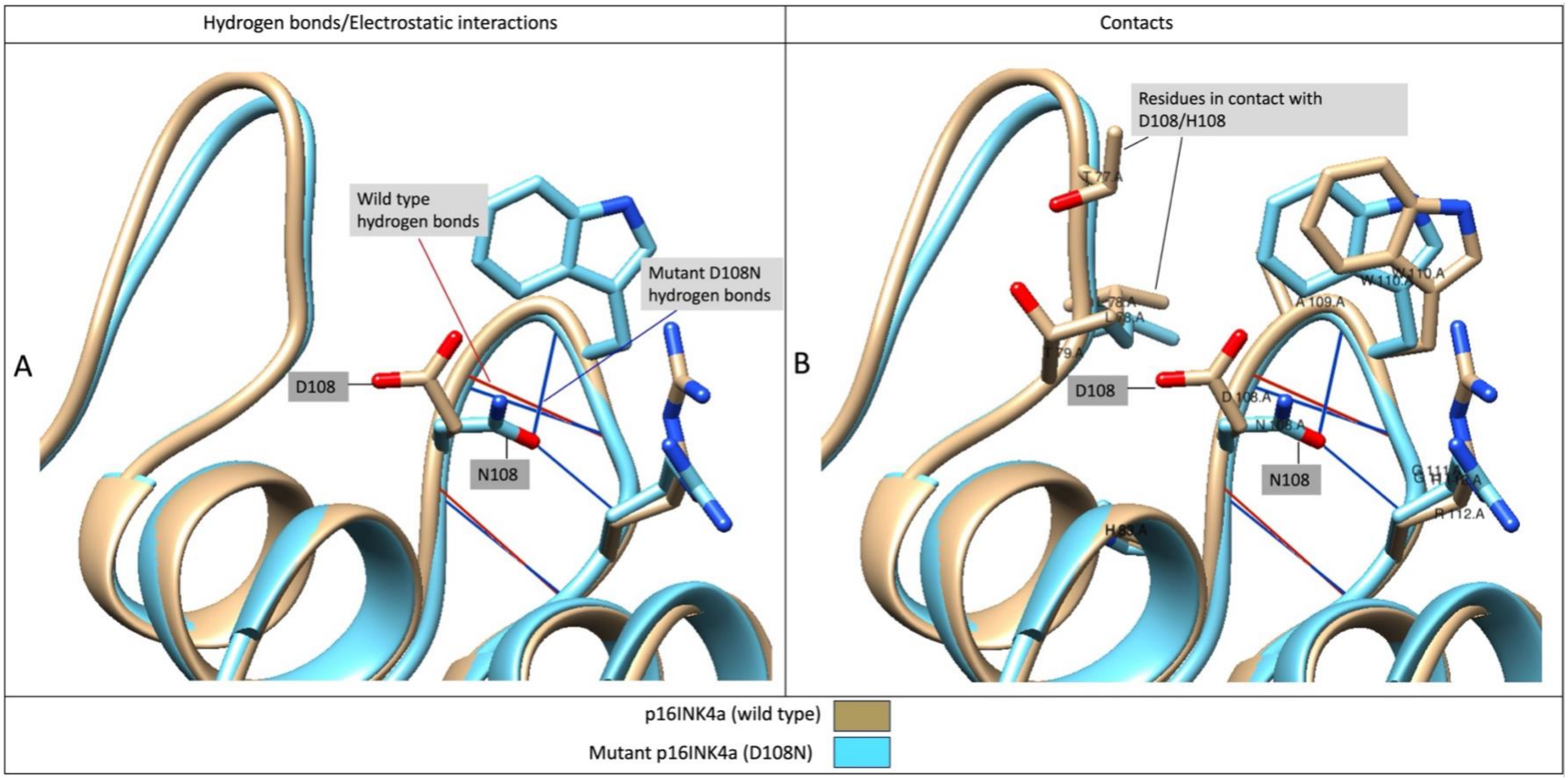


Figure 4.17 Structural analysis of the p.D108N variant on the structure of p16INk4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the D108 residue forms two salt bridges (indicated in red) between the backbone atoms of aspartic acid 108 with arginine 112, and glycine 111 backbone atoms. In the p.D108N variant, in addition to similar electrostatic interactions (indicated in blue) occurring between the same backbone atoms of the same residues found in the wild-type protein (asparagine 108 with arginine 112, and glycine 111 backbone atoms), the p.D108N variant leads to the formation of new salt bridges between the side chain atoms of asparagine 108 and the backbone atoms of arginine 112, and tryptophan 110. The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a protein. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a. The introduction of the asparagine 108 amino acid leads to increased contacts in the mutant compared to the wild-type. Residues in contact within the wild-type are depicted in brown, while residues in contact within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

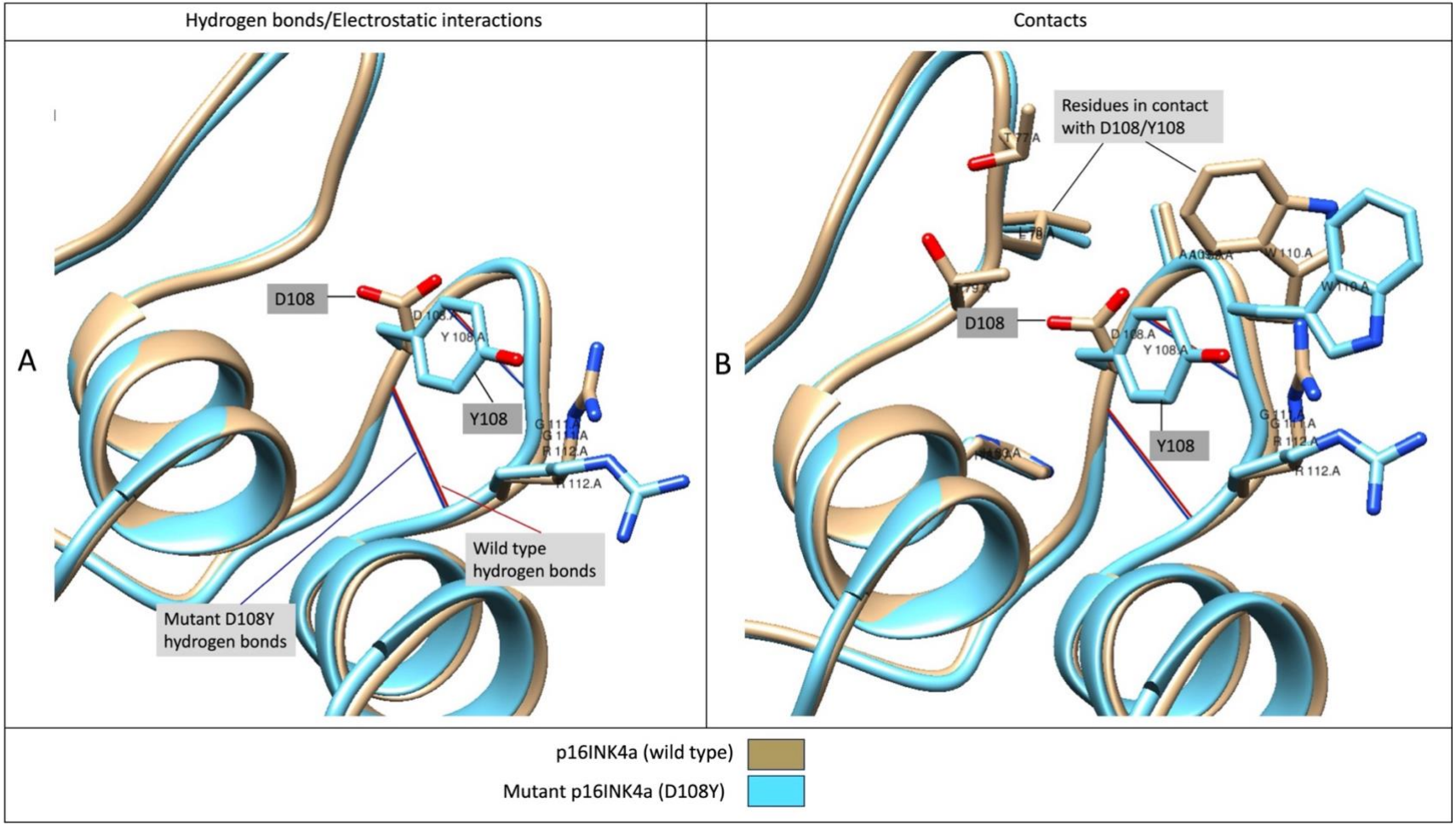


Figure 4.18 Structural analysis of the p.D108Y variant on the structure of p16INK4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the D108 residue forms two salt bridges (indicated in red) between the backbone atoms of aspartic acid 108 with arginine 112, and glycine 111 backbone atoms. Similarly, in the p.D108HY variant, same electrostatic interactions (indicated in blue) occur between the same backbone atoms of the same residues (tyrosine 108 with arginine 112, and glycine 111). The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a protein. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a. The introduction of the tyrosine 108 amino acid leads to increased contacts in the mutant compared to the wild-type. Residues in contact within the wild-type are depicted in brown, while residues in contact within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

In the wild-type p16INK4a, the L130 residue forms two salt bridges involving the backbone atoms of leucine 130, and valine 126 and alanine 134. However, the L130P variant disrupts one of these salt bridges (specifically, the interaction between leucine 130 (donor) and valine 126 (acceptor) interaction), potentially affecting the folding of the ankyrin repeat 4 helices (Figure 4.19A). Furthermore, at the L130 residue, we detected 31 contacts, encompassing polar and nonpolar interactions. Consequently, the introduction of an aromatic residue results in reduction of contacts in the mutant compared to the wild-type, reducing from 28 to 23 interactions (Figure 4.19B). This reduction may result in a loss of various interactions crucial for protein structure, potentially contributing to the decreased stability of the protein.

In summary, the comparative analysis between wild-type and mutant p16INK4a proteins reveals significant structural alterations and potential functional implications. In the p.A68V mutant, the larger, more hydrophobic side chain disrupts electrostatic interactions, leading to increased contacts, potentially compromising protein folding and stability. Similarly, the introduction of uncharged asparagine at position 84 (D84N) alters electrostatic interactions and reduces contacts within the protein. The p.D108N variant introduces new salt bridges, likely to impact protein folding, while p.D108Y replaces a charged residue with a hydrophobic one, potentially compromising protein function. Moreover, mutations such as p.D108H, p.D108N, and p.D108Y lead to increased contacts compared to the wild-type, likely due to the larger side chain amino acids. Additionally, the L130P variant disrupts salt bridges and reduces contacts, potentially contributing to decreased protein stability. These structural alterations may disrupt interactions crucial for p16INK4a stability and function, possibly affecting cell cycle regulation and promoting tumorigenesis.

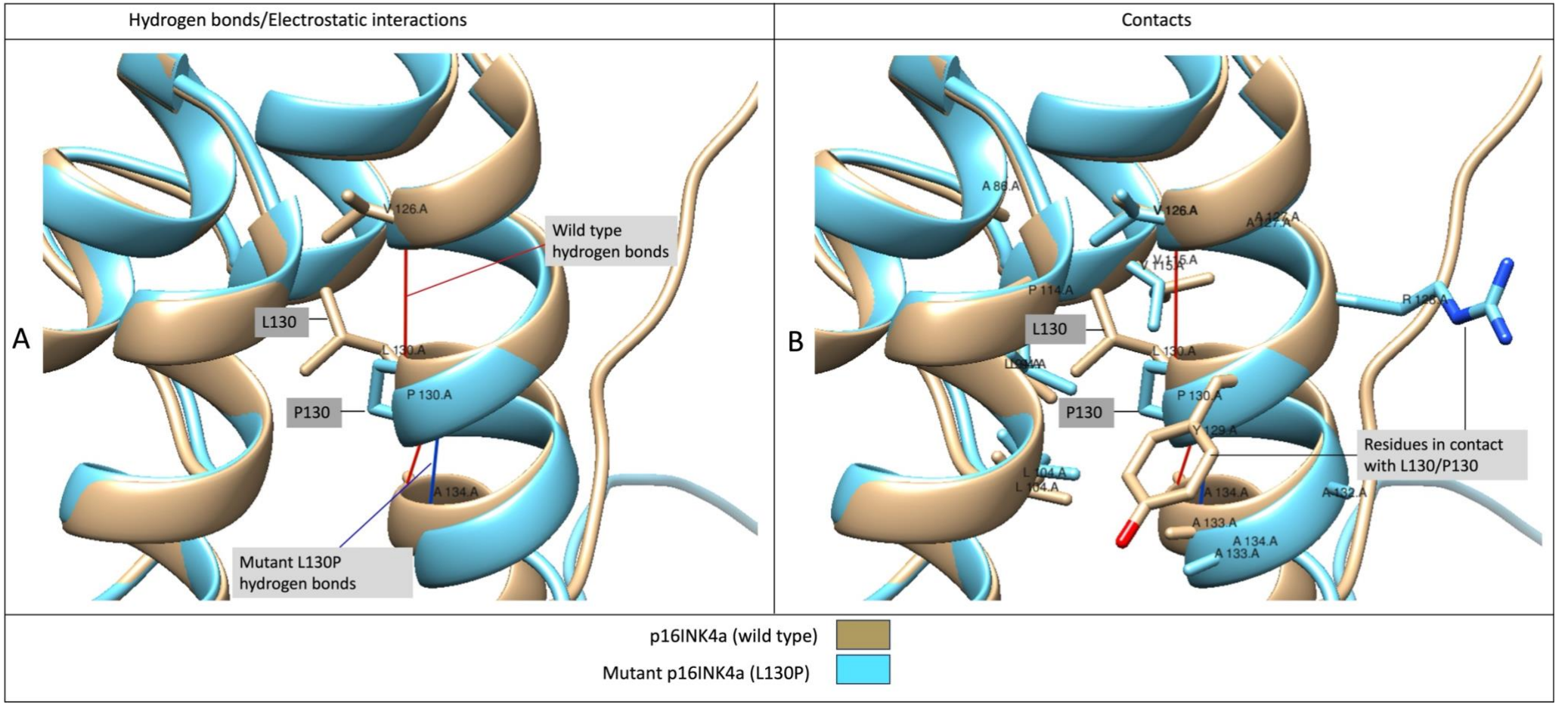


Figure 4.19 Structural analysis of the p.L130P variant on the structure of p16INK4a.

Comparison of the structures of wild-type and mutant p16INK4a proteins. **A**) Analysis of hydrogen bonds and electrostatic interactions analysis in the superimposed structures of wild-type and mutant p16INK4a. In the wild-type, the L130 residue forms two salt bridges (indicated in red) between the backbone atoms of leucine 130, and valine 126 and alanine 134 backbone atoms. In contrast, in the p.L130P variant, leads to the elimination of one of the two salt bridges found in the wild-type. The backbone structure is depicted in either brown, representing the wild-type p16INK4a protein, or blue, indicating the mutant p16INK4a protein. **B**) Analysis of contacts interactions in the superimposed structures of wild-type and mutant p16INK4a. The introduction of the proline amino acid at position 130 led to decreased contacts in the mutant compared to the wild-type. Residues in contact within the wild-type are depicted in brown, while residues in contact within the mutant are shown in blue. Contacts include various direct interactions: polar and nonpolar, favourable, and unfavourable interactions. Mutant PDB files for p16INK4a were modelled with AlphaFold2, together with ColabFold [397]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

4.2.4.1.2 p14ARF

p14ARF exhibits a distinctive molecular profile, characterized by its high basicity and hydrophobicity. Its basic nature is attributed to its high arginine content, constituting over 20% of its amino acid composition, while notably lacking in lysine amino acids [169, 178, 332]. Interestingly, most of the hydrophobic residues are predominantly packed in the exon 1β-encoded N-terminal region of the molecule, which encompasses the secondary structure elements (β-pleated sheets and α-helices) of the protein. In contrast, most of the charged residues are situated in the exon 2 encoded C-terminal region (Figure 4.20 and Figure 4.21A).

Moreover, a significant portion of the hydrogen bond network is observed within the N-terminal region of the molecule (Figure 4.21A), playing a crucial role in stabilizing the structure and facilitating interaction with other molecules, such as MDM2. Notably, p14ARF lacks structural motifs unless it forms complexes with other targets, which facilitate its folding process [178]. Proteins that bind to p14ARF, such as MDM2, typically interact with the first 64 amino acid residues of p14ARF [167, 404]. Furthermore, its noteworthy that the N-terminal domain of p14ARF binds to MDM2 with similar affinity as the full-length protein, indicating that the N-terminal 64 amino acids, encoded by exon 1β, are essential and sufficient for binding to MDM2 [167].

In our cohort, we identified missense variants resulting in alterations such as p.P72L, p.P94L, p.R98Q, p.R122P, p.R122Q, and p.R122L. Importantly, these mutations occur outside the N-terminal domain, which contains the first 64 residues crucial for binding to *p14ARF* targets such as *MDM2*. Furthermore, most of these alterations involve substituting a charged arginine residue with an uncharged or hydrophobic residue, suggesting that these alterations could potentially affect the structural interaction or dynamics of the protein.

Further analysis revealed that none of these mutated residues formed hydrogen bonds or electrostatic interactions with other residues within the protein (Figure 4.21B). Moreover, when comparing the structures of the wild-type and mutant p14ARF proteins, our analysis revealed no hydrogen bonds or electrostatic interactions with surrounding residues in any of the mutants (results not shown). The absence of such interactions suggests that these mutations may not directly affect the stability or conformation of p14ARF through disrupting specific interactions.

In summary, while these mutations may not directly affect the binding of p14ARF to its targets, they could contribute to cancer development through alternative mechanisms. These mechanisms may include influencing protein localization, or interactions with other cellular components, leading to dysregulation of critical pathways involved in tumorigenesis.

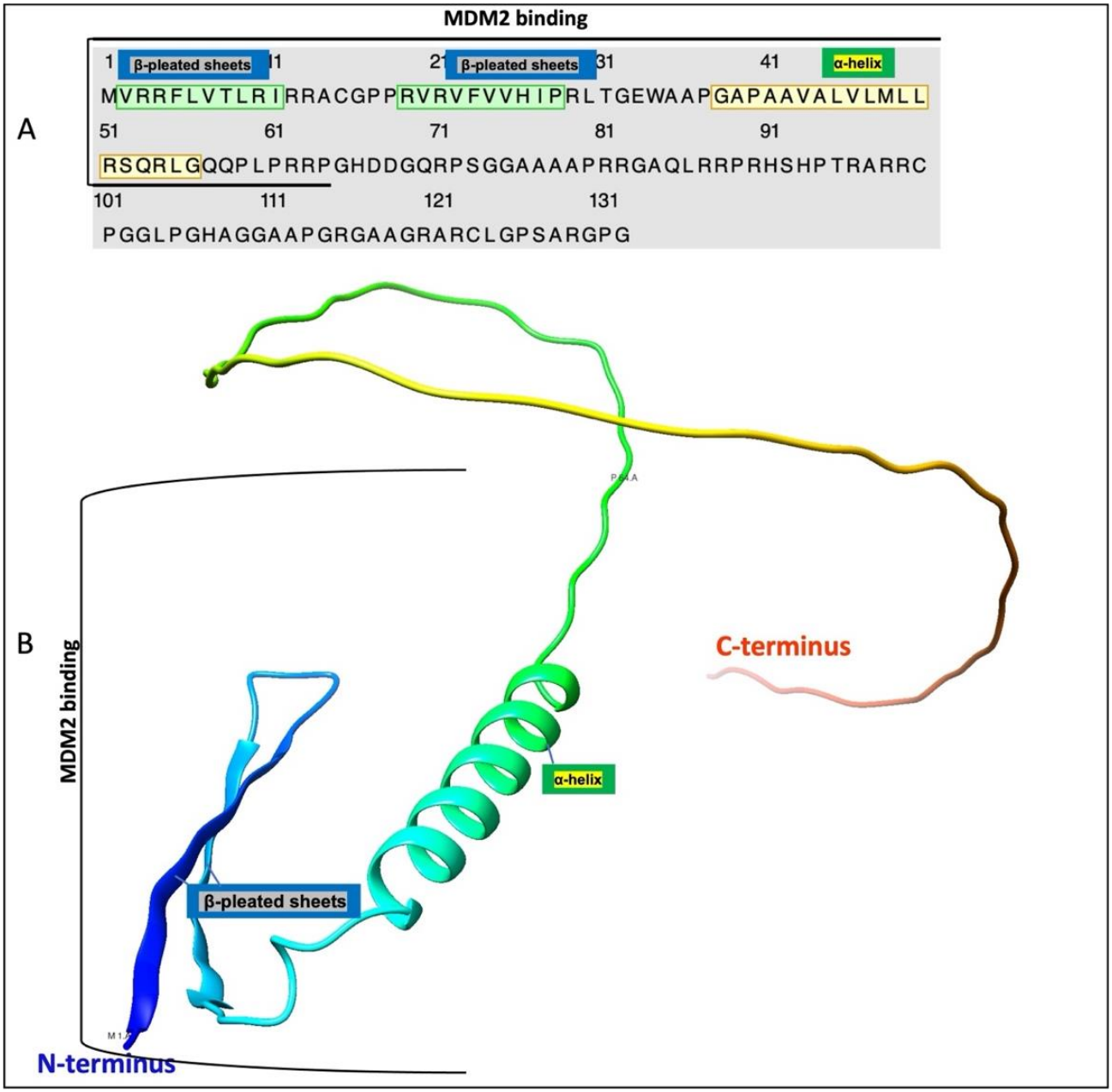


Figure 4.20 Sequence and structure of p14ARF.

p14ARF sequence and structure. **A**) The primary amino acid sequence of p14ARF. In the sequence, regions of secondary structure are highlighted in green for β-pleated sheets and yellow for α-helices, respectively. **B**) The structure of p14ARF. The wildtype PDB files for p14ARF (AF_AFQ8N726F1) was downloaded from RCSB Protein Data Bank, (https://www.rcsb.org) [396]. UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

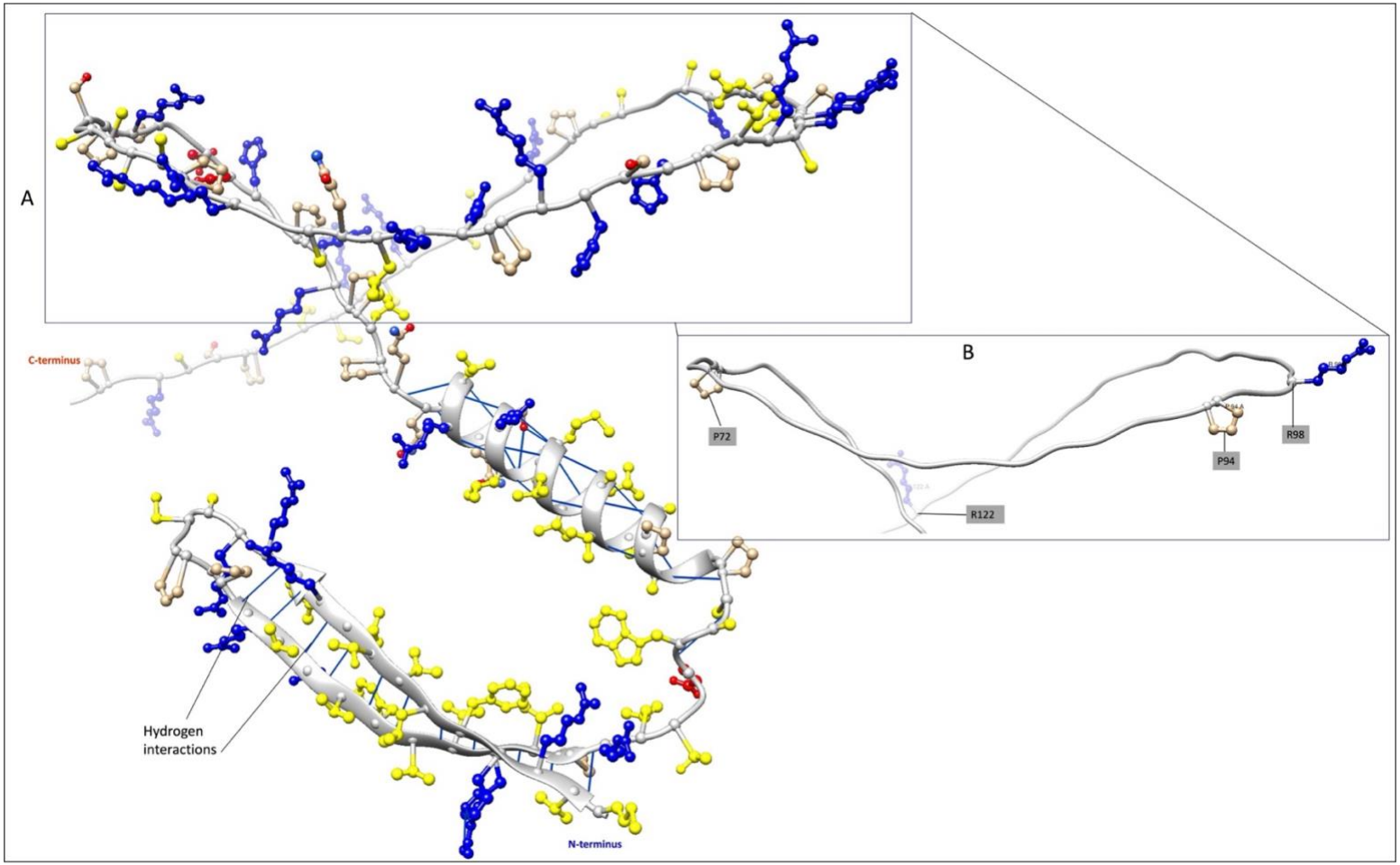


Figure 4.21 Molecular interaction pattern of wildtype p14ARF.

The hydrogen bond interactions for wild-type p14ARF protein. **A**) The hydrogen bond interactions within the p14ARF protein, highlighting hydrophobicity and electrostatic interactions. Positively and negatively charged residues are marked in blue and red, respectively, while the hydrophobic residues are denoted in yellow. The backbone structure is depicted in grey. p14ARF exhibits a high arginine content, with over 20% of its amino acids being arginine, and lacks lysine amino acids. Interestingly, the majority of the hydrophobic residues are predominantly packed in the exon 1β-encoded N-terminal region of the molecule, encompassing the secondary structure regions of the protein , while most of the charged residues are situated in the exon 2 encoded C-terminal region. Hydrogen bonds are represented in blue and defined by solid lines, with the interacting protein residues represented in ball-and-stick format. **B**) Focus on specific residues (p.P72, p.P94, p.R98 and p.R122) identified with missense variants in exon 2 of p14ARF. No hydrogen bond interactions are formed between these residues. Residues of interest are labelled in grey. The wildtype PDB files for p14ARF (AF_AFQ8N726F1) was downloaded from RCSB Protein Data Bank, (https://www.rcsb.org) [396]. The UCSF Chimera tool [399] was utilized for protein visualization and molecular structure analysis.

4.2.4.2 Analysis of p14ARF and p16INK4a missense mutations using In-Silico bioinformatics tools

In addition to our structural analysis, we used three computational tools (PolyPhen-2, I-mutant2.0, and SIFT) to assess the potential impact of missense mutations on p16INK4a and p14ARF protein stability. This integrated approach enables us to not only understand the structural implications of these mutations but also predict their functional consequences. These tools predict the effects and consequences of mutations in proteins, enhancing our structural findings by providing insights into how mutations could interfere with protein stability and function.

PolyPhen-2 (Polymorphism Phenotyping v2) [405] predicts the pathogenicity of the missense variants based on several sequence-based and structure-based features characterizing the substitution. It analyses various sequence-based features affected by the missense variant, and considers structural information of the protein, such as from X-ray crystallography or homology modelling, to assess the potential impact on protein structure. Factors like changes in protein stability, alterations in protein-protein interactions, or disruptions in active sites or functional domains are considered. Furthermore, variants with higher scores are predicted probably damaging, while those with lower scores are predicted to be benign. Variants with "probably damaging" predictions are often considered strong candidates for further functional studies or clinical investigations, especially if they occur in genes associated with human diseases [405-407].

I-Mutant v2.0 [408] predicts the effects of these mutations on protein stability by extracting features from the protein sequence and mutation(s) provided. It considers factors such as amino acid properties, structural information, and sequence conservation to generate predictions for stability change caused by the mutations. The output typically includes a stability score or deltadeltaG (ΔΔG) value, where positive values suggest destabilizing mutations and negative values suggest stabilizing mutations [408].

SIFT (Sorting Intolerant From Tolerant) [409] predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. It aligns the protein sequence with homologous sequences from related organisms or proteins with similar functions to identify conserved regions, where substitutions are more

likely to disrupt protein function. SIFT classifies substitutions as damaging or tolerated by considering the physical properties of amino acids, such as size, charge, and hydrophobicity, and evaluates how the substitution affects these properties and thus the function of the protein [409]. All analyses were conducted using the default settings of the programs [407].

The analysis revealed interesting insights regarding the impact of missense variants on both p14ARF and p16INK4a proteins. In case of p14ARF, four out of six missense variants were predicted to be deleterious, potentially affecting protein stability according to PolyPhen-2 and I-mutant2.0. However, SIFT predicted these variants to be tolerable amino acid substitutions. In contrast, SIFT predicted the remaining two p14ARF mutations as intolerant variants, while PolyPhen-2 either classified their impact on p14ARF's structure and function as benign or as not determinable (ND) (Table 4.5). An "ND" result from PolyPhen-2 indicates that the tool couldn't assess the mutation's impact due to various reasons such as insufficient data or unusual mutation types [405].

These p14ARF results align with our initial structural analysis, as these mutations occur outside the first 64 residues crucial for binding to *p14ARF* targets including *MDM2*, furthermore, the fact that none of these mutated residues formed hydrogen bonds or electrostatic interactions with other residues within the protein may explain why most of these mutations were predicted tolerable.

Table 4.5 Comparison of the effects of *p14ARF* mutations using three different predictive tools.

Gene [a]	Pos (GRCh37) [b]	Mutation [c]	Exon [d]	rsID number [e]	PolyPhen-2 [f]	I-mutant2.0 [f]	SIFT [f]
P14ARF	g.21971186G>A	c.215C>T [p.P72L]	2	rs121913387	Probably damaging	Stability decrease	Tolerated
	g.21971120G>A	c.281C>T [p.P94L]	2	rs121913388	ND	Stability decrease	Not tolerated
	g.21971108C>T	c.293G>A [p.R98Q]	2	rs11552822	Benign	Stability decrease	Not tolerated
	g.21971036C>G	c.365G>C [p.R122P]	2	rs121913381	Probably damaging	Stability increase	Tolerated
	g.21971036 C>T	c.365G>A [p.R122Q]	2	rs121913381	Probably damaging	Stability decrease	Tolerated
	g.21971036 C>A	c.365G>T [p.R122L]	2	rs121913381	Probably damaging	Stability decrease	Tolerated

Impact of the *CDKN2A* mutations on p14ARF protein function.
[a] The protein submitted for analysis.
[b] The genomic location of the substitution using the GRCh37 (Genome Reference Consortium Human Build 37).
[c] The codon that has been mutated, showing both the altered base within the coding sequence and the affected codon within the protein sequence.
[d] The exon of *p14ARF* where the mutation is found.

[e] The rsID number (Reference SNP cluster ID). If the substitutions have a variant overlapping at the same position, the rsID is displayed. rsID numbers are obtained from https://www.ncbi.nlm.nih.gov/snp/. However, the alleles may not be the same.
[f] The three tools utilized for the predictions.

In the case of p16INK4a, four out of six p16INK4a missense variants were predicted by all three tools to be intolerant and deleterious substitutions, and potentially affecting protein stability, which aligns well with our initial structural analysis. Meanwhile, SIFT identified the remaining two variants to be tolerable amino acid substitutions (Table 4.6). Interestingly, our initial analysis showed that the two variants (p.D84N and p.D108N) predicted as tolerable by SIFT altered electrostatic interactions within the protein, potentially contributing to decreased protein stability, a prediction similar to that of the I-Mutant v2.0 tool. However, it is worth noting that SIFT classifies substitutions based on the physical properties of amino acids, such as size, charge, and hydrophobicity [409], rather than considering the specific position of the mutation. Given that these residues are present in ankyrin repeat 3 and in the linking loop 3, for p.D84N and p.D108N, respectively, these residues are involved in interactions with CDK [401]. This distinction could explain why these mutations were predicted tolerable despite potentially affecting protein stability.

Table 4.6 Comparison of the effects of *p16INK4a* mutations using the three different predictive tools.

Gene [a]	Pos (GRCh37) [b]	Mutation [c]	Exon [d]	rsID number [e]	PolyPhen-2 [f]	I-mutant2.0 [f]	SIFT [f]
P16INK4a	g.21971155G>A	c.203C>T [p.A68V]	2	rs1060501260	Probably damaging	Stability increase	Not tolerated
	g.21971108C>T	c.250G>A [p.D84N]	2	rs11552822	Probably damaging	Stability decrease	Tolerated
	g.21971036C>G	c.322G>C [p.D108H]	2	rs121913381	Probably damaging	Stability decrease	Not tolerated
	g.21971036C>T	c.322G>A [p.D108N]	2	rs121913381	Probably damaging	Stability increase	Tolerated
	g.21971036C>A	c.322G>T [p.D108Y]	2	rs121913381	Probably damaging	Stability decrease	Not tolerated
	g.21970969A>G	c.389T>C [p.L130P]	2	rs2131092580	Probably damaging	Stability decrease	Not tolerated

Impact of the *CDKN2A* mutations at the p16INK4a protein function.
[a] The protein submitted for analysis.
[b] The genomic location of the substitution using the GRCh37 (Genome Reference Consortium Human Build 37).
[c] The codon that has been mutated, showing both the altered base within the coding sequence and the affected codon within the protein sequence.
[d] The exon of *p16INK4a* where the mutation is found.
[e] The rsID number (Reference SNP cluster ID). If the substitutions have a variant overlapping at the same position, the rsID is displayed. rsID numbers are obtained from https://www.ncbi.nlm.nih.gov/snp/. However, the alleles may not be the same.
[f] The three tools utilized for the predictions.

In summary, several variants identified in these proteins, especially p16INK4a have the potential to disrupt protein stability and function. Furthermore, these findings highlight the complexity of predicting the functional impact of missense variants and emphasize the

importance of considering multiple computational tools and structural analyses to obtain a comprehensive understanding. Further experimental validation is necessary to elucidate the precise functional consequences of these variants in cancer biology.

4.3 Discussion

CDKN2A is an important tumour suppressor gene which plays an important role in cell cycle regulation and prevents tumour development [166, 170]. The *CDKN2A* locus contains two unrelated proteins both capable of inducing cell cycle arrest when overexpressed, namely p14ARF and p16INK4a [164, 165, 196, 300, 332]. In previous chapters (Chapter 3, section 3.2.2), we found variable levels of *p14ARF* and *p16INK4a* mRNA in OSCC tumours compared to their normal adjacent tissues, *p16INK4a* was significantly reduced or even absent in more than 60% of the OSCC tumours, while *p14ARF* mRNA levels were significantly lower in 48% tumours of the OSCC tumours. Hence, in part of the study, we carried out siRNA-mediated knockdown of *p14ARF* and *p16INK4a* to explore their roles in OSCC cell lines. We investigated the consequences of *p14ARF* and *p16INK4a* deficiency on various signalling pathways, including those involved in cell cycle regulation, apoptosis, anti-apoptosis, and the NFE2L2-KEAP pathway. Furthermore, we performed an in-silico analysis to examine the pathogenicity of *p14ARF* and *p16INK4a* missense variants. Additionally, we investigated the effects of damaging missense variants on the structure of the p16INK4a protein.

The initial investigations involved screening the indicated cell lines for *CDKN2A* exon 2 mutations detected in patients by WGS and WES techniques (as discussed in Chapter 2). Additionally, we assessed the mRNA levels of *p16INK4a* and *p14ARF* in OSCC cell lines, using a human telomerase immortalized oesophageal cell line, EPC2, was used as control cells. One drawback of this analysis is using a single control cell line, particularly a telomerase immortalized cell line as our normal control. Ideally, employing multiple control cell lines representing different cellular contexts would provide a more comprehensive comparison and enhance the robustness of our findings and strengthen the validity of the results. Our analysis revealed that only one of the eleven mutations examined (p16INK4a c.358G>T [p.E120*]) was present in KYSE30. Four cell lines: KYSE150, KYSE450, WHCO1 and WHCO6 were identified lacking exon 2 amplification of *CDKN2A* based on our PCR amplification analysis. Our findings are partly consistent with results reported by Li etal. [410], who found *CDKN2A* deletions and point mutations in 95.8% of OSCC cell lines analysed in their study including KYSE450 and KYSE180, as well as *CDKN2A* gains and point mutations in KYSE30 and KYSE150 cells. It has been reported that *p16INK4a* and *p14ARF* expression is absent or significantly reduced in OSCCs [174, 175, 192]. The analysis of *p14ARF* and *p16INK4a*

mRNA levels by RT-qPCR analysis showed significantly lower to absent mRNA levels in most of the OSCC cell lines, including cell lines where exon 2 could not be amplified. Of note, *p16INK4a* mRNA levels were significantly lower in all the OSCC cell lines except KYSE30, when compared to the control cells. On the other hand, *p14ARF* mRNA levels were significantly lower in four and five OSCC cell lines, respectively. Notably, *p14ARF* mRNA levels were significantly higher in KYSE30 and WHCO5 cells. Furthermore, there was no significant difference *p16INK4a* mRNA levels between KYSE30 cells and EPC2 cells, indicating that the truncating mutation in KYSE30 had no significant effects on the *p16INK4a* mRNA levels in these cells. These results suggest that there is some level to which truncating mutation have a signification impact on gene expression, and truncating mutations occurring at the end of a coding region might have no to little effect on gene expression or mRNA stability. Additionally, mutations causing premature stop codons may activate nonsense-mediated mRNA decay, leading to mRNA degradation [382]. Nevertheless, if the mutation occurs in a region where nonsense-mediated mRNA decay does not occur, or if the mRNA remains stable enough to sustain detectable levels, this might explain why *p16INK4a* mRNA levels are not significantly affected in KYSE30 cells. The differential expression of *p14ARF* and *p16INK4a* in OSCC cell lines suggests a potential significance of exon 2 mutations in *CDKN2A* and their different impacts on *p14ARF* and *p16INK4a* mRNA levels in OSCC cells.

Both *p14ARF* and *p16INK4a* play important roles in regulating cell cycle, senescence and apoptosis [38, 163, 166, 168, 170, 178, 179, 196, 314, 332]. Our analysis of the effects of *p14ARF* and *p16INK4a* knockdown in signalling pathways in KYSE30 cells showed altered expression of cell cycle regulator mRNA levels. Specifically, there was an upregulation of *Rb*, *MDM2* and *p21* mRNA levels, along with downregulation *CCND1* mRNA levels. *Rb* and *CCND1* and *MDM2* and *p21*, belong to two distinct yet overlapping tumour suppressor pathways: p16INK4a/Rb and p14ARF/p53/p21 pathways, respectively [163, 167, 168, 179]. Furthermore, the regulation of these pathways differs distinctly [167]. In the p14ARF/p53/p21 pathway, p14ARF and MDM2 interaction promote the MDM2 protein degradation, thereby inhibiting MDM2-dependent p53 degradation and leading to increased expression and stabilization of p53. This increases p53 activity and increased expression of p53-regulated genes, such as p21 [167, 169]. The knockdown of *p14ARF* and *p16INK4a* in KYSE30 cells significantly elevated *MDM2* mRNA levels, suggesting that loss of *p14ARF* and *p16INK4a* lead to upregulation of *MDM2*. This upregulation of *MDM2* could promote the rapid degradation of p53 and inhibit the ability of p53 to promote cell cycle arrest in KYSE30 cells.

In contrast, *p21* mRNA levels were upregulated in *p14ARF* and *p16INK4a* knockdown cells. Although p21 is primarily associated with p53 in terms of its cell cycle arrest function [318], our results confirm the possibility of p53-independent pathways leading to p21 expression in response to cellular stress, such as DNA damage or oxidative stress [318, 411]. Furthermore, a truncation mutation in *p16INK4a* was detected in KYSE30 cell line, it is unclear how *p16INK4a* knockdown leads to elevated *p21* mRNA levels, suggesting the existence of alternative pathways or mechanisms in KYSE30 cells that regulate *p21* expression independently of p16INK4a and p53 interactions. Several studies have demonstrated a positive correlation between the tumour suppressor proteins p16INK4a and p53. It has been shown that p16INK4a acts as an effective positive regulator of p53 expression at the protein level, mediated via p16INK4a-dependent stabilization of the p53 protein through suppressing the expression of MDM2 in both human and mouse cells [412]. We have shown that although p53 mRNA levels were not significantly altered, *MDM2* mRNA levels were upregulated in *p16INK4a* deficient cells, consistent with previous findings [412]. This suggests that the levels of p53 and *p16INK4a* are positively correlated, and lack of *p16INK4a* expression upregulates *MDM2* expression, rendering the p53 protein vulnerable to MDM2-mediated degradation.

Previous studies have shown that overexpression of the anti-apoptotic proteins *BCL-XL* or *BCL2* inhibits cell death by inhibiting caspase activation [391, 392]. Additionally, the overexpression of p14ARF and p16INK4a downregulates the expression of anti-apoptotic genes such as MCL1, BCL-XL and BCL2 in p53 protein-deficient cells that facilitates mitochondrial apoptosis induced by p14ARF or p16INK4a via BAK/BAX proteins in various cancers [386, 388, 390]. Our analysis of apoptotic genes showed elevated *Caspase-3* and *Caspase-9* mRNA levels in *p14ARF* and *p16INK4a* knockdown cells. This suggests that *p14ARF* and *p16INK4a* may exert a suppressive or regulatory effect on the expression of these pro-apoptotic genes. The increased expression of *Caspase-3* and *Caspase-9* would likely lead to an increase in apoptosis, a mechanism for controlling aberrant cell growth [413]. However, the specific mechanism by which *p16INK4a* knockdown leads to elevated *Caspase-3* and *Caspase-9* mRNA levels remains unclear and requires further investigation. Despite the elevated transcription of pro-apoptotic *Caspase-3* and *Caspase-9* by knockdown of *p14ARF* and *p16INK4a*, the mRNA levels of anti-apoptotic genes; *BCL-XL* and *BCL2* mRNA levels were significantly upregulated in the same cells. The pro-apoptotic activity of BAX/BAK is counteracted by the expression of anti-apoptotic BCL2-related proteins such as BCL-XL and BCL2, which inhibit Caspase activation, thus blocking the Caspase-dependent pathway of

apoptosis [391, 392]. This overexpression of *BCL-XL* and *BCL2* in the same cells could thereby contribute to evading apoptosis in KYSE30 cells by inhibiting Caspase activation, subsequently attenuating p16INK4a and p14ARF-induced apoptosis, and ultimately leading to cell survival. Similarly, siRNA-mediated depletion *p14ARF* and *p16INK4a* mRNA resulted in significantly elevated *NFE2L2* mRNA levels, which could potentially lead to upregulation of NFE2L2 expression. High *NFE2L2* expression contribute to increased expression of *NFE2L2* target genes conferring advantages in terms of stress resistance, cell survival and proliferation [214, 217].

In addition to knockdown studies, we conducted *in silico* analysis to evaluate the potential impact of missense mutations on the stability and pathogenicity p14ARF and p16INK4a proteins to distinguish between deleterious and non-deleterious mutations. Our comparative analysis between wild-type and mutant p16INK4a proteins revealed that all examined mutations were predicted to be damaging variants. These mutations induced significant structural alterations, such as introduction of larger, more hydrophobic, or uncharged residues, disrupting hydrogen bonds and electrostatic interactions. Consequently, these alterations increased contacts, potentially compromising protein folding and stability [402, 403]. Furthermore, variants such as p.A68V and p.D84N were identified in crucial regions, in the linking loop 2 and ankyrin repeats 3, respectively. These regions, especially in the second and third ankyrin repeats, and the linking loops, play crucial roles in interacting with CDK4 for cell cycle regulation [400, 401]. Alterations in the electrostatic interactions within these regions affect the ability of the protein to interact with other molecules, such as cyclin-dependent kinases, thereby inhibiting the p16INK4a activity. In previous studies, structural modifications induced by various mutations such as D84H, G101W, and H123Q led to decreased to undetectable p16INK4a activity [400]. Notably, the p.L130P mutation located in the fourth ankyrin repeat, which along with the first ankyrin repeat (including the flexible N- and C-termini), crucial for stabilizing the overall structure of p16INK4a [400, 401], may disrupt interactions important for p16INK4a stability and function, potentially affecting cell cycle regulation and promoting tumorigenesis [401].

In case of p14ARF, most missense variants were predicted to be tolerable substitutions. These findings align with our initial structural analysis, as these mutations occur outside the first 64 residues that are crucial for binding to *p14ARF* targets such as *MDM2*. None of these mutated residues were involved in hydrogen bonds or electrostatic interactions with other residues

within the protein, explaining why most of these mutations were predicted tolerable. These mechanisms may include influencing protein localization, or interactions with other cellular components, leading to dysregulation of critical pathways involved in tumorigenesis.

In conclusion, our findings reveal the significance of reduced mRNA levels of *p14ARF* and *p16INK4a* in KYSE30 cells, which contribute to the dysregulation of multiple cancer signalling pathways. This dysregulation involves altered expression of cell cycle regulators, including the upregulation of *MDM2* mRNA levels, as well as the upregulation of anti-apoptotic genes *BCL2* and *BCL-XL* mRNA levels. Consequently, the upregulation of *MDM2* may promote cell cycle progression and hinder G1 cell cycle arrest mediated by p53. Moreover, the observed upregulation of anti-apoptotic genes and *NFE2L2* mRNA levels in *p14ARF* and *p16INK4a* mRNA-depleted cells may contribute to cell survival, and growth of apoptosis-resistant and stress-resistant cells, all of which are pivotal events in tumorigenesis. Furthermore, our in-silico mutation analysis reveals how damaging missense mutations such as p.A68V, p.D84N, p.D108H, p.D108N, p.D108Y, and p.L130P directly impact the native structure of the p16INK4a protein, potentially influencing its function. These findings collectively underscore the complex roles of *p14ARF* and *p16INK4a* in modulating key pathways implicated in cancer development.

Chapter 5
Discussion and conclusion

5.1. Overall discussion and conclusion

OSCC is one of the most prevalent cancers in Sub-Saharan Africa. Although genomic alterations associated with OSCC have been extensively studied in Western (American and European) and Eastern (Chinese, Japanese, Korean and Indian) populations [77, 124, 125, 127, 128, 131-133, 136, 142, 154-157], data from Sub-Saharan Africa are limited [22, 134]. In this study, we performed an integrative analysis of whole genome and whole exome sequence analysis on 31 and 67 OSCC genomes, respectively, from South African patients to investigate the mutational landscape of South African OSCC patients. Our analysis revealed frequent genomic mutations, identified key driver genes, and uncovered novel mutational signatures and molecular subtypes associated with OSCC in South African patients. Mutations in genes such as *p14ARF* and *p16INK4a* were implicated in OSCC development. Further experiments demonstrated that knockdown of *p14ARF* and *p16INK4a* mRNA in KYSE30 cells lead to the dysregulation of various cell cycle signalling pathways. Additionally, we observed that missense mutations in *p16INK4a*, including p.A68V, p.D108N, p.D108H, p.D108Y, and p.L130P, disrupt electrostatic interactions, potentially affecting the protein's folding, stability, and probably function.

The observed non-silent mutation load (2.5 mutations/Mb) in both WGS and WES analyses was consistent with reported mutation loads ranging from 1.9 to 3 mutations per Mb in OSCC [22, 123, 125, 128, 135, 139, 250]. Analysis of transition and transversion rates in coding sequences in our samples showed that C:G> T:A transitions were the most common mutations, followed by C:G>A:T and C:G>G:C transversions, similar to previous studies of OSCC [131, 138]. The C:G> T:A transitions predominance is consistent with spontaneous cytosine deamination to thymine [227], as a major mutagenic process in OSCC, as previously observed [77, 122, 123, 125, 128, 131, 137, 138, 142, 154]. The most frequently mutated genes including *TP53*, *TTN*, *CDKN2A*, *NOTCH1*, *KMT2D*, *NFE2L2*, *DMD*, *PCLO*, *CSMD3*, *PIK3CA*, *LRP1B*, *FAT3*, *FAT2*, *KMT2C*, *RYR2* and *MUC16*, which have been previously reported in the Asian [77, 123-125, 127, 128, 131, 135, 136, 138, 139, 142, 143, 156] and Malawian populations [22]. Furthermore, through bioinformatics analysis, we identified several driver genes in our cohort. Six genes showed significant recurring mutations across OSCC genomes including

TP53, *CDKN2A* (*p16INK4a*) identified in both WGS and WES analyses, while *CDKN2A* (*p14ARF*), and *KMT2D* were identified in only WGS, and *NFE2L2*, *ZNF750* while *NOTCH1* were detected only in WES analysis. These driver genes align with previous studies findings, highlighting the roles of these genes in OSCC [77, 124, 125, 127, 128, 131-133, 136, 142, 154-157]. Further analysis revealed genetic variants associated with specific pathways and molecular events involved in "cellular responses to stimuli", "disease pathway", "gene expression (transcription)", "extracellular matrix organization", "metabolism of proteins", "signal transduction", "neuronal system" and "cell-cell communication". Many of these pathways were previously reported to be involved in OSCC-associated pathways including cell cycle regulation, DNA repair, and epigenetic modifications, the NOTCH signalling pathway, KEAP1-NFE2L2 pathway [15, 71, 115-126, 128-135].

A novel finding highlighting the heterogeneity and complex molecular mechanism of OSCC was the identification of three distinct groups based on mutation spectrum of our samples: cluster 1, cluster 2a and cluster 2b. These clusters were characterised by varying frequencies of *TP53* alterations and the mutations per Mb sequenced. Cluster 1 tumours, which constituted most of the samples in our cohort, had *TP53* mutations and exhibited a relatively high somatic mutation rate per Mb. In contrast, cluster 2 lacked *TP53* mutations and was more prevalent among black female patients. Further, in WES analyses, cluster 2 could be further divided into subclusters 2a and 2b. Cluster 2a displayed a high mutation rate per Mb, whereas cluster 2b displayed fewer genomic alterations. This distribution of high and low mutation rates has been reported in Malawian OSCC patients and African American OSCC populations [22, 132, 137, 297]. Comparative analysis of mutation rates between our cluster 2a and cluster 2b and previously identified OSCC subgroups showed that our cluster 2b resembled the Malawian subtype 1b [22] as well as the Subtype 3 (OSCC3) among African Americans [132]. These subgroups are characterized by a lower prevalence of *TP53* mutations and fewer genomic alterations with the lowest somatic mutations per Mb [22, 132, 297]. This underscores the diverse molecular profiles within OSCC, influenced by genetic factors and probably environmental factors, highlighting the complex nature of this disease.

By integrating the mutation signatures generated from WES and WGS, we identified distinct signature profiles in OSCC and their correlation with patient demographic variables, behavioural or lifestyle factors as well as their mutation profiles. Importantly, among the seven mutation signatures identified, clock-like signature (SBS1 and SBS5) and APOBEC-mediated

mutation signatures (SBS2 and SBS13) significantly contributed to the mutation burden in our cohort, consistent with previous OSCC studies [22, 77, 123-125, 128, 131, 135, 136, 138, 139, 142, 238-240]. This suggests the diverse mechanisms contributing to mutagenesis, with aging and APOBEC activation playing crucial roles in OSCC tumour development. In addition, our cohort revealed the enrichment of SBS6, SBS10b and SBS15 mutation signatures, suggesting a potential association between defects in DNA repair pathways and OSCC [146], highlighting the roles of defects DNA repair mechanisms in OSCC development. WGS analysis revealed three novel mutational signatures that had not been previously identified. Interestingly, these signatures were not observed in the samples analysed by WES, even though the WES cohort included a larger sample size. The significance of these novel mutational signatures remains unclear and may warrant further investigation.

Evaluation of *p16INK4a* and *p14ARF* mRNA levels in OSCC tumour samples compared to their adjacent normal tissues showed variable expression patterns of these genes. *p16INK4a* and *p14ARF* mRNA levels were significantly lower in 61% and 48% of tumour samples, respectively, while elevated levels of these genes were found in 16% and 25% of tumours, respectively. Similarly, *p14ARF* and *p16INK4a* mRNA levels in OSCC cell lines showed variable expression patterns of *p14ARF* and *p16INK4a*. *p16INK4a* mRNA levels were significantly lower in all OSCC cell lines except KYSE30 cells when compared to the control EPC2 cell line. On the other hand, *p14ARF* mRNA levels were significantly lower in five OSCC cell lines. Knocking down *p14ARF* and *p16INK4a* in KYSE30 cells revealed that reduced mRNA levels of these genes contribute to dysregulation of multiple cancer signalling pathways. This dysregulation included altered expression of cell cycle regulators, notably the upregulation of *MDM2* mRNA levels, and anti-apoptotic genes *BCL2* (only in p16INK4a knockdown) and *BCL-XL* mRNA levels. The upregulation of *MDM2* may promote cell cycle progression and hinder G1 cell cycle arrest mediated by p53. Moreover, the observed upregulation of anti-apoptotic genes and *NFE2L2* mRNA levels in *p14ARF* and *p16INK4a* mRNA-depleted cells may contribute to cell survival, and growth of apoptosis-resistant and stress-resistant cells, all of which are pivotal events in tumorigenesis. *In-sillico* analyses of several *p16INK4a* missense mutations have revealed significant structural alterations with potential functional implications. Mutations introducing larger, more hydrophobic side chain (such as p.A68V and p.D108Y) disrupts electrostatic interactions, potentially compromising protein folding and stability by increasing molecular contacts. On the other hand, mutations introducing uncharged residues (D84N, p.D108N, and L130P) alter electrostatic interactions

and reduce contacts within the protein, which could contribute to decreased stability of the protein. Previous studies have shown that structural changes in p16INK4a resulting from mutations like D84H have led to decreased or even undetectable p16INK4a activity [400], potentially disrupting the regulation of the cell cycle, increasing susceptibility to oncogenic stimuli, and promoting tumour progression [169, 401].

In summary, our study presents the first integrated analysis combining genome and exome sequencing to explore the mutational landscape of OSCC in a South African cohort. We identified key driver genes, mutation signatures, and OSCC subtypes are associated with OSCC development in this population. We observed genomic alterations that align with findings from diverse populations, including those from Malawi, African Americans, Brazilians, and Asians. Furthermore, our study highlights both similarities and potentially distinct genetic characteristics unique to South African and Sub-Saharan African populations. Our integrated study deepened our understanding of the molecular alterations and provides crucial insights into the molecular mechanisms driving OSCC pathogenesis within the South African population.

5.2 Study limitations and future work

Despite the smaller sample size analysed by WGS (n=31) compared to WES (n=67), the data quality from WGS was high and showed strong alignment with the WES results. One of the major limitations of the study was the relatively small sample size available for the Kaplan-Meier analysis. Furthermore, gene expression analysis was conducted solely at the mRNA level, without corresponding analysis at the protein level. While mRNA levels provide valuable insights into gene expression patterns and potential dysregulation, protein expression levels can further validate these findings and provide a more direct link to functional activity.

Going forward, larger and more diverse patient cohorts are needed to improve the reliability and generalizability of findings regarding genetic alterations, gene expression patterns, and their correlations with clinical outcomes in OSCC. Furthermore, specific mutations identified in candidate driver genes such as p16INK4a should be prioritized for validation through functional studies, including approaches such as genome editing with CRISPR-Cas9, to elucidate their roles in OSCC development.

In summary, while the present study provides valuable insights, addressing these limitations will enhance our understanding of OSCC pathogenesis and pave the way for more targeted therapeutic strategies in the future.

Chapter 6
Materials and Methods

6.1 Materials

6.1.1 Sample collection

6.1.1.1 Sample cohort

A total of 177 OSCC patients were enrolled in this study. Samples from patients diagnosed with OSCC were obtained from two hospitals (Groote Schuur Hospital in Cape Town, Western Cape and Charlotte Maxeke Hospital in Johannesburg, Gauteng, South Africa). Written informed consent was obtained from all patients before recruitment into the study. Ethics approval was obtained from the University of Cape Town/Groote Schuur Hospital Human Research Ethics Committee (Cape Town, South Africa; approval number: HEC040/2005). Biopsies from oesophageal tumour tissue and adjacent normal tissues (10 cm from tumour site) and blood samples from 177 patients by a research nurse. Samples were transported back to the laboratories at the University of Cape Town (UCT) and the University of the Witwatersrand (Charlotte Maxeke) for processing and storage. Epidemiological data for each individual were collected, which included gender, age at diagnosis, ethnicity, smoking status and alcohol consumption. The inclusion criteria of eligible patients were as follows: 1. Histologically confirmed OSSC, 2. Participants did not receive any of cancer treatment before surgery, 3. Written informed consent and ability to understand the nature of the study. Patients meeting any of the following criteria are not eligible for this study (exclusion criteria): 1. Patients who had received any form of radio or chemotherapy, 2. Patients who were too ill to go through the procedure. 3. Patients with histologically confirmed forms of cancer other than OSCC.

6.1.1.2 DNA and RNA extraction and processing

DNA and RNA was extracted as described in section 6.2.1.2- 6.2.1.4 and subjected to whole genome sequencing or whole exome sequencing at the Wellcome Sanger Institute in Cambridge in the United Kingdom. The whole genome sequencing was performed on DNA from 31 pairs of OSCC tumours and their matched blood samples as indicated in Table 6.1. The cohort included 20 cases from Groote Schuur Hospital and 11 cases from Charlotte Maxeke Johannesburg Academic Hospital. The WGS cohort comprised 13 men and 18 females. The mean age of the patients was 58 years, ranging from 37 to 81 years. For smoking status, 10 cases were smokers, 5 cases were ex-smokers, 16 cases were non-smokers. Except

for 12 patients with unknown alcohol consumption histories, 13 patients had a history of regular alcohol consumption, while 6 patients had no history of alcohol consumption. There were 29 black patients and 2 mixed ancestry patients. Black subjects were mainly Xhosa or Zulu speakers from the Western Cape province of South Africa who migrated from the Eastern Cape over the past 1–2 generations (cases from GSH) or from KwaZulu-Natal province of South Africa (cases from Charlotte Maxeke), respectively. The mixed ancestry subjects were from the Western Cape.

Table 6.1 A summary of the study samples analysed using WGS.

	Patient Number	Age	Sex	Smoking status	Alcohol	Ethnicity	Histology
Groote Schuur Hospital, UCT, Cape Town	PD39445	57	Female	No	No	Black	OSCC
	PD39446	45	Male	Ex-smoker	Yes	Black	OSCC
	PD39447	41	Male	Yes	Yes	Black	OSCC
	PD39448	52	Male	Ex-smoker	Yes	Black	OSCC
	PD39449	79	Female	No	No	Black	OSCC
	PD39450	50	Female	Ex-smoker	Yes	Black	OSCC
	PD39451	71	Male	Ex-smoker	Yes	Black	OSCC
	PD39452	53	Female	No	Yes	Black	OSCC
	PD39453	37	Male	Yes	Yes	Black	OSCC
	PD39455	48	Female	No	No	Black	OSCC
	PD39456	41	Male	Yes	No	Black	OSCC
	PD39457	57	Male	No	Yes	Black	OSCC
	PD39458	60	Female	Yes	Yes	Black	OSCC
	PD39459	64	Female	No	No	Black	OSCC
	PD39460	56	Male	Yes	Yes	Black	OSCC
	PD50649	66	Female	No	No	Black	OSCC
	PD50650	60	Female	Yes	Yes	Black	OSCC
	PD50651	70	Male	Ex-smoker	Yes	Black	OSCC
	PD50653	57	Female	Yes	No info	Mixed ancestry	OSCC
	PD51372	60	Male	Yes	Yes	Mixed ancestry	OSCC
Charlotte Maxeke Hospital, WITS, Johannesburg	PD44691	70	Female	No	No info	Black	OSCC
	PD44694	54	Female	No	No info	Black	OSCC
	PD44695	63	Female	No	No info	Black	OSCC
	PD44696	54	Female	No	No info	Black	OSCC
	PD44697	38	Male	Yes	No info	Black	OSCC
	PD44698	45	Female	No	No info	Black	OSCC
	PD44699	81	Female	No	No info	Black	OSCC
	PD44701	69	Female	No	No info	Black	OSCC
	PD44702	65	Female	No	No info	Black	OSCC
	PD44703	78	Male	No	No info	Black	OSCC
	PD44704	56	Male	Yes	No info	Black	OSCC

This table provides a concise summary of the demographic and clinical characteristics of the cohort analysed using WGS for OSCC tumours. It includes details on sample size, site of collection, gender distribution, age, smoking and alcohol consumption histories, and race distribution among the patients studied.
UCT – University of Cape Town, WITS - University of the Witwatersrand, OSCC – oesophageal squamous cell carcinoma

The whole exome sequencing was performed on 67 pairs of OSCC tumours and matched blood samples (Table 6.2). The cohort included 29 cases from Groote Schuur Hospital cohort and 38 cases from Charlotte Maxeke Johannesburg Academic Hospital cohort. The WES cohort comprised 39 men and 28 females. The mean age of the patients was 61 years, ranging between 28 to 89 years. Except for 4 patients with unknown smoking histories, 24 cases were smokers, 16 cases were ex-smokers, 23 cases were non-smokers. For alcohol consumption, 37 patients had a history of regular alcohol consumption (either drinking at time of diagnosis or in the past), 20 patients had no history of alcohol consumption and 10 patients with unknown history of alcohol consumption. There were 45 black patients, 1 white, 17 mixed ancestry, and 4 of unknown race.

Table 6.2 A summary of the study samples analysed using WES.

Groote Schuur Hospital, UCT, Cape Town

Patient Number	Age	Sex	Smoking status	Alcohol	Ethnicity	Histology
PD55952a	55	Male	Yes	Yes	Mixed ancestry	OSCC
PD55953a	49	Female	Yes	Yes	Mixed ancestry	OSCC
PD55955a	67	Female	No	No	Mixed ancestry	OSCC
PD55956a	47	Male	Yes	Yes	Mixed ancestry	OSCC
PD55958a	86	Female	Ex-smoker	No	Mixed ancestry	OSCC
PD55959a	58	Male	Yes	Yes	Mixed ancestry	OSCC
PD55960a	48	Female	No	Yes	White	OSCC
PD55961a	41	Female	Yes	Yes	Mixed ancestry	OSCC
PD55962a	28	Female	No	No	Black	OSCC
PD55963a	58	Male	Yes	Yes	Black	OSCC
PD55964a	47	Male	Yes	Yes	Mixed ancestry	OSCC
PD55965a	62	Female	Ex-smoker	Yes	Black	OSCC
PD55966a	61	Female	Ex-smoker	No	Black	OSCC
PD55967a	59	Male	Ex-smoker	Yes	Mixed ancestry	OSCC
PD55968a	43	Male	No	No	Mixed ancestry	OSCC
PD55969a	54	Male	Ex-smoker	No info	Mixed ancestry	OSCC
PD55971a	75	Female	No	No	Black	OSCC
PD55972a	58	Male	Yes	No info	Mixed ancestry	OSCC
PD55973a	83	Female	Ex-smoker	Yes	Mixed ancestry	OSCC
PD55974a	52	Male	Yes	Yes	Mixed ancestry	OSCC
PD55977a	57	Female	Yes	Yes	Mixed ancestry	OSCC
PD55978a	41	Female	No	Yes	Black	OSCC
PD55979a	69	Male	Ex-smoker	In the past	Mixed ancestry	OSCC
PD55982a	58	Male	Yes	Yes	Mixed ancestry	OSCC
PD55983a	64	Male	Yes	Yes	Black	OSCC
PD55987a	49	Female	No	No info	No info	OSCC
PD55990b	56	Female	Yes	No info	No info	OSCC
PD55991b	64	Female	No	No info	No info	OSCC
PD55992b	46	Male	Yes	No info	No info	OSCC

Charlotte Maxeke Hospital, WITS, Johannesburg

Patient Number	Age	Sex	Smoking status	Alcohol	Ethnicity	Histology
PD55994a	66	Male	No info	No info	Black	OSCC
PD55995a	64	Female	No info	No info	Black	OSCC
PD55996a	65	Female	No	No	Black	OSCC
PD55997a	63	Female	No	In the past	Black	OSCC
PD55998a	52	Male	Yes	In the past	Black	OSCC
PD55999a	77	Male	No	Yes	Black	OSCC
PD56000a	72	Male	Yes	No	Black	OSCC
PD56001a	61	Male	Yes	In the past	Black	OSCC
PD56002a	60	Male	Ex-smoker	In the past	Black	OSCC
PD56003a	75	Female	No	No	Black	OSCC
PD56004a	54	Male	Ex-smoker	In the past	Black	OSCC
PD56005a	62	Male	No	No	Black	OSCC
PD56006a	56	Male	Yes	In the past	Black	OSCC
PD56007a	60	Male	Ex-smoker	In the past	Black	OSCC
PD56008a	65	Male	Ex-smoker	In the past	Black	OSCC
PD56009a	73	Male	Ex-smoker	In the past	Black	OSCC
PD56010a	65	Male	No	No	Black	OSCC
PD56011a	68	Male	No	In the past	Black	OSCC
PD56012a	60	Female	No	No	Black	OSCC
PD56013b	54	Male	Ex-smoker	In the past	Black	OSCC
PD56014b	89	Male	Ex-smoker	No	Black	OSCC
PD56015b	86	Female	No	No	Black	OSCC
PD56016b	59	Male	Ex-smoker	In the past	Black	OSCC
PD56017b	56	Female	Ex-smoker	In the past	Black	OSCC
PD56018b	59	Male	Yes	In the past	Black	OSCC
PD56019b	61	Female	No info	No info	Black	OSCC
PD56020b	82	Male	No	No	Black	OSCC
PD56021b	66	Male	Yes	In the past	Black	OSCC
PD56022b	69	Male	Yes	In the past	Black	OSCC
PD56023b	84	Female	No	No	Black	OSCC
PD56024b	81	Male	No	No	Black	OSCC
PD56025b	62	Male	Yes	Yes	Black	OSCC
PD56026b	64	Male	Yes	In the past	Black	OSCC
PD56027b	66	Male	Yes	In the past	Black	OSCC
PD56028b	64	Female	No	No	Black	OSCC
PD56029b	51	Female	No	No	Black	OSCC
PD56030b	60	Female	No	No	Black	OSCC
PD56031b	52	Female	No info	No info	Black	OSCC

This table provides a concise summary of the demographic and clinical characteristics of the cohort analysed using WES for OSCC tumours. It includes details on sample size, site of collection, gender distribution, age, smoking and alcohol consumption histories, and race distribution among the patients studied.
UCT – University of Cape Town, WITS - University of the Witwatersrand, OSCC – oesophageal squamous cell carcinoma

To investigate the relationship between gene expression and survival outcome in chapter 3, we evaluated 79 OSCC tumour-normal pairs, distinct from those analysed using WGS and WES. The cohort comprised 36 females and 43 males, with a median age of 59, ranging between 28–

82 years. For smoking status, 35 patients smoked in the past and 25 patients were smokers at the time of diagnosis and 19 patients were non-smokers. All tumours were histologically confirmed as OSCC, of these, 10 tumours were poorly differentiated, 54 were moderately differentiated, 6 were well-differentiated and in 9 cases the differentiation status was unknown. There were 47 black patients and 32 mixed ancestries. All patients were followed until either death or the last follow-up date. During the follow-up period, 46 patients were reported to have passed on, while 10 patients were alive, and 23 patients failed to follow up on their last follow up (Table 6.3).

Table 6.3 A summary of the study samples for gene expression and Kaplan Meier curve analysis.

Groote Schuur Hospital, UCT, Cape Town

Patient Number	Age	Sex	Smoking status	Differentiation	Ethnicity	Survival	Histology
279	43	Male	In the past	Moderate	Mixed ancestry	LTF	OSCC
292	62	Female	Yes	Moderate	Mixed ancestry	Dead	OSCC
299	65	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
329	71	Female	Never	Moderate	Black	Dead	OSCC
380	53	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
389	52	Female	In the past	No info	Black	LTF	OSCC
392	56	Male	In the past	Well	Black	LTF	OSCC
411	79	Male	In the past	Moderate	Black	LTF	OSCC
412	58	Female	Yes	Moderate	Mixed ancestry	Dead	OSCC
413	68	Male	Yes	No info	Mixed ancestry	LTF	OSCC
416	81	Male	Yes	Poor	Mixed ancestry	LTF	OSCC
417	54	Female	Never	Moderate	Black	Dead	OSCC
422	51	Female	Yes	Moderate	Mixed ancestry	LTF	OSCC
425	45	Male	Yes	No info	Mixed ancestry	Dead	OSCC
426	47	Male	Never	Poor	Black	Dead	OSCC
427	66	Female	Never	Moderate	Black	Dead	OSCC
431	64	Male	In the past	Poor	Black	LTF	OSCC
435	79	Female	In the past	Well	Black	LTF	OSCC
437	52	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
438	78	Male	Never	Poor	Mixed ancestry	Dead	OSCC
439	57	Female	Never	Poor	Black	Dead	OSCC
440	59	Female	Never	Moderate	Black	Dead	OSCC
441	69	Male	In the past	No info	Black	LTF	OSCC
443	55	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
444	60	Male	Yes	Moderate	Black	LTF	OSCC
445	76	Male	In the past	Moderate	Mixed ancestry	Dead	OSCC
449	59	Male	In the past	Moderate	Black	Dead	OSCC
452	59	Female	In the past	Moderate	Black	Dead	OSCC
453	56	Male	Never	Moderate	Black	LTF	OSCC
455	50	Male	In the past	Well	Black	Dead	OSCC
465	61	Female	Never	Moderate	Black	LTF	OSCC
467	66	Female	In the past	No info	Black	Dead	OSCC
474	57	Female	Never	Moderate	Black	Dead	OSCC
481	63	Male	In the past	Moderate	Mixed ancestry	Dead	OSCC
482	81	Female	In the past	Moderate	Black	LTF	OSCC
493	59	Male	In the past	Moderate	Black	Dead	OSCC
495	53	Female	In the past	Poor	Black	LTF	OSCC
497	75	Female	In the past	Moderate	Black	LTF	OSCC
498	62	Female	Never	Moderate	Black	LTF	OSCC

Patient Number	Age	Sex	Smoking status	Differentiation	Ethnicity	Survival	Histology
501	68	Female	In the past	No info	Black	Dead	OSCC
503	68	Female	Never	Moderate	Black	Dead	OSCC
505	57	Female	Never	Moderate	Black	Dead	OSCC
507	78	Female	In the past	Moderate	Black	Dead	OSCC
508	59	Male	In the past	Moderate	Mixed ancestry	Dead	OSCC
511	54	Male	In the past	Moderate	Black	LTF	OSCC
512	67	Female	In the past	Moderate	Black	Dead	OSCC
514	55	Male	In the past	Moderate	Black	Dead	OSCC
515	71	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
519	45	Female	In the past	Poor	Mixed ancestry	Dead	OSCC
522	48	Female	Never	Moderate	Black	Dead	OSCC
524	69	Male	Yes	No info	Mixed ancestry	Dead	OSCC
525	52	Male	Yes	Moderate	Black	LTF	OSCC
526	53	Male	In the past	Moderate	Black	LTF	OSCC
527	62	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
528	55	Male	In the past	Poor	Black	Dead	OSCC
529	82	Female	In the past	Well	Mixed ancestry	Dead	OSCC
530	75	Female	In the past	Moderate	Black	LTF	OSCC
531	76	Male	In the past	Poor	Mixed ancestry	Dead	OSCC
532	56	Female	In the past	Moderate	Black	LTF	OSCC
533	45	Male	Yes	Moderate	Mixed ancestry	Dead	OSCC
534	64	Male	In the past	Poor	Mixed ancestry	Dead	OSCC
535	54	Male	Yes	Moderate	Black	Dead	OSCC
538	61	Female	Never	Moderate	Black	Dead	OSCC
539	66	Male	Yes	Moderate	Mixed ancestry	LTF	OSCC
540	62	Female	Yes	Moderate	Black	Dead	OSCC
541	52	Female	In the past	No info	Mixed ancestry	Dead	OSCC
580	28	Female	Never	Moderate	Black	Alive	OSCC
590	58	Male	Yes	Moderate	Black	Dead	OSCC
595	61	Female	In the past	Moderate	Black	Alive	OSCC
602	43	Male	Never	Moderate	Mixed ancestry	Alive	OSCC
616	58	Male	Yes	Moderate	Mixed ancestry	Alive	OSCC
627	52	Male	Yes	Moderate	Mixed ancestry	Alive	OSCC
634	57	Female	Yes	Moderate	Mixed ancestry	Alive	OSCC
635	57	Male	In the past	Moderate	Mixed ancestry	Alive	OSCC
636	41	Female	Never	Well	Black	Dead	OSCC
638	69	Male	In the past	Well	Mixed ancestry	Dead	OSCC
648	60	Male	Ye	Moderate	Mixed ancestry	Alive	OSCC
653	64	Male	Yes	Moderate	Black	Alive	OSCC
657	59	Female	Never	No info	Black	Alive	OSCC

Groote Schuur Hospital, UCT, Cape Town

This table provides a comprehensive overview of the demographic characteristics, clinical features, and follow-up outcomes of the study cohort used for gene expression analysis and Kaplan Meier curve analysis in the context of OSCC. It includes details on sample size, gender distribution, age range, smoking status, histological differentiation, ethnicity, follow-up outcomes.UCT – University of Cape Town, LTF – lost to follow up, OSCC – oesophageal squamous cell carcinoma

6.1.1.3 Ethics and consent

Ethical approval for the project was obtained from the UCT/Groote Schuur Hospital Human Research Ethics Committee (Ethics number: 040/2005).

6.2 Methods

6.2.1 Genomic DNA and RNA extraction

6.2.1.1 Patient blood processing

Blood samples collected in EDTA tubes were stored under different conditions: at 4°C for a few days, or at -20°C or -80°C for a few weeks, before genomic DNA extraction. Upon collection, blood was collected in EDTA tubes and centrifuged at 2000xg for 10 minutes to separate the blood into three layers. The upper clear-to-pale-yellow plasma layer was aspirated using a disposable transfer pipette, and stored in 1.5 ml aliquots at -80°C. The middle grey-white buffy coat interphase layer was removed using a transfer pipette and transferred to a 1.5 ml cryovial for storage, also at -80°C. Finally, the dark red bottom layer was removed and aliquoted into 3-4 cryotubes for storage at -80°C.

6.2.1.2 DNA extraction from blood samples

Genomic DNA was extracted as previously described [414]. Briefly, the buffy coat interphase layer samples were defrosted at room temperature and transferred to a sterile polypropylene 50 ml tube and diluted with 2 volumes of 1x phosphate buffered saline (PBS), and centrifuged at 3000xg for 15 min. The supernatant was decanted, and the pellet suspended in 25 ml of Sucrose Triton X-100 Lysing Buffer and vortexed to mix thoroughly. Tubes were then placed on ice for 5 minutes, followed by centrifugation for 5 minutes at 3000xg and the decanted. The pellet was resuspended in 3 ml of T20E5 solution (20 mM Tris-HCl, 5 mM EDTA), 300µl of 10% Sodium dodecyl sulphate (SDS) and 75 µl of 10 mg/ml proteinase K and incubated at 45°C overnight. The following day, 1 ml of saturated NaCl was added, mixed vigorously for 15 seconds and the samples centrifuged for 40 min at 3200xg. The supernatant containing the DNA was transferred to a clean tube and 2 volumes of absolute ethanol was added to pellet the DNA. The tube was agitated gently and centrifuged for 30 minutes at 3000xg at 4°C to precipitate the DNA. The pellet was washed in 1 ml of 70% ice-cold ethanol, tubes were centrifuged at 7000xg for 5 min and pellets (DNA) were dissolved in 400 µl of 1X Tris-EDTA buffer by shaking overnight at 4°C. The following day the samples were vortexed gently and incubated at 60°C for a further 10 minutes to ensure the complete dissolution of the DNA. DNA was then quantitated on a NanoDrop 2000/2000c Spectrophotometer (Thermo Scientific, IL, USA) and the samples stored at –20°C until needed.

6.2.1.3 DNA extraction from tissue biopsies

Tissue biopsies collected from patients were stored at -80°C until ready for DNA extraction. Purification of DNA was performed using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit (Qiagen, 80224, Hilden, Germany) as described by the manufacturers. Tubes containing frozen biopsy samples were thawed on ice, biopsies were weighed, transferred to a sterile P2 hood and dissected into several small pieces using a sharp surgical blade. A section was removed for haematoxylin and eosin staining. Tissue samples were disrupted and homogenised at room temperature with lysis buffer (Buffer RLT) containing β-mercaptoethanol in a tissue rupture probe at full speed until the tissue was uniformly homogenised. Lysates were then centrifuged at room temperature and the supernatant transferred to an AllPrep DNA spin column placed in a 2 ml collection tube and centrifuged at 20 000xg for 30 seconds. The flow-through was set aside for RNA extraction (see section 6.2.1.4). A further 300 µl lysis buffer (Buffer RLT) was then added to the spin column followed by centrifugation. The spin column was placed in a clean 2ml collection tube, 350 µl of wash buffer 1 (Buffer AW1) was added and centrifuged at 20 000xg for 30 seconds to wash the membrane. The flow through was discarded. The DNA spin column was transferred to a new 2 ml collection tube and 80 µl of Proteinase-K/Buffer wash buffer 1 (Buffer AW1) mix was added and the tube incubated for 5 minutes at room temperature. The spin column was then centrifuged with wash buffer 2 (Buffer AW2) at 20 000xg for 30 seconds and transferred to a clean 2ml collection tube followed by the addition of 500 µl wash buffer 2 and centrifuged for 2 minutes. Flow through was discarded. Finally, the DNA spin column was placed in a new 1.5 ml microfuge tube and 100 µl of elution buffer (Buffer EB) was added directly onto the centre of the spin column membrane. The column was incubated at room temperature for 1 minute and centrifuged at 8000xg for 1 minute to elute the DNA. This step was repeated with a further 50-100 µl elution buffer. DNA yield was calculated 260nm using a NanoDrop2000 spectrophotometer, and samples were stored at -20°C until further use.

6.2.1.4 RNA extraction from tissue biopsies

The flow-through obtained from the initial steps in section 2.2.1.3 was used for RNA extraction according to manufacturer's instructions (Qiagen AllPrep DNA/RNA/miRNA Universal Kit). 50 - 80 µl Proteinase K was added (depending on the lysate volume) to the flow-through together with 200 - 350 µl of 100% ethanol and mixed well. Samples were incubated at room temperature for 10 minutes followed by the addition of 400 – 750 µl of 100% ethanol and

mixed well. Up to 700 µl of the sample, including any precipitate that may have formed, was then transferred to a RNeasy® spin column placed in a 2 ml collection tube (supplied with kit) and centrifuged at 20 000xg for 15 seconds. This step was repeated until the all the lysate was used. 500 µl of wash buffer (Buffer RPE) was added to the spin column and centrifuged at 20 000xg for 15 seconds. 10 µl DNase 1X stock solution was added to 70 µl of digestion buffer (Buffer RDD) and mixed gently by inverting the tube, 80 µl of the mix was added directly to the spin column membrane and incubated at room temperature for 15 minutes. 500 µl RNA buffer (Buffer FRN) was added to the spin column and centrifuged at 20 000xg for 15 seconds and the flow through was set aside. The spin column was placed in a new 2 ml collection tube and the flow-through reapplied to the column and centrifuged for 15 seconds at 20 000xg. 500 µl wash buffer (Buffer RPE) was added to the RNeasy spin column and centrifuged for 15 seconds followed by the addition of 500 µl 100% ethanol to the spin column and re-centrifuged for 2 minutes at 20 000xg. The spin column was then placed in a new 1.5 ml collection tube and 30-50 µl of RNAase free water was added directly to the spin column membrane. The sample was centrifuged for 1 minute at 8000xg to elute the RNA. This step was repeated to elute further RNA by reapplying the eluate to the column. RNA yield was calculated at 260nm using a NanoDrop 2000 spectrophotometer, and samples were stored at -20°C until use.

6.2.1.5 Preparation of agarose gels for electrophoresis

1% Agarose gels were prepared by adding 5g agarose powder (SeaKem®, Lonza, Rockland, ME, USA) in 500 ml 1X TBE (Tris-Borate-EDTA) and the solution was dissolved by heating in a microwave oven for 5-10 min and cooled down to 40-50°C before pouring into the gel tank and inserting the combs after it had cooled to room temperature. 1X TBE electrophoresis buffer was added to the tank.

6.2.1.6 DNA integrity and quantification

50 ng of DNA was electrophoresed on a 1% agarose gel (SeaKem®, Lonza, Rockland, ME, USA) together with 1 µl Novel Juice (Bio-Helix, Taipei, Taiwan) detection dye and ddH_2O to make 6 µl. A gene-ladder was loaded into the gel (GeneRuler ™ 100bp Plus 40 DNA Ladder (ThermoFisher, Vilnius, Lithuania). Gels were immersed in 1X TBE buffer in a gel-electrophoresis system (ADVANCE Mupid®-One 077388, Tokyo, Japan) and electrophoresed for 35 minutes at 100V. Gels were then examined under ultraviolet light using UVP BioSpectrum ImagingSystem (UVP, USA).

6.2.1.7 RNA integrity and quantification

100 ng of RNA was electrophoresed on a 1% agarose gel (SeaKem®, Lonza, Rockland, ME, USA) with 5µl Formaldehyde loading dye (160µl 10X MOPS-EDTA buffer, 100µl sterile ddH_2O, 100µl ethidium bromide, 160µl sterile glycerol and 160µl saturated bromophenol blue in ddH_2O) and ddH_2O to make 10 µl. Prior electrophoresis, sample was heated at 60°C for 10 min and cooled at room temperature. Gels were immersed in 1X TBE buffer (Tris-Borate-EDTA) containing 0.5 µg of Ethidium Bromide (EthBr)/ ml of gel solution from stock solution (10 mg/ml) and electrophoresed for 90 minutes at 100V. Gels were examined under ultraviolet light using UVP BioSpectrum ImagingSystem (UVP, USA) to visualise the RNA. A good RNA sample should have sharp, clear 28S and 18S rRNA bands and the 28S rRNA band should be approximately twice as intense as the 18S rRNA band.

6.2.2 DNA Sequencing and data analysis

6.2.2.1 Whole genome sequencing and Whole exome sequencing

Thirty-one DNA samples and sixty-seven RNA samples isolated from paired tumour biopsies and blood samples were subjected to WGS and WES, respectively. The library preparation, capture, and sequencing were performed at the Wellcome Sanger Institute in Cambridge, UK. Samples were genotyped for single nucleotide polymorphisms (SNP) using a Fluidigm chip array to confirm that the tumour and normal samples were matched. Samples were sequenced on an Illumina HiseqX10 using 150 bp paired end reads to a depth 40x coverage for WGS and 92× for WES. Raw pair-end sequencing reads were aligned with Burrows-Wheeler Aligner (BWA) [415] to the Genome Reference Consortium Human Build 38 (GRCh38, also known as hg38) or Genome Reference Consortium Human Build 37 (GRCh37, also known as hg19) for WGS or WES, respectively. Somatic single base somatic substitutions were called using *CaVEMan* (https://github.com/cancerit/CaVEMan), an expectation maximisation–based somatic substitution detection algorithm [288]. *CaVEMan* compares sequencing reads from tumour and matched normal samples and uses a naive Bayesian model and expectation-maximization approach to calculate the probability of a somatic variant at each base (https://github.com/cancerit/CaVEMan). Small insertion and deletion (indel) detection was performed using the *cgp-pindel* pipeline https://github.com/genome/pindel) [289]. After calling the full set of variants, off-target variants and false positive variants were filtered with a set of standard CaVEMan filters (Table 6.4). Post-processing filters required that the following criteria were met to call a somatic substitution:

Table 6.4 Summary of CaVEMan filters for variant calling and filtering criteria.

Filter name	id	description
depth	DTH	Less than 1/3 mutant alleles were >= 25 base quality
readPosition	RP	Coverage was less than 8 and no mutant alleles were found in the first 2/3 of a read (shifted 0.08 from the start and extended 0.08 more than 2/3 of the read length
matchedNormal	MN	More than 0.05 of mutant alleles that were >= 15 base quality found in the matched normal
pentamericMotif	PT	Mutant alleles all on one direction of read (1rd allowed on opposite strand) and in second half of the read. Second half of read contains the motif GGC[AT]G in sequenced orientation and the mean base quality of all bases after the motif was less than 20
avgMapQual	MQ	Mean mapping quality of the mutant allele reads was < 21
simpleRepeat	SR	Position falls within a simple repeat using the supplied bed file
centromericRepeat	CR	Position falls within a centromeric repeat using the supplied bed file
phasing	PH	Mutant reads were on one strand (permitted proportion on other strand: 0.04), and mean mutant base quality was less than 21
hiSeqDepth	HSD	Position falls within a high sequencing depth region using the supplied bed file
germlineIndel	GI	Position falls within a germline indel using the supplied bed file
unmatchedNormalVcf	VUM	Position has >= 3 mutant allele present in at least 1 percent unmatched normal samples in the unmatched VCF
singleEnd	SE	Coverage is >= 10 on each strand but mutant allele is only present on one strand
matchedNormalProportion	MNP	Tumour sample mutant allele proportion - normal sample mutant allele proportion < 0.2

A summary of the variant calling and filtering criteria applied after calling the full set of variants. Off-target and false positive variants were filtered using standard CaVEMan filters. Post-processing filters were subsequently applied to call somatic substitutions, ensuring rigorous criteria were met. These criteria include thresholds related to base quality, read coverage, allele presence in matched normal samples, read directionality, motif presence, mapping quality, region specificity (such as simple repeats, centromeric regions, high-depth regions, and germline insertions/deletions), and comparison with unmatched normal samples.

6.2.2.2 Mutation signatures analysis

Single nucleotide variants (SNVs) in all tumours were annotated based on the 96 possible trinucleotide context substitutions, encompassing 6 types of substitutions multiplied by 4 possible flanking 5′ bases and 4 possible flanking 3′ bases. The contribution of mutation signatures (single base substitutions, SBS) was estimated using the R package MutationalPatterns [290], validated with deconstructSigs [291], and the hierarchical Dirichlet process (HDP) method [292]. VCF files served as input, and the mutations_from_vcf() function was employed to extract the base substitutions. To identify the optimal contribution of known signatures, mutation signature profiles specific to human cancers (COSMIC v.2 and version 3.3) as defined by Alexandrov et al., [230] were downloaded from the respective website. The similarity between the mutational profiles of our samples and the COSMIC signatures was calculated using the cos_sim_matrix() function and visualized using the plot_cosine_heatmap() function.

6.2.2.3 Identification of significantly mutated genes

Driver genes were selected using the dNdScv algorithm [149] (v0.0.1.0, https://github.com/im3sanger/dndscv) implemented in the dNdScv R package, a suite of maximum-likelihood dN/dS methods designed to search for genes with significant recurrent mutations and quantify selection in cancer and somatic evolution. dNdScv estimates the ratio

of non-synonymous to synonymous mutations across genes, controlling for the sequence composition of the gene and the mutation signatures, using trinucleotide context-dependent substitution matrices to avoid common mutation biases affecting dN/dS. Values of dN/dS significantly higher than 1 indicate an excess of nonsynonymous mutations in that particular gene and therefore imply positive selection, whereas dN/dS values significantly lower than 1 suggest negative selection. For this analysis, genes with false discovery rate (FDR) $q \leq 0.05$ were declared to be significantly mutated. Using an FDR of $q \leq 0.05$ ensures that the expected fraction of false positives in our analysis does not exceed 5%. This well-established statistical procedure allows one to increase statistical power to detect true positives, while controlling the proportion of false positives [293].

6.2.2.4 Genomic data visualizations and interpretation

The genomic data from both WGS and WES were summarized and plotted to illustrate the frequency of mutations in genes across cohorts. This visualization order emphasizes frequently mutated genes and depicts the distribution of mutations among the samples. These analyses were conducted using the Genomic Visualizations in R (GenVisR) package [295] in R software (https://github.com/griffithlab/GenVisR). The top 70 frequently mutated genes identified through WGS and WES are listed in Table 6.5.

Table 6.5 Top 70 frequently mutated genes by WGS and WES.

WGS		
1. TP53	24. PTPRD	47. FLG
2. AHNAK2	25. PKHD1L1	48. DMD
3. MUC4	26. PIK3CA	49. CSMD3
4. CDKN2A	27. PCDHA11	50. CSMD1
5. TTN	28. OBSCN	51. CNTNAP2
6. AHNAK	29. NF1	52. CNNM2
7. NOTCH1	30. NACA	53. CHD8
8. PCLO	31. MUC5B	54. CACNA1C
9. KMT2D	32. MUC5AC	55. C3
10. PLEC	33. MAGEC1	56. AMY2B
11. FAT2	34. LRP1B	57. ADGRV1
12. USH2A	35. LAMA5	58. ADGRB3
13. RYR2	36. IGFN1	59. ZNF316
14. PKD1L1	37.HSPG2	60. ZNF268
15. MUC16	38. HMCN2	61. ZFHX3
16. FNDC1	39. HECTD4	62. ZC3H4
17. DNAH14	40. GOLGA3	63. VWA5A
18. CDK11A	41. FSIP2	64. UBR5
19. VPS13B	42. FLT4	65. TRIOBP
20. TNIK	43. FLG	66. TOX4
21. TNC	44. DMD	67. TMEM266
22. TCHH	45. CSMD3	68. TMEM132D
23. SCN2A	46. FLT4	69. TG
		70. TAS2R43

WES		
1. TP53	24. NRXN1	47. PTCH1
2. TTN	25. MUC17	48. LRP1
3. NOTCH1	26. KMT2C	49. LAMA3
4. MUC16	27. KDM6A	50. KIAA1244
5. NFE2L2	28. DOCK1	51. KCNA4
6. KMT2D	29. DNM1L	52. HSPG2
7. CSMD3	30. DMD	53. FCGBP
8. PCLO	31. CACNA1C	54. FBN2
9. AHNAK2	32. AHNAK	55. FAT1
10. ZFHX4	33. ZNF208	56. DSP
11. OBSCN	34. USP34	57. DPP6
12. MUC12	35. UNC80	58. DNAH5
13. FAT3	36. ST18	59. DCAF4L2
14. BIRC6	37. SI	60. ASXL3
15. MYO18B	38. SHANK2	61. ANK3
16. MROH2B	39. RYR1	62. ACE
17. LRP1B	40. RTEL1	63. ZNF236
18. CTNND2	41. RP1	64. XIRP2
19. ABCC1	42. PTPRQ	65. USH2A
20. ZFHX3	43. PTCH1	66. UBR4
21. RYR2	44. LRP1	67. TRPM3
22. RELN	45. LAMA3	68. TRANK1
23. PLEC	46. PTPRQ	69. TENM2
		70. SLITRK4

The top 70 frequently mutated genes identified through WGS and WES. These genes were prioritized based on their mutation frequency across the studied cohorts. WGS - Whole Genome Sequencing, WES - Whole Exome Sequencing.

6.2.2.5 Transitions and transversions rates analysis

Transitions and transversions rates for both WGS and WES cohorts were analysed using Genomic Visualizations in R (GenVisR) package [295] of the R software (https://github.com/griffithlab/GenVisR).

6.2.2.6 Pathway analysis

The enrichment analysis was performed against the Reactome (human) pathway database webtool (version 86), (web link: https://reactome.org/PathwayBrowser/) [294]. UniProt identifiers for the top 70 frequently mutated genes (Table 6.5) were used for the mapping in this analysis. The Reactome database is curated by expert biologists and maintained by Reactome's team of editors, ensuring high-quality data integration from various sources such as NCBI, Ensembl, UniProt, KEGG, ChEBI, PubMed, and GO. REAC analysis was performed as described previously [294, 416]. The analysis parameters were set as follows: The analysis parameters were set as follows: Homo sapiens (human) as the organism, with exclusion of interactors from the IntAct database to focus on manually curated and biologically significant pathways. Enriched biological pathways were identified based on statistical significance (p-value < 0.05, FDR < 0.1). Pathways enriched from the gene list are represented in a colour scale (from olive to yellow) indicating the corrected probability (FDR). The top 25 sub-pathways affected by these mutations were specifically noted in the analysis.

6.2.2.7 Validation of bioinformatics data by PCR product sequencing

Validation studies were performed to determine precision, accuracy, sensitivity, specificity, and limit of detection of the bioinformatics data by using PCR product sequencing. Samples with known mutations in *CDKN2A*, *NFE2L2*, *PIK3CA*, *ERCC6*, *TOPBP1* and *C20orf196* found by WGS were PCR amplified using Taq DNA polymerase with 50 ng of genomic DNA as template. PCR samples and gel band extracts were subjected to Sanger sequencing to verify the selected mutations identified from the WES data. The primers are listed in Table 6.6.

Table 6.6 Mutation validation and mutation screening primers.

Gene	Primer sequence	Annealing temp. (°C)
CDKN2A	FW 5′- AGC TTC CTT TCC GTC ATG C -3′	58
	RV 5′- GGA AGC TCT CAG GGT ACA AAT TC -3′	
NFE2L2	FW 5′- TAA TCT CCC CAC TTC CCA CC -3′	58
	RV 5′- GAA AGG CAA AGC TGG AAC TC -3′	
PIK3CA	FW 5′- GCT TTT TCT GTA AAT CAT CTG -3′	58
	RV 5′- CAT TTA ATG TGC CAA CTA CC-3′	
TOPBP1	FW 5′- CCT TAG ACA AGT TAC CCA TG -3′	60
Set 1	RV 5′- CTG CTG TTA GGC TGT ATT AG -3′	
TOPBP1	FW 5′- ACC CTC TTG TCT TGC CTG C -3′	58
Set 2	RV 5′- AAG TGG TTG CTA GAG TGT TTC AG -3′	
ERCC6	FW 5′- GCT GGA CAG AGA CTC TAA CTA C -3′	60
Set 1	RV 5′- CAC ATC GAC CGA CAT CAG -3′	
ERCC6	FW 5′- CCT GGC TGG GTC TTT CTC -3′	58
Set 2	RV 5′- GGA TCA GAT GTT CAG ACA CC -3′	
C20orf196	FW 5′- GAT CCT GCT ATG TGG TGC C -3′	62
Set 1	RV 5′- TGT CCT TTG AAG TTC CTC CTG -3′	
C20orf196	FW 5′- CTC GTT GTG AGC AAG CAA G -3′	58
Set 2	RV 5′- AGG GCA TAC TTT TGG CTC TCC -3′	

6.2.3 Cell Culture

6.2.3.1 Cell lines

Seven OSCC cell lines were used in this study including WHCO1, WHCO5 and WHCO6 cells originally established from South African patient biopsy samples of oesophageal squamous cell carcinoma [417]. KYSE30, KYSE150, KYSE180 and KYSE450 cell lines were Japanese derived. These cell lines were obtained from the American Type Culture Collection (ATCC). The non-cancerous oesophageal cell line EPC2 a telomerase immortalised epithelial cell line was used as a control cell line. EPC2 cells was a kind gift from Dr. Anil Rustgi (University of Pennsylvania, Philadelphia, PA, USA) [418].

6.2.3.2 Cell culture and maintenance

All cell lines were cultured in a humidified environment containing 5% CO_2 at 37˚C. OSCC cell lines were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% foetal bovine serum (FBS) and 1% of penicillin (100 U/ ml) and streptomycin (100 µg/ ml) (complete medium). The non-cancerous EPC2 cell line was maintained in Keratinocyte Serum-Free Median (KSFM) supplemented with 50 µg/ml Bovine Pituitary Extract (BPE), 1 ng/ml Epidermal Growth Factor (EGF) and 1% penicillin-streptomycin. The medium was replaced every 2 to 3 days by aspirating the used medium and washing the cells

with 10ml of pre-warmed 1X Phosphate Buffered Saline (PBS) and adding fresh complete medium. Cells were sub-cultured every 4 to 5 days or when 80% confluent by discarding the used medium and washing the cells with 1X PBS. Cells were trypsinized with 0.05% trypsin-EDTA for 2-3 minutes at 37°C and once they had detached, the trypsin-EDTA was inactivated by adding an equal volume of complete medium. Cells were pelleted by centrifugation at 4000 rpm for 3 minutes, re-suspended in 3 ml of complete medium and 1 ml aliquots were re-plated in fresh plates.

6.2.3.3 Freezing and thawing cells

Frozen cells were collected from long-term storage in liquid nitrogen and thawed in a 37°C water bath. The vials were wiped with 70% ethanol, cells were transferred to 12 ml Falcon tubes, containing 5 ml of fresh growth media and centrifuged at 4000 rpm for 3 minutes. The supernatant was aspirated, and the cell pellet were then resuspended in fresh complete medium and plated in a 100 mm dish containing 10ml of complete medium. For long term storage, cells were resuspended in freezing down media (70% DMEM, 20 % FBS and 10% DMSO) (v/v) by gentle pipetting action. The cells were aliquoted into 2 ml cryotubes and stored at -80°C for 48 hours before being transferred to liquid nitrogen.

6.2.3.4 DNA extraction from cell lines

Extraction of genomic DNA from cell lines was performed as described by Strauss (1998) [419]. Cells were harvested by trypsinisation and pelleted by centrifugation, as previously described [419]. The cell pellets were resuspended and washed in 10 ml of ice-cold 1X PBS. After centrifugation, the supernatant was discarded. The pellets were resuspended in 600 µl of digestion buffer (100 mM NaCl, 10 mM Tris pH8, 25 mM Na2EDTA pH8, 0.5% SDS and 0.1 mg/ml Proteinase K). Samples were incubated overnight in a shaking water bath at 50°C. The following day, an equal volume (600 µl) of phenol:chloroform:isoamylalcohol (25:24:1) was added and contents transferred to 2 ml microcentrifuge tubes. The tubes were vortexed to mix the contents thoroughly and centrifuged at 14 000 rpm for 10 minutes at 4°C to separate the phases. The clear upper aqueous phase was transferred to a clean 1.5 ml microcentrifuge tube and one-half volume of 7.5 M ammonium acetate (pH5.5) and two volumes of ice-cold 100% ethanol was added. DNA was pelleted by centrifugation at 14 000 rpm for 30 minutes. The pellet was washed with 1 ml of 70% ethanol, vortexed gently and centrifuged at 14 000 rpm for 10 minutes. The supernatant was removed, the DNA pellets airdried for 60 minutes at room temperature and resuspended in 50-100 µl TE-buffer. The DNA yield was measured using the NanoDrop 2000 spectrophotometer. DNA was stored at -80°C until used.

6.2.3.5 RNA extraction from cell cell lines

RNA was extracted from cells using QIAzol Lysis Reagent (QIAGEN, Hilden, Germany). The medium was aspirated, and cells were washed with cold 1X PBS. Appropriate volume of QIAzol relative to the size of the plate was added onto the cells (1 ml for 10 cm plates and 0.5 ml for 6-well plates). The cells were lysed by scrapping them off the plate using a cell scrapper and transferred to 1.5 ml Eppendorf tubes. The homogenized samples were incubated for 5 minutes at room temperature. 200 µl of chloroform was added, followed by vigorous shaking for 15 seconds to mix well and incubated for 3 minutes at room temperature. The samples were centrifuged at 14000 rpm for 15 minutes at 4°C. The upper aqueous layer was transferred to a fresh labelled tube and 0.5 ml of isopropanol was added, mixed thoroughly by shaking (do not vortex) for 15 seconds and incubated at room temperature for 10 minutes. The samples were then centrifuged at 14000rpm for 20 minutes at 4°C to pellet the RNA. The supernatant was carefully removed, 1ml 75% DEPC-ethanol was added and vortexed on low for 5-10 seconds to wash the pellet thoroughly, centrifuged at 12000 rpm for 5 minutes at 4°C to re-pellet the RNA. The supernatant was carefully removed, the pellet was air dried at room temperature for 10 minutes and dissolved in 20- 30 µl DEPC treated-water by gentle pipetting and incubated at 55°C for 5 minutes. RNA was quantitated on a NanoDrop 2000/2000c Spectrophotometer (Thermo Scientific, IL, USA) and stored at -80°C until used.

6.2.4 Primer design

Specific primer pairs for the selected genes to be validated were designed using the Primer3 algorithm (version 4.1.0) [420], one of the most widely used primer designing tools. Primer sequences were checked for the presence of possible secondary structure or possible duplexes with hairpins, self-dimers and heterodimers, formation using OligoAnalyzer tool [421]. The resulting amplicons were checked for sequence similarity throughout the human genome using the Primer-BLAST tool [422].

6.2.5 Polymerase chain reaction (PCR) protocol

PCR reactions were prepared under sterile conditions on ice in 25 µl volumes as per either Table 6.7 or Table 6.8. Thermocycling was carried out using an Applied Biosystems SimpliAmp Thermo cycler machine (Applied Biosystems by Thermo Fisher Scientific, Singapore). The standard thermocycling protocol conditions are listed in Table 6.9. Magnesium, primer concentration and DNA concentration titrations or the addition of

enhancers such as DMSO and a combination of Tris, KCl and Gelatine were all tested to optimise PCR conditions. Primers and annealing temperatures are shown in Table 6.6.

Table 6.7 PCR master mix for other genes other than *CDKN2A*.

	X1(μl)
10X PCR buffer w/o $MgCl_2$	2.5
25mM $MgCl_2$	2
10mM dNTPs	0.5
Forward Primer (10uM)	1.5
Reverse Primer (10uM)	1.5
10mM Tris-HCl pH 8.3	2.5
50mM KCl	0.5
0.01% gelatin	1
Taq polymerase	0.2
DNA	2
ddH_2O	10.8
Total	25

Table 6.8 PCR master mix for *CDKN2A*.

	X1(μl)
10X PCR buffer w/o $MgCl_2$	2.5
25mM $MgCl_2$	2
10mM dNTPs	0.5
Forward Primer (10uM)	1.5
Reverse Primer (10uM)	1.5
10mM Tris-HCl pH 8.3	2.5
50mM KCl	0.5
0.01% gelatin	1
5% DMSO	1.25
Taq polymerase	0.2
DNA	2
ddH_2O	9.55
Total	25

Table 6.9 Standard PCR thermocycling conditions.

Conditions	Temperature (°C)	Time	Cycles
Initial denaturation	94	10 min	1
Denaturation Annealing Extension	94 X 72	1 min 1 min 1 min	45
Final extension	72	4 min	1
Cooling	4	Hold	

X – Annealing temperature (°C) for each primer set as specified in table 6.6.

6.2.6 Post-PCR DNA sequencing

PCR products were electrophoresed on 1% Agarose gel to confirm the product size. PCR samples including those from the gel slices were subjected to bi-directional Sanger sequencing. Chromatograms were analysed using Chromas v2.6.6 (available at http://technelysium.com.au/wp/chromas/) a free trace viewer for simple DNA sequencing projects.

6.2.7 Extraction of gel bands from 1% agarose gel

Following PCR amplification, 5 µl PCR product was electrophoresed on a 1% agarose gel (SeaKem®, Lonza, Rockland, ME, USA) together with 2.5 µl Novel Juice (Bio-Helix, Taipei, Taiwan) detection dye. Gels were immersed in 1XTBE buffer in a gel-electrophoresis system (ADVANCE Mupid®-One 077388, Tokyo, Japan) and electrophoresed for 120 minutes at 100V and visualised under UV light. Target DNA bands were excised from the gel using a sterile surgical blade and placed into sterile microcentrifuge tubes.

6.2.8 cDNA synthesis and Real-Time quantitative PCR (RT-qPCR) analysis

6.2.8.1 cDNA synthesis

Conversion of mRNA to cDNA was performed using the ImProm-II™ Reverse Transcription System (Promega, WI, USA), following manufacturer's instructions. Briefly, 0.1µg-1µg template mRNA, 1 µl oligo dT or random primer and dH_2O up to 9 µl were added together. This mixture was heated for 10 minutes at 70 ˚C to denature any secondary structure in the RNA followed by annealing to the oligo dT primers. The mixture was chilled on ice for 5 minutes.

For each sample, 16 µl of the second master mix (see Table 6.10) was added to the first master mix (above) and incubated for 2 hours at 42°C, followed by 10 minutes at 70°C to inactivate the reverse transcriptase. The cDNA was stored at 4°C until needed.

Table 6.10 cDNA synthesis master mix 2.

Reagents	Volume (µl)
5X first strand synthesis buffer	5
dNTPs mix	1
RNase inhibitor	1
$MgCL_2$	2
Impromp II Reverse Transcriptase	1
ddH2O	6
Total	16

6.2.8.2 RT-qPCR analysis

Quantitative real-time PCR was performed on a Roche Lightcycler 480 II (96- or 384-well plates), Roche, or QuantStudio 3 Real time PCR instrument (96-well 0.2ml block), (ThermoFisher Scientific, Waltham, Massachusetts, United States). To a mixture of SYBR Green PCR Master Mix (KAPA SYBR Fast qPCR Kit, KAPA Biosystems), 10 uM of forward and reverse primers, and 1 µl of cDNA was added, in a total volume of 12.5 µl (Table 6.11) for other genes. For *p14ARF* and *p16INK4a*, 5% of DMSO was added in the master mix (Table 6.12). *GAPDH* was used as an internal control. Primers and melting temperatures are shown in Table 6.13.

Table 6.11 RT-qPCR mixture set up for each gene other than *p14ARF* and *p16INK4a*.

Reagents	Volume (µl)
SYBR Green PCR Master Mix	6.25
Forward primer (10 uM)	0.5
Reverse primer (10 uM)	0.5
ddH2O	4.25
cDNA	1
Total	12.5

Table 6.12 RT-qPCR mixture set up for *p14ARF* and *p16INK4a*.

Reagents	Volume (µl)
SYBR Green PCR Master Mix	6.25
Forward primer (10 uM)	0.5
Reverse primer (10 uM)	0.5
ddH2O	3.62
cDNA	1
5% DMSO	0.63
Total	12.5

Table 6.13 qPCR primers.

Gene	Primer sequence	Annealing temp. (°C)
Exon 3 primers	FW 5'- CCC GCT TTC GTA GTT TTC AT -3' RV 5'- TTA TTT GAG CTT TGG TTC TG -3'	58
p14ARF	FW 5'- GTG GCC CTC GTG CTG ATG -3' RV 5'- AGC ACC ACC AGC GTG TCC -3'	68
p16INK4a	FW 5'- GGT GCG GGC GCT GCT GGA -3' RV 5'- AGC ACC ACC AGC GTG TCC -3'	66
WFE2L2	FW 5'-GCG ACG GAA AGA GTA TGA GC -3' RV 5'- TGG GAG TAG TTG GCA GAT CC -3'	58
PIK3CA	FW 5'- TCA GCA GTG TGG TAA AGT TC -3' RV 5'- CAG TCC AGA AGT TCC ATA GC -3'	50
AKAP4	FW 5'-CAT CAA GGC TGT TTC CGA TG -3' RV 5'- AGT CCT CAC AGT GGT AGC AC-3'	60
TGFBR1	FW 5'- CAT CTT CAA CCC CTG ACA GC -3' RV 5'- GCT CAG AGT ATT GTG TGG G -3'	60
ERCC6	FW 5'- CCA GCA GAG ACA TCA ACA GG -3' RV 5'- AGC TCT TCC GAG GCA GTC TC -3'	60
C20orf196	FW 5'- AGG CTC TGC AAA CTC ACT CTC -3' RV 5'- TTC CTT CCT CCA GTG GGT AG -3'	60
BAX	FW 5'- GGT TGT CGC CCT TTT CTA CT -3' RV 5'- AAG TCC AAT GTC CAG CCC AT -3'	60
BCL2	FW 5'- CTG CAC CTG ACG CCC TTC ACC -3' RV 5'- CAC ATG ACC CCA CCG AAC TCA AAG -3'	60
BCL-XL	FW 5'- GAT CCC CAT GGC AGC AGT AAA GCA AG -3' RV 5'- CCC CAT CCC GGA AGA GTT CAT TCA CT -3'	60
Caspase-3	FW 5'- ACA TGA CTC AGC CTG TTC C -3' RV 5'- GCC TCA CCA CCT TTA GAA C -3'	60
Caspase-9	FW 5'- GTG AAC TTC TGC CGT GAG TC -3' RV 5'- GCA AAG CCA GCA CCA TTT TC -3'	60
CCDN1	FW 5'- TGG GGT TCT AGG CAT CTC TG -3' RV 5'- CTG GAT GGT TTG TTG GGG TG -3'	60
MDM2	FW 5'- CTC ACA GAT TCC AGC TTC GG -3' RV 5'- CAG AGA AGC TTG GCA CGC -3'	60
p21	FW 5'- ACC TCA CCT GCT CTG CTG C -3' RV 5'- ATT AGG GCT TCC TCT TGG AGA -3'	60
Rb	FW 5'- GAG ACA CAA GCA ACC TCA GC -3' RV 5'- GGT GTG CTG GAA AAG GGT C -3'	60
PS3	FW 5'- CTG CTC AGA TAG CGA TGG TCT G -3' RV 5'- TTG TAG TGG ATG GTG GTA CAG TCA -3'	60
GAPDH	FW 5'- GGC TCT CCA GAA CAT CAT CC -3' RV 5'- GCC TGC TTC ACC ACC TTC -3'	60

6.2.8.3 Analysis of RT-qPCR data

Real Time PCR data were analysed using the comparative $2^{-\Delta\Delta CT}$ method for relative quantification. Results of the real-time RT-qPCR data are presented as CT values, where CT is defined as the threshold PCR cycle number at which an amplified product is first detected. The average CT was calculated for each gene evaluated and GAPDH, and the ΔCT was determined as the mean of the triplicate CT values for the evaluated gene minus the mean of the triplicate CT values for GAPDH. The ΔΔCT represents the difference between the paired tissue samples, as calculated by the formula ΔΔCT = (ΔCT of tumour - ΔCT of normal). The N-fold differential expression of the evaluated gene for a tumour sample compared with its normal adjacent tissue was expressed as $2^{-\Delta\Delta CT}$, which represents the fold change in the target gene expression in tumour normalized to an internal control gene (*GAPDH*) and relative to the normal control.

6.2.9 siRNA transfection assay

For the inhibition of gene expression, siRNA targeting either *p14ARF* or *p16INK4*a was designed from either exon 1β or exon 1α, respectively (Table 6.14). Short interfering RNAs (siRNAs) were purchased from Merck Life Science (Merck Life Science (Pty) Ltd, Modderfontein, Johannesburg, South Africa). Lipofectamine® RNAiMAX transfection reagent (ThermoFisher Scientific, Waltham, Massachusetts, United States) was used in the transient transfection assays. KYSE30 cells were transfected according to the manufacturer protocol. Briefly, cells in DMEM containing 10% FBS without antibiotics were seeded to 50-60% confluence at transfection in 6-well plates. 12 µl (120 pmol) of 10 µM siRNA were diluted in 150 µl in DMEM containing 10% FBS and was dissolved in transfection diluent (12 µl of Lipofectamine® RNAiMAX reagent in 150 µl in DMEM containing 10% FBS). The diluted siRNA was added to diluted Lipofectamine® RNAiMAX Reagent (1:1 ratio) and incubated for 5 minutes at room temperature. The siRNA-lipid complex was added to cells and incubated for 48 hours in a humidified environment containing 5% CO_2 at 37˚C. RNA extracts were harvested 48 hours post-transfection (as described in Section 6.2.3.5), and the downstream effects of cellular gene expression were examined by RT-qPCR (as described in Section 6.2.8).

Table 6.14 *p14ARF* and *p16INK4a* siRNA sequences.

Gene		siRNA sequence	Exon
p14ARF	*mRNA sequence (5'-3')*	GAGGGTTTTCGTGGTTCACATCC	1β
	siRNA sequence (Sense) (5'-3')	GGGUUUUCGUGGUUCACAU[dT][dT]	
	siRNA sequence (Antisense) (5'-3')	AUGUGAACCACGAAAACCC[dT][dT]	
p16INK4a	*mRNA sequence (5'-3')*	AACGCACCGAATAGTTACGGTCG	1α
	siRNA sequence (Sense) (5'-3')	CGCACCGAAUAGUUACGGU[dT][dT]	
	siRNA sequence (Antisense) (5'-3')	ACCGUAACUAUUCGGUGCG[dT][dT]	

6.2.10 Identifying the most deleterious missense variants

We used three distinct bioinformatics tools namely, PolyPhen-2 (Polymorphism Phenotyping v2) (http://genetics.bwh.harvard.edu/pph2/index.shtml) [405], I-Mutant v2.0 (https://folding.biofold.org/cgi-bin/i-mutant2.0.cgi) [408] and SIFT (Sorting Intolerant From Tolerant) (https://sift.bii.a-star.edu.sg) [409, 423] to predict the functional effects of the missense variants in p14ARF and p16INK4a proteins. AlphaFold2 [424] together with ColabFold [397] were used to predict the three-dimensional (3D) structures of p16INK4a protein for both wild-type and mutant 3D structures. The Chimera tool [399] was used for protein interactive visualization

6.2.11 Overall survival analysis by Kaplan-Meier

Prognostic analysis of differential expressed genes Kaplan–Meier survival analysis was employed to analyse the association between differential expressed genes and overall survival. The difference was compared by the Log-rank method. Samples without last day of follow up were not included in the Kaplan–Meier survival analysis. Kaplan–Meier survival analysis was carried out with the or Prism10 (GraphPad) software. P-value <0.05 was considered to be statistically significant.

6.2.12 Experimental statistical analyses

Statistical analyses were performed using GraphPad Prism 3.0 (GraphPad Software Incorporated, San Diego, CA, USA). The unpaired Student's t-test was performed to identify significant differences amongst studied genes' mRNA levels in the tumours and their paired normal tissues or the relationships between the levels of studied genes in untreated KYSE30 cells and *p16ARF* and *p16INK4a* knockdown KYS30 cells. Data are presented as mean ± SD. Statistical significance was computed with the test indicated in each figure legend. Statistical significance was assessed at *p < 0.05, **p < 0.01, ***p <0.001, ****p < 0.0001.

6.2.13 Preparation of buffers and reagents

All bottles and beakers were autoclaved beforehand. Distilled, autoclaved H_2O was used in all preparations.

6.2.13.1 Tissue culture solutions

Dulbecco's Modified Eagle Medium (DMEM) Culture Medium (pH 7.4)

DMEM powder was dissolved in 800ml distilled water and brought to pH 7.2 using $NaHCO_3$. The solution was filtered using a TPP "rapid" 500 0.2 μm polyethersulfone (PES) filter top in a sterile tissue culture hood. Bottles of filtered medium were incubated at 37°C for 2-3 days to check for bacterial contamination.

13.5 g (1pack) of DMEM powder

3.7 g Sodium Hydrogen Carbonate ($NaHCO_3$)

Add ddH_2O to a final volume of 1000 ml

Adjust pH to 7.4 using concentrated HCl or NaOH.

Incubate at 37°C for 2-3 days.

Store at 4°C

Penicillin/Streptomycin (1000 UI)

6.04 g Penicillin

13.16 g Streptomycin

Add ddH_2O to a final volume of 1000 ml

Filter sterilizes and store at 20°C

Complete DMEM medium

50 ml Fetal bovine serum (FBS)

450 ml DMEM cultured medium

5 ml Penicillin/Streptomycin (1000 UI)

Complete Keratinocyte Serum-Free Growth Medium (KSFM) medium

25 mg Bovine Pituitary Extract (BPE)

500 ml KSFM media

500 ng Epidermal Growth Factor (EGF)

5 ml Penicillin/Streptomycin (1000 UI)

Freezing media

70 % Fetal bovine serum (FBS)

20 % complete DMEM medium

10 % Dimethylsulphoxide (DMSO)

10X PBS (pH 7.4)

80 g NaCl

2 g KCl

14.4 g Na_2HPO_4

2.4 g KH_2PO_4

Adjust pH to 7.4 using concentrated HCl or NaOH. Bring to 1L with ddH_2O and autoclave.

1X PBS (pH 7.4)

50 ml of 10X PBS

450 ml of ddH_2O

1 X Trypsin-EDTA solution (0.05% Trypsin-0.02% EDTA)
0.5 g trypsin
0.4 g EDTA
Add 1X PBS to a final volume of 1000 ml

6.2.13.2 RNA and DNA solutions

Sterile DEPC treated water

1 ml DEPC in 1 litre ddH_2O

Incubate for 1 hour at 30°C and autoclave.

Buffer for DNA extraction from cell cultures

200 µl of 5M NaCl

100 µl of 1M Tris (pH 8)

500 µl of 0.5M EDTA (pH 8)

50 µl of 1% SDS

100 µl of Proteinase K (10mg/ml)

Add ddH_2O to a final volume of 10 ml

Phenol: Chloroform: Isoamyl alcohol (25:24:1)

Mix 50 ml of phenol, 48 ml of chloroform and 2 ml isoamyl alcohol (25:24:1)

7.5 M Ammonium acetate (pH 7.4)

Dissolve 57.81 g ammonium acetate in 60 ml of ddH_2O

Adjust pH to 7.4 with ammonia solution

Make up to 100 ml with ddH_2O and sterilise by filtration using a filter of 0.45 μm pore size.

10X TBE buffer

Dissolve 108 g Tris in 800 ml ddH_2O

Add 55 g Boric Acid (Mix)

Add 40 ml 0.5 Na_2EDTA (pH 8.0)

Adjust volume to 1 L, store at room temperature.

1% Agarose gel

5 g of Agarose powder

500 ml of 1X TBE buffer

10 % SDS

Dissolve 10 g sodium dodecyl sulphate in 80 ml ddH_2O (heat to 80°C), make up to 100 ml with ddH_2O

1 % SDS

Dissolve 1 g sodium dodecyl sulphate in 80 ml ddH_2O (heat to 80°C), make up to 100 ml with ddH_2O

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* 2021;71(3):209-249.
2. Morgan E, Soerjomataram I, Rumgay H, Coleman HG, Thrift AP, Vignat J, et al. The Global Landscape of Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma Incidence and Mortality in 2020 and Projections to 2040: New Estimates from Globocan 2020. *Gastroenterology*. 2022;163(3):649-658. e642.
3. Kachala R. Systematic Review: Epidemiology of Oesophageal Cancer in Sub-Saharan Africa. *Malawi Medical Journal*. 2010;22(3):65-70.
4. Jain S, Dhingra S. Pathology of Esophageal Cancer and Barrett's Esophagus. *Annals of Cardiothoracic Surgery*. 2017;6(2):99-109.
5. Napier KJ, Scheerer M, Misra S. Esophageal Cancer: A Review of Epidemiology, Pathogenesis, Staging Workup and Treatment Modalities. *World Journal of Gastrointestinal Oncology*. 2014;6(5):112-120.
6. Sewram V, Sitas F, O'Connell D, Myers J. Tobacco and Alcohol as Risk Factors for Oesophageal Cancer in a High Incidence Area in South Africa. *Cancer Epidemiology*. 2016;41:113-121.
7. Cunha L, Fontes F, Come J, Lobo V, Santos LL, Lunet N, et al. Risk Factors for Oesophageal Squamous Cell Carcinoma in Mozambique. *Ecancermedicalscience*. 2022;16:1437.
8. Machoki M, Saidi H, Raja A, Ndonga A, Njue A, Biomdo I, et al. Risk Factors for Esophageal Squamous Cell Carcinoma in a Kenyan Population. *Annals of African Surgery*. 2015;12(1).
9. Kaimila B, Mulima G, Kajombo C, Salima A, Nietschke P, Pritchett N, et al. Tobacco and Other Risk Factors for Esophageal Squamous Cell Carcinoma in Lilongwe Malawi: Results from the Lilongwe Esophageal Cancer Case: Control Study. *PLOS Global Public Health*. 2022;2(6):e0000135.
10. Alsop BR, Sharma P. Esophageal Cancer. *Gastroenterology Clinics*. 2016;45(3):399-412.
11. Enzinger PC, Mayer RJ. Esophageal Cancer. *New England Journal of Medicine*. 2003;349(23):2241-2252.
12. Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal Carcinoma. *The Lancet*. 2013;381(9864):400-412.
13. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Icd-10, 2008 Edition. Geneva; 2009.
14. Wakhisi J, Patel K, Buziba N, Rotich J. Esophageal Cancer in North Rift Valley of Western Kenya. *African Health Sciences*. 2005;5(2):157-163.
15. Tollefson L. The Use of Epidemiology, Scientific Data, and Regulatory Authority to Determine Risk Factors in Cancer of Some Organs of the Digestive System: 2. Esophageal Cancer. *Regulatory Toxicology and Pharmacology*. 1985;5(3):255-275.
16. African Esophageal Cancer Consortium. Expanding Oesophageal Cancer Research and Care in Eastern Africa. *Nature Reviews Cancer*. 2022;22(5):253-254.
17. Parkin DM, Bray F, Ferlay J, Pisani P. Global Cancer Statistics, 2002. *CA: A Cancer Journal for Clinicians* 2005;55(2):74-108.

18. Hendricks D, Parker MI. Oesophageal Cancer in Africa. *IUBMB Life*. 2002;53(4-5):263-268.
19. Murphy G, McCormack V, Abedi-Ardekani B, Arnold M, Camargo M, Dar N, et al. International Cancer Seminars: A Focus on Esophageal Squamous Cell Carcinoma. *Annals of Oncology*. 2017;28(9):2086-2093.
20. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global Cancer Statistics 2022: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*. 2024;74(3):229-263.
21. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today Lyon, France: International Agency for Research on Cancer; 2024 [Available from: https://gco.iarc.who.int/today.
22. 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*. 2016;1(16):e88755.
23. Codipilly DC, Qin Y, Dawsey SM, Kisiel J, Topazian M, Ahlquist D, et al. Screening for Esophageal Squamous Cell Carcinoma: Recent Advances. *Gastrointestinal Endoscopy*. 2018;88(3):413-426.
24. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in Globocan 2012. *International Journal of Cancer* 2015;136(5):E359-E386.
25. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* 2018;68(6):394-424.
26. Lipenga T, Matumba L, Vidal A, Herceg Z, McCormack V, De Saeger S, et al. A Concise Review Towards Defining the Exposome of Oesophageal Cancer in Sub-Saharan Africa. *Environment International*. 2021;157:106880.
27. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global Incidence of Oesophageal Cancer by Histological Subtype in 2012. *Gut*. 2015;64(3):381-387.
28. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer Statistics for the Year 2020: An Overview. *International Journal of Cancer*. 2021;149(4):778-789.
29. The Global Cancer Observatory. South Africa Source: Globocan 2020. Globocan 20202020.
30. Narh CT, Dzamalala CP, Mmbaga BT, Menya D, Mlombe Y, Finch P, et al. Geophagia and Risk of Squamous Cell Esophageal Cancer in the African Esophageal Cancer Corridor: Findings from the Esccape Multicountry Case-Control Studies. *International Journal of Cancer*. 2021;149(6):1274-1283.
31. Van Loon K, Mwachiro MM, Abnet CC, Akoko L, Assefa M, Burgert SL, et al. The African Esophageal Cancer Consortium: A Call to Action. *Journal of Global Oncology*. 2018;4:1-9.
32. Matejcic M, Mathew CG, Parker MI. The Relationship between Environmental Exposure and Genetic Architecture of the 2q33 Locus with Esophageal Cancer in South Africa. *Frontiers in Genetics*. 2019;10:406.
33. Dandara C, Li D-P, Walther G, Parker MI. Gene–Environment Interaction: The Role of Sult1a1 and Cyp3a5 Polymorphisms as Risk Modifiers for Squamous Cell Carcinoma of the Oesophagus. *Carcinogenesis*. 2006;27(4):791-797.

34. Vogelsang M, Wang Y, Veber N, Mwapagha LM, Parker MI. The Cumulative Effects of Polymorphisms in the DNA Mismatch Repair Genes and Tobacco Smoking in Oesophageal Cancer Risk. *PLOS One* 2012;7(5):e36962.
35. Huang FL, Yu SJ. Esophageal Cancer: Risk Factors, Genetic Association, and Treatment. *Asian Journal of Surgery*. 2018;41(3):210-215.
36. Engel LS, Chow WH, Vaughan TL, Gammon MD, Risch HA, Stanford JL, et al. Population Attributable Risks of Esophageal and Gastric Cancers. *Journal of the National Cancer Institute*. 2003;95(18):1404-1413.
37. Parkin DM, Boyd L, Walker L. 16. The Fraction of Cancer Attributable to Lifestyle and Environmental Factors in the Uk in 2010. *British Journal of Cancer*. 2011;105(S2):S77-S81.
38. Vizcaino A, Parkin D, Skinner M. Risk Factors Associated with Oesophageal Cancer in Bulawayo, Zimbabwe. *British Journal of Cancer*. 1995;72(3):769-773.
39. Mlombe Y, Rosenberg N, Wolf L, Dzamalala C, Challulu K, Chisi J, et al. Environmental Risk Factors for Oesophageal Cancer in Malawi: A Case-Control Study. *Malawi Medical Journal*. 2015;27(3):88-92.
40. McCormack V, Menya D, Munishi M, Dzamalala C, Gasmelseed N, Leon Roux M, et al. Informing Etiologic Research Priorities for Squamous Cell Esophageal Cancer in Africa: A Review of Setting-Specific Exposures to Known and Putative Risk Factors. *International Journal of Cancer*. 2017;140(2):259-271.
41. Zhang H-Z, Jin G-F, Shen H-B. Epidemiologic Differences in Esophageal Cancer between Asian and Western Populations. *Chinese Journal of Cancer*. 2012;31(6):281.
42. Morita M, Kumashiro R, Kubo N, Nakashima Y, Yoshida R, Yoshinaga K, et al. Alcohol Drinking, Cigarette Smoking, and the Development of Squamous Cell Carcinoma of the Esophagus: Epidemiology, Clinical Findings, and Prevention. *International Journal of Clinical Oncology*. 2010;15(2):126-134.
43. Shen Y, Xie S, Zhao L, Song G, Shao Y, Hao C, et al. Estimating Individualized Absolute Risk for Esophageal Squamous Cell Carcinoma: A Population-Based Study in High-Risk Areas of China. *Frontiers in Oncology*. 2021;10:598603.
44. Saadaat R, Abdul-Ghafar J, Haidary AM, Atta N, Ali TS. Esophageal Carcinoma and Associated Risk Factors: A Case-Control Study in Two Tertiary Care Hospitals of Kabul, Afghanistan. *Cancer Management and Research*. 2022:2445-2456.
45. Martincorena I, Fowler JC, Wabik A, Lawson AR, Abascal F, Hall MW, et al. Somatic Mutant Clones Colonize the Human Esophagus with Age. *Science*. 2018;362(6417):911-917.
46. 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. *Environment International*. 2021;152:106485.
47. Mmbaga EJ, Mushi BP, Deardorff K, Mgisha W, Akoko LO, Paciorek A, et al. A Case–Control Study to Evaluate Environmental and Lifestyle Risk Factors for Esophageal Cancer in Tanzaniarisk Factors for Esophageal Cancer in Tanzania. *Cancer Epidemiology, Biomarkers & Prevention*. 2021;30(2):305-316.
48. Buckle GC, Mmbaga EJ, Paciorek A, Akoko L, Deardorff K, Mgisha W, et al. Risk Factors Associated with Early-Onset Esophageal Cancer in Tanzania. *JCO Global Oncology*. 2022;8:e2100256.
49. Okello S, Akello SJ, Dwomoh E, Byaruhanga E, Opio CK, Zhang R, et al. Biomass Fuel as a Risk Factor for Esophageal Squamous Cell Carcinoma: A Systematic Review and Meta-Analysis. *Environmental Health*. 2019;18(1):1-11.

50. Li K, Yu P. Food Groups and Risk of Esophageal Cancer in Chaoshan Region of China: A High-Risk Area of Esophageal Cancer. *Cancer Investigation*. 2003;21(2):237-240.
51. De Stefani E, Brennan P, Boffetta P, Ronco AL, Mendilaharsu M, Deneo-Pellegrini H. Vegetables, Fruits, Related Dietary Antioxidants, and Risk of Squamous Cell Carcinoma of the Esophagus: A Case-Control Study in Uruguay. *Nutrition and Cancer*. 2000;38(1):23-29.
52. Launoy G, Milan C, Day NE, Pienkowski MP, Gignoux M, Faivre J. Diet and Squamous-Cell Cancer of the Oesophagus: A French Multicentre Case-Control Study. *International Journal of Cancer*. 1998;76(1):7-12.
53. Loomis D, Guyton KZ, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, et al. Carcinogenicity of Drinking Coffee, Mate, and Very Hot Beverages. *Lancet Oncology*. 2016;17(7):877-878.
54. Rapozo DC, Blanco TC, Reis BB, Gonzaga IM, Valverde P, Canetti C, et al. Recurrent Acute Thermal Lesion Induces Esophageal Hyperproliferative Premalignant Lesions in Mice Esophagus. *Experimental and Molecular Pathology*. 2016;100(2):325-331.
55. Islami F, Pourshams A, Nasrollahzadeh D, Kamangar F, Fahimi S, Shakeri R, et al. Tea Drinking Habits and Oesophageal Cancer in a High Risk Area in Northern Iran: Population Based Case-Control Study. *BMJ : British Medical Journal*. 2009;338:b929.
56. Islami F, Poustchi H, Pourshams A, Khoshnia M, Gharavi A, Kamangar F, et al. A Prospective Study of Tea Drinking Temperature and Risk of Esophageal Squamous Cell Carcinoma. *International Journal of Cancer*. 2020;146(1):18-25.
57. Dar NA, Islami F, Bhat GA, Shah IA, Makhdoomi MA, Iqbal B, et al. Poor Oral Hygiene and Risk of Esophageal Squamous Cell Carcinoma in Kashmir. *British Journal of Cancer*. 2013;109(5):1367-1372.
58. Xu W, Liu Z, Bao Q, Qian Z. Viruses, Other Pathogenic Microorganisms and Esophageal Cancer. *Gastrointestinal Tumors* 2015;2(1):2-13.
59. Kamangar F, Chow WH, Abnet CC, Dawsey SM. Environmental Causes of Esophageal Cancer. *Gastroenterology Clinics of North America*. 2009;38(1):27-57, vii.
60. Bye H, Prescott NJ, Lewis CM, Matejcic M, Moodley L, Robertson B, et al. Distinct Genetic Association at the Plce1 Locus with Oesophageal Squamous Cell Carcinoma in the South African Population. *Carcinogenesis*. 2012;33(11):2155-2161.
61. 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;40(4):513-520.
62. Gao Y, Hu N, Han X, Giffen C, Ding T, Goldstein A, et al. Family History of Cancer and Risk for Esophageal and Gastric Cancer in Shanxi, China. *BMC Cancer*. 2009;9(1):1-10.
63. Akbari MR, Malekzadeh R, Nasrollahzadeh D, Amanian D, Sun P, Islami F, et al. Familial Risks of Esophageal Cancer among the Turkmen Population of the Caspian Littoral of Iran. *International Journal of Cancer*. 2006;119(5):1047-1051.
64. Hu N, Dawsey S, Wu M, Taylor P. Family History of Oesophageal Cancer in Shanxi Province, China. *European Journal of Cancer* 1991;27(10):1336.
65. Tran GD, Sun XD, Abnet CC, Fan JH, Dawsey SM, Dong ZW, et al. Prospective Study of Risk Factors for Esophageal and Gastric Cancers in the Linxian General Population Trial Cohort in China. *International Journal of Cancer*. 2005;113(3):456-463.
66. 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. *Frontiers in Genetics*. 2019;10:642.

67. Abnet CC, Arnold M, Wei W-Q. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology*. 2018;154(2):360-373.
68. Bye H, Prescott NJ, Matejcic M, Rose E, Lewis CM, Parker MI, et al. Population-Specific Genetic Associations with Oesophageal Squamous Cell Carcinoma in South Africa. *Carcinogenesis*. 2011;32(12):1855-1861.
69. Hecht SS. Tobacco Carcinogens, Their Biomarkers and Tobacco-Induced Cancer. *Nature Reviews Cancer*. 2003;3(10):733-744.
70. Vaish R, Bajpai J. Alcohol and Cancer: Waiting for the Storm to Pass or Dancing in the Rains! *Indian Journal of Medical and Paediatric Oncology*. 2020;41(04):473-475.
71. Koop CE, Luoto J. " The Health Consequences of Smoking: Cancer," Overview of a Report of the Surgeon General. *Public Health Reports*. 1982;97(4):318.
72. Ishiguro S, Sasazuki S, Inoue M, Kurahashi N, Iwasaki M, Tsugane S, et al. Effect of Alcohol Consumption, Cigarette Smoking and Flushing Response on Esophageal Cancer Risk: A Population-Based Cohort Study (Jphc Study). *Cancer Letters*. 2009;275(2):240-246.
73. Hashibe M, Boffetta P, Janout V, Zaridze D, Shangina O, Mates D, et al. Esophageal Cancer in Central and Eastern Europe: Tobacco and Alcohol. *International Journal of Cancer*. 2007;120(7):1518-1522.
74. Zambon P, Talamini R, La Vecchia C, Dal Maso L, Negri E, Tognazzo S, et al. Smoking, Type of Alcoholic Beverage and Squamous-Cell Oesophageal Cancer in Northern Italy. *International Journal of Cancer*. 2000;86(1):144-149.
75. Yu MC, Garabrant DH, Peters JM, Mack TM. Tobacco, Alcohol, Diet, Occupation, and Carcinoma of the Esophagus. *Cancer Research*. 1988;48(13):3843-3848.
76. Prabhu A, Obi KO, Rubenstein JH. The Synergistic Effects of Alcohol and Tobacco Consumption on the Risk of Esophageal Squamous Cell Carcinoma: A Meta-Analysis. *The American Journal of Gastroenterology*. 2014;109(6):822-827.
77. Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P, et al. Genomic Analyses Reveal Mutational Signatures and Frequently Altered Genes in Esophageal Squamous Cell Carcinoma. *The American Journal of Human Genetics*. 2015;96(4):597-611.
78. White MC, Holman DM, Boehm JE, Peipins LA, Grossman M, Henley SJ. Age and Cancer Risk: A Potentially Modifiable Relationship. *American Journal of Preventive Medicine*. 2014;46(3):S7-S15.
79. DePinho RA. The Age of Cancer. *Nature*. 2000;408(6809):248-254.
80. American Cancer Society. Cancer Prevention & Early Detection: Facts & Figures: American Cancer Society; 2000.
81. Giri PA, Singh KK, Phalke DB. Study of Socio-Demographic Determinants of Esophageal Cancer at a Tertiary Care Teaching Hospital of Western Maharashtra, India. *South Asian Journal of Cancer*. 2014;3(01):054-056.
82. Wu M, Zhao J-K, Hu X-S, Wang P-H, Qin Y, Lu Y-C, et al. Association of Smoking, Alcohol Drinking and Dietary Factors with Esophageal Cancer in High-and Low-Risk Areas of Jiangsu Province, China. *World Journal of Gastroenterology*. 2006;12(11):1686.
83. Khazaei S, Ayubi E, Mansori K, Gholamaliee B, Khazaei S, Khosravi Shadmani F, et al. Geographic, Sex and Age Distribution of Esophageal Cancer Incidence in Iran: A Population-Based Study. *Middle East Journal of Cancer*. 2017;8(2):103-108.
84. Zeng H, Zheng R, Zhang S, Zuo T, Xia C, Zou X, et al. Esophageal Cancer Statistics in C Hina, 2011: Estimates Based on 177 Cancer Registries. *Thoracic Cancer*. 2016;7(2):232-237.

85. Liu X, Wang X, Lin S, Lao X, Zhao J, Song Q, et al. Dietary Patterns and the Risk of Esophageal Squamous Cell Carcinoma: A Population-Based Case–Control Study in a Rural Population. *Clinical Nutrition*. 2017;36(1):260-266.
86. Leon ME, Assefa M, Kassa E, Bane A, Gemechu T, Tilahun Y, et al. Qat Use and Esophageal Cancer in Ethiopia: A Pilot Case-Control Study. *PLOS One*. 2017;12(6):e0178911.
87. Sammon AM. Protease Inhibitors and Carcinoma of the Esophagus. *Cancer*. 1998;83(3):405-408.
88. Playford R, Woodman A, Vesey D, Deprez P, Calam J, Watanapa P, et al. Effect of Luminal Growth Factor Preservation on Intestinal Growth. *The Lancet*. 1993;341(8849):843-848.
89. Itakura Y, Sasano H, Shiga C, Furukawa Y, Shiga K, Mori S, et al. Epidermal Growth Factor Receptor Overexpression in Esophageal Carcinoma. An Immunohistochemical Study Correlated with Clinicopathologic Findings and DNA Amplification. *Cancer*. 1994;74(3):795-804.
90. Maghsudlu M, Farashahi Yazd E. Heat-Induced Inflammation and Its Role in Esophageal Cancer. *Journal of Digestive Diseases*. 2017;18(8):431-444.
91. Abedi-Ardekani B, Kamangar F, Hewitt SM, Hainaut P, Sotoudeh M, Abnet CC, et al. Polycyclic Aromatic Hydrocarbon Exposure in Oesophageal Tissue and Risk of Oesophageal Squamous Cell Carcinoma in North-Eastern Iran. *Gut*. 2010;59(9):1178-1183.
92. Mirvish SS. Role of N-Nitroso Compounds (Noc) and N-Nitrosation in Etiology of Gastric, Esophageal, Nasopharyngeal and Bladder Cancer and Contribution to Cancer of Known Exposures to Noc. *Cancer Letters*. 1995;93(1):17-48.
93. Ludwig J, Marufu L, Huber B, Andreae M, Helas G. Domestic Combustion of Biomass Fuels in Developing Countries: A Major Source of Atmospheric Pollutants. *Journal of Atmospheric Chemistry*. 2003;44(1):23-37.
94. Abdel-Shafy HI, Mansour MS. A Review on Polycyclic Aromatic Hydrocarbons: Source, Environmental Impact, Effect on Human Health and Remediation. *Egyptian Journal of Petroleum*. 2016;25(1):107-123.
95. Hakami R, Mohtadinia J, Etemadi A, Kamangar F, Nemati M, Pourshams A, et al. Dietary Intake of Benzo (a) Pyrene and Risk of Esophageal Cancer in North of Iran. *Nutrition and Cancer*. 2008;60(2):216-221.
96. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Some Non-Heterocyclic Polycyclic Aromatic Hydrocarbons and Some Related Exposures. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. 2010;92:1-853.
97. Hashimoto AH, Amanuma K, Hiyoshi K, Takano H, Masumura Ki, Nohmi T, et al. In Vivo Mutagenesis in the Lungs of Gpt-Delta Transgenic Mice Treated Intratracheally with 1, 6-Dinitropyrene. *Environmental and Molecular Mutagenesis*. 2006;47(4):277-283.
98. Finkelman RB, Belkin HE, Zheng B. Health Impacts of Domestic Coal Use in China. *Proceedings of the National Academy of Sciences of the United States of America*. 1999;96(7):3427-3431.
99. Siwińska E, Mielżyńska D, Bubak A, Smolik E. The Effect of Coal Stoves and Environmental Tobacco Smoke on the Level of Urinary 1-Hydroxypyrene. *Mutation Research*. 1999;445(2):147-153.
100. Kamangar F, Strickland PT, Pourshams A, Malekzadeh R, Boffetta P, Roth MJ, et al. High Exposure to Polycyclic Aromatic Hydrocarbons May Contribute to High Risk of Esophageal Cancer in Northeastern Iran. *Anticancer Research*. 2005;25(1B):425-428.

101. Fagundes RB, Abnet CC, Strickland PT, Kamangar F, Roth MJ, Taylor PR, et al. Higher Urine 1-Hydroxy Pyrene Glucuronide (1-Ohpg) Is Associated with Tobacco Smoke Exposure and Drinking Mate in Healthy Subjects from Rio Grande Do Sul, Brazil. *BMC Cancer*. 2006;6:1-7.
102. Chen T, Cheng H, Chen X, Yuan Z, Yang X, Zhuang M, et al. Family History of Esophageal Cancer Increases the Risk of Esophageal Squamous Cell Carcinoma. *Scientific Reports*. 2015;5(1):1-9.
103. 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 and Therapy*. 2020;13:8795-8802.
104. Hu N, Dawsey S, Wu M, Bonney G, He L, Han X, et al. Familial Aggregation of Oesophageal Cancer in Yangcheng County, Shanxi Province, China. *International Journal of Epidemiology*. 1992;21(5):877-882.
105. Yang P-W, Lin M-C, Huang P-M, Wang C-P, Chen T-C, Chen C-N, et al. Risk Factors and Genetic Biomarkers of Multiple Primary Cancers in Esophageal Cancer Patients. *Frontiers in Oncology*. 2021;10:585621.
106. Rosenberg PS, Alter BP, Ebell W. Cancer Risks in Fanconi Anemia: Findings from the German Fanconi Anemia Registry. *Haematologica*. 2008;93(4):511-517.
107. 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. *Human Genetics*. 2011;129(5):573-582.
108. Ellis A, Field J, Field E, Friedmann P, Fryer A, Howard P, et al. Tylosis Associated with Carcinoma of the Oesophagus and Oral Leukoplakia in a Large Liverpool Family—a Review of Six Generations. *European Journal of Cancer Part B: Oral Oncology*. 1994;30(2):102-112.
109. Blaydon DC, Etheridge SL, Risk JM, Hennies H-C, Gay LJ, Carroll R, et al. Rhbdf2 Mutations Are Associated with Tylosis, a Familial Esophageal Cancer Syndrome. *The American Journal of Human Genetics*. 2012;90(2):340-346.
110. Su H, Hu N, Shih J, Hu Y, Wang Q-H, Chuang EY, et al. Gene Expression Analysis of Esophageal Squamous Cell Carcinoma Reveals Consistent Molecular Profiles Related to a Family History of Upper Gastrointestinal Cancer. *Cancer Research*. 2003;63(14):3872-3876.
111. Hemminki K, Rawal R, Chen B, Bermejo JL. Genetic Epidemiology of Cancer: From Families to Heritable Genes. *International Journal of Cancer*. 2004;111(6):944-950.
112. Ghadirian P. Familial History of Esophageal Cancer. *Cancer*. 1985;56(8):2112-2116.
113. Matejcic M, Gunter MJ, Ferrari P. Alcohol Metabolism and Oesophageal Cancer: A Systematic Review of the Evidence. *Carcinogenesis*. 2017;38(9):859-872.
114. Yokoyama A, Omori T. Genetic Polymorphisms of Alcohol and Aldehyde Dehydrogenases and Risk for Esophageal and Head and Neck Cancers. *Alcohol*. 2005;35(3):175-185.
115. Edenberg HJ. The Genetics of Alcohol Metabolism: Role of Alcohol Dehydrogenase and Aldehyde Dehydrogenase Variants. *Alcohol Research & Health*. 2007;30(1):5.
116. Druesne-Pecollo N, Tehard B, Mallet Y, Gerber M, Norat T, Hercberg S, et al. Alcohol and Genetic Polymorphisms: Effect on Risk of Alcohol-Related Cancer. *The Lancet Oncology*. 2009;10(2):173-180.
117. Suwaki H, Ohara H. Alcohol-Induced Facial Flushing and Drinking Behavior in Japanese Men. *Journal of studies on alcohol*. 1985;46(3):196-198.

118. Shibuya A, Yasunami M, Yoshida A. Genotypes of Alcohol Dehydrogenase and Aldehyde Dehydrogenase Loci in Japanese Alcohol Flushers and Nonflushers. *Human Genetics*. 1989;82:14-16.
119. Zhang J, Zhang S, Song Y, Ma G, Meng Y, Ye Z, et al. Facial Flushing after Alcohol Consumption and the Risk of Cancer: A Meta-Analysis. *Medicine*. 2017;96(13):e6506.
120. Yu C, Guo Y, Bian Z, Yang L, Millwood IY, Walters RG, et al. Association of Low-Activity Aldh2 and Alcohol Consumption with Risk of Esophageal Cancer in Chinese Adults: A Population-Based Cohort Study. *International Journal of Cancer*. 2018;143(7):1652-1661.
121. Andrici J, Hu SX, Eslick GD. Facial Flushing Response to Alcohol and the Risk of Esophageal Squamous Cell Carcinoma: A Comprehensive Systematic Review and Meta-Analysis. *Cancer Epidemiology*. 2016;40:31-38.
122. Agrawal N, Jiao Y, Bettegowda C, Hutfless SM, Wang Y, David S, et al. Comparative Genomic Analysis of Esophageal Adenocarcinoma and Squamous Cell Carcinoma. *Cancer Discovery*. 2012;2(10):899-905.
123. Gao Y-B, Chen Z-L, Li J-G, Hu X-D, Shi X-J, Sun Z-M, et al. Genetic Landscape of Esophageal Squamous Cell Carcinoma. *Nature Genetics*. 2014;46(10):1097-1102.
124. Lin D-C, Hao J-J, Nagata Y, Xu L, Shang L, Meng X, et al. Genomic and Molecular Characterization of Esophageal Squamous Cell Carcinoma. *Nature Genetics*. 2014;46(5):467.
125. 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;509(7498):91-95.
126. Wang K, Johnson A, Ali SM, Klempner SJ, Bekaii-Saab T, Vacirca JL, et al. Comprehensive Genomic Profiling of Advanced Esophageal Squamous Cell Carcinomas and Esophageal Adenocarcinomas Reveals Similarities and Differences. *The Oncologist*. 2015;20(10):1132-1139.
127. Qin H-D, Liao X-Y, Chen Y-B, Huang S-Y, Xue W-Q, Li F-F, et al. Genomic Characterization of Esophageal Squamous Cell Carcinoma Reveals Critical Genes Underlying Tumorigenesis and Poor Prognosis. *The American Journal of Human Genetics*. 2016;98(4):709-727.
128. 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;150(5):1171-1182.
129. Erkizan HV, Johnson K, Ghimbovschi S, Karkera D, Trachiotis G, Adib H, et al. African-American Esophageal Squamous Cell Carcinoma Expression Profile Reveals Dysregulation of Stress Response and Detox Networks. *BMC Cancer*. 2017;17(1):1-13.
130. 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. *Nature Communications*. 2017;8(1):1-9.
131. Du P, Huang P, Huang X, Li X, Feng Z, Li F, et al. Comprehensive Genomic Analysis of Oesophageal Squamous Cell Carcinoma Reveals Clinical Relevance. *Scientific Reports*. 2017;7(1):15324.
132. Cancer Genome Atlas Research N, Analysis Working Group: Asan U, Agency BCC, Brigham, Women's H, Broad I, et al. Integrated Genomic Characterization of Oesophageal Carcinoma. *Nature*. 2017;541(7636):169-175.
133. Lin D-C, Dinh HQ, Xie J-J, Mayakonda A, Silva TC, Jiang Y-Y, et al. Identification of Distinct Mutational Patterns and New Driver Genes in Oesophageal Squamous Cell Carcinomas and Adenocarcinomas. *Gut*. 2018;67(10):1769-1779.

134. 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;20(1):1-10.
135. 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. *Frontiers in Oncology*. 2020;10:1457.
136. Cui Y, Chen H, Xi R, Cui H, Zhao Y, Xu E, et al. Whole-Genome Sequencing of 508 Patients Identifies Key Molecular Features Associated with Poor Prognosis in Esophageal Squamous Cell Carcinoma. *Cell Research*. 2020;30(10):1-12.
137. Erkizan HV, Sukhadia S, Natarajan TG, Marino G, Notario V, Lichy JH, et al. Exome Sequencing Identifies Novel Somatic Variants in African American Esophageal Squamous Cell Carcinoma. *Scientific Reports*. 2021;11(1):1-15.
138. Wang L, Jia Y-M, Zuo J, Wang Y-D, Fan Z-S, Feng L, et al. Gene Mutations of Esophageal Squamous Cell Carcinoma Based on Next-Generation Sequencing. *Chinese Medical Journal*. 2021;134(06):708-715.
139. Park S, Won D, Kim DJ, Park SY, Lee S-T. Genetic Alterations of Esophageal Squamous Cell Carcinoma in Korean Patients. *Research Square*. 2021.
140. Munari FF, Dos Santos W, Evangelista AF, Carvalho AC, Pastrez PA, Bugatti D, et al. Profile of Esophageal Squamous Cell Carcinoma Mutations in Brazilian Patients. *Scientific Reports*. 2021;11(1):1-13.
141. Zou B, Guo D, Kong P, Wang Y, Cheng X, Cui Y. Integrative Genomic Analyses of 1,145 Patient Samples Reveal New Biomarkers in Esophageal Squamous Cell Carcinoma. *Frontiers in Molecular Biosciences*. 2022;8:792779.
142. Chang J, Tan W, Ling Z, Xi R, Shao M, Chen M, et al. Genomic Analysis of Oesophageal Squamous-Cell Carcinoma Identifies Alcohol Drinking-Related Mutation Signature and Genomic Alterations. *Nature Communications*. 2017;8(1):1-11.
143. Zhang N, Shi J, Shi X, Chen W, Liu J. Mutational Characterization and Potential Prognostic Biomarkers of Chinese Patients with Esophageal Squamous Cell Carcinoma. *OncoTargets and Therapy*. 2020;13:12797–12809.
144. Alexandrov LB, Stratton MR. Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes. *Current Opinion in Genetics & Development*. 2014;24(100):52-60.
145. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*. 2013;3(1):246-259.
146. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of Mutational Processes in Human Cancer. *Nature*. 2013;500(7463):415.
147. Hu J, Cao J, Topatana W, Juengpanich S, Li S, Zhang B, et al. Targeting Mutant P53 for Cancer Therapy: Direct and Indirect Strategies. *Journal of Hematology & Oncology*. 2021;14:1-19.
148. Tang Y-Y, Wei P-J, Zhao J-p, Xia J, Cao R-F, Zheng C-H. Identification of Driver Genes Based on Gene Mutational Effects and Network Centrality. *BMC Bioinformatics*. 2021;22(3):1-16.
149. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-1041.e1021.
150. Pon JR, Marra MA. Driver and Passenger Mutations in Cancer. *Annual Review of Pathology: Mechanisms of Disease*. 2015;10(1):25-50.

151. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes. *Nature*. 2013;499(7457):214-218.
152. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. Oncodriveclust: Exploiting the Positional Clustering of Somatic Mutations to Identify Cancer Genes. *Bioinformatics*. 2013;29(18):2238-2244.
153. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. Oncodrivefml: A General Framework to Identify Coding and Non-Coding Regions with Cancer Driver Mutations. *Genome Biology*. 2016;17(1):1-13.
154. Li X, Wang M, Yang M, Dai H, Zhang B, Wang W, et al. A Mutational Signature Associated with Alcohol Consumption and Prognostically Significantly Mutated Driver Genes in Esophageal Squamous Cell Carcinoma. *Annals of Oncology*. 2018;29(4):938-944.
155. Dutta M, Nakagawa H, Kato H, Maejima K, Sasagawa S, Nakano K, et al. Whole Genome Sequencing Analysis Identifies Recurrent Structural Alterations in Esophageal Squamous Cell Carcinoma. *PeerJ*. 2020;8:e9294.
156. Li M, Zhang Z, Wang Q, Yi Y, Li B. Integrated Cohort of Esophageal Squamous Cell Cancer Reveals Genomic Features Underlying Clinical Characteristics. *Nature Communications*. 2022;13(1):1-15.
157. Dai W, Ko JMY, Choi SSA, Yu Z, Ning L, Zheng H, et al. Whole-Exome Sequencing Reveals Critical Genes Underlying Metastasis in Oesophageal Squamous Cell Carcinoma. *The Journal of Pathology*. 2017;242(4):500-510.
158. Zhang X, Wang Y, Meng L. Comparative Genomic Analysis of Esophageal Squamous Cell Carcinoma and Adenocarcinoma: New Opportunities Towards Molecularly Targeted Therapy. *Acta Pharmaceutica Sinica B*. 2022;12(3):1054-1067.
159. Kang N, Wang Y, Guo S, Ou Y, Wang G, Chen J, et al. Mutant Tp53 G245c and R273h Promote Cellular Malignancy in Esophageal Squamous Cell Carcinoma. *BMC Cell Biology*. 2018;19:1-11.
160. Li J, Yang L, Gaur S, Zhang K, Wu X, Yuan YC, et al. Mutants Tp 53 P. R273h and P. R273c but Not P. R273g Enhance Cancer Cell Malignancy. *Human Mutation*. 2014;35(5):575-584.
161. Kang HJ, Chun S-M, Kim K-R, Sohn I, Sung CO. Clinical Relevance of Gain-of-Function Mutations of P53 in High-Grade Serous Ovarian Carcinoma. *PLOS One*. 2013;8(8):e72609.
162. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-Related Remodelling of Oesophageal Epithelia by Mutated Cancer Drivers. *Nature*. 2019;565(7739):312-317.
163. Sharpless NE, DePinho RA. The Ink4a/Arf Locus and Its Two Gene Products. *Current Opinion in Genetics and Development*. 1999;9(1):22-30.
164. Ouelle DE, Zindy F, Ashmun RA, Sherr CJ. Alternative Reading Frames of the Ink4a Tumor Suppressor Gene Encode Two Unrelated Proteins Capable of Inducing Cell Cycle Arrest. *Cell*. 1995;83(6):993-1000.
165. Stone S, Jiang P, Dayananth P, Tavtigian SV, Katcher H, Parry D, et al. Complex Structure and Regulation of the P16 (Mts1) Locus. *Cancer Research*. 1995;55(14):2988-2994.
166. Serrano M, Hannon GJ, Beach D. A New Regulatory Motif in Cell-Cycle Control Causing Specific Inhibition of Cyclin D/Cdk4. *Nature*. 1993;366(6456):704-707.
167. Zhang Y, Xiong Y, Yarbrough WG. Arf Promotes Mdm2 Degradation and Stabilizes P53: Arf-Ink4a Locus Deletion Impairs Both the Rb and P53 Tumor Suppression Pathways. *Cell*. 1998;92(6):725-734.

168. Bates S, Phillips AC, Clark PA, Stott F, Peters G, Ludwig RL, et al. P14arf Links the Tumour Suppressors Rb and P53. *Nature*. 1998;395(6698):124-125.
169. Fontana R, Ranieri M, La Mantia G, Vivo M. Dual Role of the Alternative Reading Frame Arf Protein in Cancer. *Biomolecules*. 2019;9(3):87.
170. Brown VL, Harwood CA, Crook T, Cronin JG, Kelsell DP, Proby CM. P16ink4a and P14arf Tumor Suppressor Genes Are Commonly Inactivated in Cutaneous Squamous Cell Carcinoma. *Journal of Investigative Dermatology*. 2004;122(5):1284-1292.
171. Hanahan D, Weinberg RA. The Hallmarks of Cancer. *Cell*. 2000;100(1):57-70.
172. Liu X, Zhang M, Ying S, Zhang C, Lin R, Zheng J, et al. Genetic Alterations in Esophageal Tissues from Squamous Dysplasia to Carcinoma. *Gastroenterology*. 2017;153(1):166-177.
173. Hu N, Wang C, Su H, Li WJ, Emmert-Buck MR, Li G, et al. High Frequency of Cdkn2a Alterations in Esophageal Squamous Cell Carcinoma from a High-Risk Chinese Population. *Genes, Chromosomes and Cancer*. 2004;39(3):205-216.
174. Smeds J, Berggren P, Ma X, Xu Z, Hemminki K, Kumar R. Genetic Status of Cell Cycle Regulators in Squamous Cell Carcinoma of the Oesophagus: The Cdkn2a (P16 Ink4a and P14 Arf) and P53 Genes Are Major Targets for Inactivation. *Carcinogenesis*. 2002;23(4):645-655.
175. Xing EP, Nie Y, Song Y, Yang G-Y, Cai YC, Wang L-D, et al. Mechanisms of Inactivation of P14 Arf, P15 Ink4b, and P16 Ink4a Genes in Human Esophageal Squamous Cell Carcinoma. *Clinical Cancer Research*. 1999;5(10):2704-2713.
176. Onozato Y, Sasaki Y, Abe Y, Sato H, Yagi M, Mizumoto N, et al. Novel Genomic Alteration in Superficial Esophageal Squamous Cell Neoplasms in Non-Smoker Non-Drinker Females. *Scientific Reports*. 2021;11(1):1-11.
177. Borg Ak, Sandberg T, Nilsson K, Johannsson O, Klinker M, Måsbäck A, et al. High Frequency of Multiple Melanomas and Breast and Pancreas Carcinomas in Cdkn2a Mutation-Positive Melanoma Families. *Journal of the National Cancer Institute*. 2000;92(15):1260-1266.
178. Ozenne P, Eymin B, Brambilla E, Gazzeri S. The Arf Tumor Suppressor: Structure, Functions and Status in Cancer. *International Journal of Cancer*. 2010;127(10):2239-2247.
179. Ruas M. The P16ink4a/Cdkn2a Tumor Suppressor and Its Relatives. *Biochimica et Biophysica Acta*. 1998;1378(2):F115-F177.
180. Li Z, Ding S, Zhong Q, Li G, Zhang Y, Huang XC. Significance of Mmp11 and P14arf Expressions in Clinical Outcomes of Patients with Laryngeal Cancer. *International Journal of Clinical and Experimental Medicine* 2015;8(9):15581.
181. Ito T, Nishida N, Fukuda Y, Nishimura T, Komeda T, Nakao K. Alteration of the P14 Arf Gene and P53 Status in Human Hepatocellular Carcinomas. *Journal of Gastroenterology*. 2004;39:355-361.
182. Cabral VD, Cerski MR, Sa Brito IT, Kliemann LM. P14 Expression Differences in Ovarian Benign, Borderline and Malignant Epithelial Tumors. *Journal of Ovarian Research*. 2016;9(1):1-7.
183. Pare R, Shin JS, Lee CS. Increased Expression of Senescence Markers P14arf and P16ink4a in Breast Cancer Is Associated with an Increased Risk of Disease Recurrence and Poor Survival Outcome. *Histopathology*. 2016;69(3):479-491.
184. Wang F, Li H, Long J, Ye S. Clinicopathological Significance of P14arf Expression in Lung Cancer: A Meta-Analysis. *OncoTargets and Therapy*. 2017:2491-2499.
185. Ming Z, Lim SY, Rizos H. Genetic Alterations in the Ink4a/Arf Locus: Effects on Melanoma Development and Progression. *Biomolecules*. 2020;10(10):1447.

186. Bottillo I, Valiante M, Menale L, Paiardini A, Papi L, Janson G, et al. A Novel Cdkn2a in-Frame Deletion Associated with Pancreatic Cancer-Melanoma Syndrome. *Dermatology Online Journal*. 2020;26(8):13030/qt13035t13022m13035gk.
187. Reed AL, Califano J, Cairns P, Westra WH, Jones RM, Koch W, et al. High Frequency of P16 (Cdkn2/Mts-1/Ink4a) Inactivation in Head and Neck Squamous Cell Carcinoma. *Cancer Research*. 1996;56(16):3630-3633.
188. Sano T, Masuda N, Oyama T, Nakajima T. Overexpression of P16 and P14arf Is Associated with Human Papillomavirus Infection in Cervical Squamous Cell Carcinoma and Dysplasia. *Pathology International*. 2002;52(5-6):375-383.
189. Zhao R, Choi BY, Lee M-H, Bode AM, Dong Z. Implications of Genetic and Epigenetic Alterations of Cdkn2a (P16ink4a) in Cancer. *EBioMedicine*. 2016;8:30-39.
190. Tam KW, Zhang W, Soh J, Stastny V, Chen M, Sun H, et al. Cdkn2a/P16 Inactivation Mechanisms and Their Relationship to Smoke Exposure and Molecular Features in Non–Small-Cell Lung Cancer. *Journal of Thoracic Oncology*. 2013;8(11):1378-1388.
191. Marchetti A, Buttitta F, Pellegrini S, Bertacca G, Chella A, Carnicelli V, et al. Alterations of P16 (Mts1) in Node-Positive Non-Small Cell Lung Carcinomas. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*. 1997;181(2):178-182.
192. de Almeida Simao T, De Bonis Almeida Simões GL, Ribeiro FS, de Paula Cidade DA, Andreollo NA, Lopes LR, et al. Lower Expression of P14arf and P16ink4a Correlates with Higher Dnmt3b Expression in Human Oesophageal Squamous Cell Carcinomas. *Human & Experimental Toxicology*. 2006;25(9):515-522.
193. Güner D, Sturm I, Hemmati P, Hermann S, Hauptmann S, Wurm R, et al. Multigene Analysis of Rb Pathway and Apoptosis Control in Esophageal Squamous Cell Carcinoma Identifies Patients with Good Prognosis. *International Journal of Cancer*. 2003;103(4):445-454.
194. Tokugawa T, Sugihara H, Tani T, Hattori T. Modes of Silencing of P16 in Development of Esophageal Squamous Cell Carcinoma. *Cancer Research*. 2002;62(17):4938-4944.
195. Siddiqui S, Libertini SJ, Lucas CA, Lombard AP, Baek HB, Nakagawa RM, et al. The P14arf Tumor Suppressor Restrains Androgen Receptor Activity and Prevents Apoptosis in Prostate Cancer Cells. *Cancer Letters*. 2020;483:12-21.
196. Mao L, Merlo A, Bedi G, Shapiro GI, Edwards CD, Rollins BJ, et al. A Novel P16ink4a Transcript. *Cancer Research*. 1995;55(14):2995-2997.
197. Li J, Li L, Zhang S. Different Expression of P16ink4a and P14arf in Cervical and Lung Cancers. *European Review for Medical and Pharmacological Sciences*. 2013;17(22):3007-3011.
198. Liu W, Zhuang C, Huang T, Yang S, Zhang M, Lin B, et al. Loss of Cdkn2a at Chromosome 9 Has a Poor Clinical Prognosis and Promotes Lung Cancer Progression. *Molecular Genetics & Genomic Medicine*. 2020;8(12):e1521.
199. Alhejaily A, Day AG, Feilotter HE, Baetz T, LeBrun DP. Inactivation of the Cdkn2a Tumor-Suppressor Gene by Deletion or Methylation Is Common at Diagnosis in Follicular Lymphoma and Associated with Poor Clinical Outcome. *Clinical Cancer Research*. 2014;20(6):1676-1686.
200. Cohen I, Birnbaum RY, Leibson K, Taube R, Sivan S, Birk OS. Znf750 Is Expressed in Differentiated Keratinocytes and Regulates Epidermal Late Differentiation Genes. *PLOS One*. 2012;7(8):e42628.

201. Nambara S, Masuda T, Tobo T, Kidogami S, Komatsu H, Sugimachi K, et al. Clinical Significance of Znf750 Gene Expression, a Novel Tumor Suppressor Gene, in Esophageal Squamous Cell Carcinoma. *Oncology Letters*. 2017;14(2):1795-1801.
202. Kong P, Xu E, Bi Y, Xu X, Liu X, Song B, et al. Novel Escc-Related Gene Znf750 as Potential Prognostic Biomarker and Inhibits Epithelial-Mesenchymal Transition through Directly Depressing Snai1 Promoter in Escc. *Theranostics*. 2020;10(4):1798.
203. Otsuka R, Akutsu Y, Sakata H, Hanari N, Murakami K, Kano M, et al. Znf750 Expression Is a Potential Prognostic Biomarker in Esophageal Squamous Cell Carcinoma. *Oncology*. 2018;94(3):142-148.
204. Otsuka R, Akutsu Y, Sakata H, Hanari N, Murakami K, Kano M, et al. Znf750 Expression as a Novel Candidate Biomarker of Chemoradiosensitivity in Esophageal Squamous Cell Carcinoma. *Oncology*. 2017;93(3):197-203.
205. Yuan X, Wu H, Xu H, Xiong H, Chu Q, Yu S, et al. Notch Signaling: An Emerging Therapeutic Target for Cancer Treatment. *Cancer Letters*. 2015;369(1):20-27.
206. Ranganathan P, Weaver KL, Capobianco AJ. Notch Signalling in Solid Tumours: A Little Bit of Everything but Not All the Time. *Nature Reviews Cancer*. 2011;11(5):338-351.
207. Ntziachristos P, Lim JS, Sage J, Aifantis I. From Fly Wings to Targeted Cancer Therapies: A Centennial for Notch Signaling. *Cancer Cell*. 2014;25(3):318-334.
208. Li Y, Li Y, Chen X. Notch and Esophageal Squamous Cell Carcinoma. *Advances in Experimental Medicine and Biology*. 2021;1287:59-68.
209. Sawangarun W, Mandasari M, Aida J, Morita K-i, Kayamori K, Ikeda T, et al. Loss of Notch1 Predisposes Oro-Esophageal Epithelium to Tumorigenesis. *Experimental Cell Research*. 2018;372(2):129-140.
210. Song B, Cui H, Li Y, Cheng C, Yang B, Wang F, et al. Mutually Exclusive Mutations in Notch1 and Pik3ca Associated with Clinical Prognosis and Chemotherapy Responses of Esophageal Squamous Cell Carcinoma in China. *Oncotarget*. 2016;7(3):3599-3613.
211. Bannister AJ, Kouzarides T. Regulation of Chromatin by Histone Modifications. *Cell Research*. 2011;21(3):381-395.
212. Froimchuk E, Jang Y, Ge K. Histone H3 Lysine 4 Methyltransferase Kmt2d. *Gene*. 2017;627:337-342.
213. Sasaki Y, Tamura M, Koyama R, Nakagaki T, Adachi Y, Tokino T. Genomic Characterization of Esophageal Squamous Cell Carcinoma: Insights from Next-Generation Sequencing. *World Journal of Gastroenterology*. 2016;22(7):2284-2293.
214. Suzuki T, Yamamoto M. Molecular Basis of the Keap1–Nrf2 System. *Free Radical Biology and Medicine*. 2015;88(Pt B):93-100.
215. Taguchi K, Motohashi H, Yamamoto M. Molecular Mechanisms of the Keap1–Nrf2 Pathway in Stress Response and Cancer Evolution. *Genes to Cells*. 2011;16(2):123-140.
216. Shibata T, Ohta T, Tong KI, Kokubu A, Odogawa R, Tsuta K, et al. Cancer Related Mutations in Nrf2 Impair Its Recognition by Keap1-Cul3 E3 Ligase and Promote Malignancy. *Proceedings of the National Academy of Sciences*. 2008;105(36):13568-13573.
217. Tong KI, Katoh Y, Kusunoki H, Itoh K, Tanaka T, Yamamoto M. Keap1 Recruits Neh2 through Binding to Etge and Dlg Motifs: Characterization of the Two-Site Molecular Recognition Model. *Molecular and Cellular Biology*. 2006;26(8):2887-2900.
218. Goldstein LD, Lee J, Gnad F, Klijn C, Schaub A, Reeder J, et al. Recurrent Loss of Nfe2l2 Exon 2 Is a Mechanism for Nrf2 Pathway Activation in Human Cancers. *Cell Reports*. 2016;16(10):2605-2617.

219. Kerins MJ, Ooi A. A Catalogue of Somatic Nrf2 Gain-of-Function Mutations in Cancer. *Scientific Reports*. 2018;8(1):12846.
220. Kim YR, Oh JE, Kim MS, Kang MR, Park SW, Han JY, et al. Oncogenic Nrf2 Mutations in Squamous Cell Carcinomas of Oesophagus and Skin. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*. 2010;220(4):446-451.
221. Shibata T, Kokubu A, Saito S, Narisawa-Saito M, Sasaki H, Aoyagi K, et al. Nrf2 Mutation Confers Malignant Potential and Resistance to Chemoradiation Therapy in Advanced Esophageal Squamous Cancer. *Neoplasia*. 2011;13(9):864-IN826.
222. Ma S, Paiboonrungruan C, Yan T, Williams KP, Major MB, Chen XL. Targeted Therapy of Esophageal Squamous Cell Carcinoma: The Nrf2 Signaling Pathway as Target. *Annals of the New York Academy of Sciences*. 2018;1434(1):164-172.
223. Beura PK, Aziz R, Sen P, Das S, Namsa ND, Feil EJ, et al. Synonymous and Non-Synonymous Transitions/Transversions Vividly Disclose Purifying Selection in Escherichia Coli Coding Sequences. *BioRxiv*. 2022:2022-2011.
224. Luo G-H, Li X-H, Han Z-J, Zhang Z-C, Yang Q, Guo H-F, et al. Transition and Transversion Mutations Are Biased Towards Gc in Transposons of Chilo Suppressalis (Lepidoptera: Pyralidae). *Genes*. 2016;7(10):72.
225. Gojobori T, Li W-H, Graur D. Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes. *Journal of Molecular Evolution*. 1982;18(5):360-369.
226. Petrov DA, Hartl DL. Patterns of Nucleotide Substitution in Drosophila and Mammalian Genomes. *Proceedings of the National Academy of Sciences*. 1999;96(4):1475-1479.
227. Barnes DE, Lindahl T. Repair and Genetic Consequences of Endogenous DNA Base Damage in Mammalian Cells. *Annual Review of Genetics*. 2004;38(1):445-476.
228. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science*. 2011;333(6046):1157-1160.
229. Cancer Genome Atlas Network. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature*. 2012;487(7407):330-337.
230. 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;578(7793):94-101.
231. Pfeifer GP. Environmental Exposures and Mutational Patterns of Cancer Genomes. *Genome Medicine*. 2010;2(8):54.
232. Wiebauer K, Neddermann P, Hughes M, Jiricny J. The Repair of 5-Methylcytosine Deamination Damage: Experientia Supplementum; 1993. 510-522 p.
233. Pena-Diaz J, Bregenhorn S, Ghodgaonkar M, Follonier C, Artola-Boran M, Castor D, et al. Noncanonical Mismatch Repair as a Source of Genomic Instability in Human Cells. *Molecular Cell*. 2012;47(5):669-680.
234. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms Underlying Mutational Signatures in Human Cancers. *Nature Reviews Genetics*. 2014;15(9):585-598.
235. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, et al. A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome. *Nature*. 2010;463(7278):191-196.
236. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco Smoke Carcinogens, DNA Damage and P53 Mutations in Smoking-Associated Cancers. *Oncogene*. 2002;21(48):7435-7451.

237. Wilbourn J, Haroun L, Heseltine E, Kaldor J, Partensky C, Vainio H. Response of Experimental Animals to Human Carcinogens: An Analysis Based Upon the Iarc Monographs Programme. *Carcinogenesis*. 1986;7(11):1853-1863.
238. Guo J, Huang J, Zhou Y, Zhou Y, Yu L, Li H, et al. Germline and Somatic Variations Influence the Somatic Mutational Signatures of Esophageal Squamous Cell Carcinomas in a Chinese Population. *BMC genomics*. 2018;19(1):538.
239. Liang J, Wang Y, Cai L, Liu J, Yan J, Chen X, et al. Comparative Genomic Analysis Reveals Genetic Variations in Multiple Primary Esophageal Squamous Cell Carcinoma of Chinese Population. *Frontiers in Oncology*. 2022;12:868301.
240. Zhang R, Li C, Wan Z, Qin J, Li Y, Wang Z, et al. Comparative Genomic Analysis of Esophageal Squamous Cell Carcinoma among Different Geographic Regions. *Frontiers in Oncology*. 2023;12:999424.
241. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-Like Mutational Processes in Human Somatic Cells. *Nature Genetics*. 2015;47(12):1402-1407.
242. Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, et al. Apobec3b Is an Enzymatic Source of Mutation in Breast Cancer. *Nature*. 2013;494(7437):366-370.
243. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An Apobec Cytidine Deaminase Mutagenesis Pattern Is Widespread in Human Cancers. *Nature Genetics*. 2013;45(9):970-976.
244. Harris RS, Petersen-Mahrt SK, Neuberger MS. Rna Editing Enzyme Apobec1 and Some of Its Homologs Can Act as DNA Mutators. *Molecular Cell*. 2002;10(5):1247-1253.
245. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. Cosmic: Exploring the World's Knowledge of Somatic Mutations in Human Cancer. *Nucleic Acids Research*. 2015;43(D1):D805-D811.
246. Dotto GP, Rustgi AK. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. *Cancer Cell*. 2016;29(5):622-637.
247. Malumbres M, Barbacid M. Cell Cycle, Cdks and Cancer: A Changing Paradigm. *Nature Reviews Cancer*. 2009;9(3):153-166.
248. Ogawa R, Ishiguro H, Kimura M, Funahashi H, Wakasugi T, Ando T, et al. Notch1 Expression Predicts Patient Prognosis in Esophageal Squamous Cell Cancer. *European Surgical Research*. 2013;51(3-4):101-107.
249. Kagawa S, Natsuizaka M, Whelan KA, Facompre N, Naganuma S, Ohashi S, et al. Cellular Senescence Checkpoint Function Determines Differential Notch1-Dependent Oncogenic and Tumor-Suppressor Activities. *Oncogene*. 2015;34(18):2347-2359.
250. Cheng C, Cui H, Zhang L, Jia Z, Song B, Wang F, et al. Genomic Analyses Reveal Fam84b and the Notch Pathway Are Associated with the Progression of Esophageal Squamous Cell Carcinoma. *Gigascience*. 2016;5(1):s13742-13015.
251. Snigdha K, Gangwani KS, Lapalikar GV, Singh A, Kango-Singh M. Hippo Signaling in Cancer: Lessons from Drosophila Models. *Frontiers in Cell and Developmental Biology*. 2019;7:85.
252. Kim W, Khan SK, Gvozdenovic-Jeremic J, Kim Y, Dahlman J, Kim H, et al. Hippo Signaling Interactions with Wnt/B-Catenin and Notch Signaling Repress Liver Tumorigenesis. *The Journal of Clinical Investigation*. 2017;127(1):137-152.
253. Yu FX, Guan KL. The Hippo Pathway, Regulators and Regulations. *Genes & Development*. 2013;27(4):355-371.

254. Schutte U, Bisht S, Heukamp LC, Kebschull M, Florin A, Haarmann J. Hippo Signaling Mediates Proliferation, Invasiveness, and Metastatic Potential of Clear Cell Renal Cell Carcinoma. *Translational Oncology*. 2014;7(2):309-321.
255. Misra JR, Irvine KD. The Hippo Signaling Network and Its Biological Functions. *Annual Review of Genetics*. 2018;52(1):65-87.
256. Proweller A, Tu L, Lepore JJ, Cheng L, Lu MM, Seykora J, et al. Impaired Notch Signaling Promotes De Novo Squamous Cell Carcinoma Formation. *Cancer Research*. 2006;66(15):7438-7444.
257. Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, et al. Integrative Genomic Characterization of Oral Squamous Cell Carcinoma Identifies Frequent Somatic Drivers. *Cancer Discovery*. 2013;3(7):770-781.
258. Wang NJ, Sanborn Z, Arnett KL, Bayston LJ, Liao W, Proby CM, et al. Loss-of-Function Mutations in Notch Receptors in Cutaneous and Lung Squamous Cell Carcinoma. *Proceedings of the National Academy of Sciences*. 2011;108(43):17761-17766.
259. Tsang C, Lo K, Nicholls JM, Huang S, Tsao S. Pathogenesis of Nasopharyngeal Carcinoma: Histogenesis, Epstein–Barr Virus Infection, and Tumor Microenvironment. *Nasopharyngeal Carcinoma*: Elsevier; 2019. p. 45-64.
260. Yuan T, Cantley L. Pi3k Pathway Alterations in Cancer: Variations on a Theme. *Oncogene*. 2008;27(41):5497-5510.
261. Wu N, Du Z, Zhu Y, Song Y, Pang L, Chen Z. The Expression and Prognostic Impact of the Pi3k/Akt/Mtor Signaling Pathway in Advanced Esophageal Squamous Cell Carcinoma. *Technology in Cancer Research & Treatment*. 2018;17:1533033818758772.
262. Osaki M, Oshimura Ma, Ito H. Pi3k-Akt Pathway: Its Functions and Alterations in Human Cancer. *Apoptosis*. 2004;9(6):667-676.
263. Liu P, Cheng H, Roberts TM, Zhao JJ. Targeting the Phosphoinositide 3-Kinase Pathway in Cancer. *Nature Reviews Drug Discovery*. 2009;8(8):627-644.
264. Lin J-w, Li X, Qiu M-l, Luo R-g, Lin J-b, Liu B. Pi3k Overexpression and Pik3ca Mutations Are Associated with Age, Tumor Staging, and Other Clinical Characteristics in Chinese Patients with Esophageal Squamous Cell Carcinoma. *Genetic Testing and Molecular Biomarkers*. 2017;21(4):236-241.
265. Cai Y, Dodhia S, Su GH. Dysregulations in the Pi3k Pathway and Targeted Therapies for Head and Neck Squamous Cell Carcinoma. *Oncotarget*. 2017;8(13):22203.
266. Brugge J, Hung M-C, Mills GB. A New Mutational Aktivation in the Pi3k Pathway. *Cancer Cell*. 2007;12(2):104-107.
267. Vivanco I, Sawyers CL. The Phosphatidylinositol 3-Kinase–Akt Pathway in Human Cancer. *Nature Reviews Cancer*. 2002;2(7):489-501.
268. Fruman DA, Chiu H, Hopkins BD, Bagrodia S, Cantley LC, Abraham RT. The Pi3k Pathway in Human Disease. *Cell*. 2017;170(4):605-635.
269. Ligresti G, Militello L, Steelman LS, Cavallaro A, Basile F, Nicoletti F, et al. Pik3ca Mutations in Human Solid Tumors: Role in Sensitivity to Various Therapeutic Approaches. *Cell Cycle*. 2009;8(9):1352-1358.
270. Zhao JJ, Cheng H, Jia S, Wang L, Gjoerup OV, Mikami A, et al. The P110α Isoform of Pi3k Is Essential for Proper Growth Factor Signaling and Oncogenic Transformation. *Proceedings of the National Academy of Sciences*. 2006;103(44):16296-16300.
271. Menegon S, Columbano A, Giordano S. The Dual Roles of Nrf2 in Cancer. *Trends in Molecular Medicine*. 2016;22(7):578-593.

272. Lau A, Villeneuve NF, Sun Z, Wong PK, Zhang DD. Dual Roles of Nrf2 in Cancer. *Pharmacological Research*. 2008;58(5-6):262-270.
273. Hirose W, Oshikiri H, Taguchi K, Yamamoto M. The Keap1-Nrf2 System and Esophageal Cancer. *Cancers*. 2022;14(19):4702.
274. He F, Ru X, Wen T. Nrf2, a Transcription Factor for Stress Response and Beyond. *International Journal of Molecular Sciences*. 2020;21(13):4777.
275. Finotello F, Di Camillo B. Measuring Differential Gene Expression with Rna-Seq: Challenges and Strategies for Data Analysis. *Briefings in Functional Genomics*. 2015;14(2):130-142.
276. Huggett J, Dheda K, Bustin S, Zumla A. Real-Time Rt-Pcr Normalisation; Strategies and Considerations. *Genes & Immunity*. 2005;6(4):279-284.
277. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and Quantifying Mammalian Transcriptomes by Rna-Seq. *Nature Methods*. 2008;5(7):621-628.
278. Schena M, Shalon D, Davis RW, Brown PO. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*. 1995;270(5235):467-470.
279. Beer DG, Kardia SL, Huang C-C, Giordano TJ, Levin AM, Misek DE, et al. Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *Nature Medicine*. 2002;8(8):816-824.
280. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene Expression Profiling Predicts Clinical Outcome of Prostate Cancer. *The Journal of Clinical Investigation*. 2004;113(6):913-923.
281. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature*. 2002;415(6870):436-442.
282. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*. 2002;415(6871):530-536.
283. Patowary P, Bhattacharyya DK, Barah P. Identifying Critical Genes in Esophageal Squamous Cell Carcinoma Using an Ensemble Approach. *Informatics in Medicine Unlocked*. 2020;18:100277.
284. Lin L, Lin D-C. Biological Significance of Tumor Heterogeneity in Esophageal Squamous Cell Carcinoma. *Cancers*. 2019;11(8):1156.
285. Gerlinger M, Swanton C. How Darwinian Models Inform Therapeutic Failure Initiated by Clonal Heterogeneity in Cancer Medicine. *British Journal of Cancer*. 2010;103(8):1139-1143.
286. Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C. Intratumor Heterogeneity: Seeing the Wood for the Trees. *Science Translational Medicine*. 2012;4(127):127ps110-127ps110.
287. Moody S, Senkin S, Islam S, Wang J, Nasrollahzadeh D, Cortez Cardoso Penha R, et al. Mutational Signatures in Esophageal Squamous Cell Carcinoma from Eight Countries with Varying Incidence. *Nature Genetics*. 2021;53(11):1553-1563.
288. Jones D, Raine KM, Davies H, Tarpey PS, Butler AP, Teague JW, et al. Cgpcavemanwrapper: Simple Execution of Caveman in Order to Detect Somatic Single Nucleotide Variants in Ngs Data. *Current Protocols in Bioinformatics*. 2016;56(1):15-10.
289. Raine KM, Hinton J, Butler AP, Teague JW, Davies H, Tarpey P, et al. Cgppindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Current Protocols in Bioinformatics*. 2015;52(1):15-17.

290. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. Mutationalpatterns: Comprehensive Genome-Wide Analysis of Mutational Processes. *Genome Medicine*. 2018;10(1):1-11.
291. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. Deconstructsigs: Delineating Mutational Processes in Single Tumors Distinguishes DNA Repair Deficiencies and Patterns of Carcinoma Evolution. *Genome Biology*. 2016;17(1):1-11.
292. Teh Y, Jordan M, Beal M, Blei D. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. *Advances in Neural Information Processing Systems*. 2004;17:1385–1392.
293. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types. *Nature*. 2014;505(7484):495-501.
294. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Research* 2018;46(D1):D649-D655.
295. Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, et al. Genvisr: Genomic Visualizations in R. *Bioinformatics*. 2016;32(19):3012-3014.
296. Li L, Wang K, Tan B, Lyu Y, He X, Wang W, et al. Genomic Profiling of Chinese Esophageal Squamous Cell Carcinoma Patients and Difference of Genomic Mutation between Chinese and American Cohorts. *Journal of Clinical Oncology*. 2021;39(15_suppl):e16108-e16108.
297. Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Oesophageal Carcinoma. *Nature*. 2017;541(7636):169.
298. Gowers K, Yoshida K, Lee-Six H, Chandrasekharan D, Maughan E, Millar F, et al. Tobacco Exposure and Somatic Mutations in Normal Human Bronchial Epithelium. *Nature*. 2020;578(7794):A4090-A4090.
299. Frankell AM, Jammula S, Li X, Contino G, Killcoyne S, Abbas S, et al. The Landscape of Selection in 551 Esophageal Adenocarcinomas Defines Genomic Biomarkers for the Clinic. *Nature Genetics*. 2019;51(3):506-516.
300. Duro D, Bernard O, Della Valle V, Berger R, Larsen C-J. A New Type of P16ink4/Mts1 Gene Transcript Expressed in B-Cell Malignancies. *Oncogene*. 1995;11(1):21-29.
301. Serrano M, Lee H-W, Chin L, Cordon-Cardo C, Beach D, DePinho RA. Role of the Ink4a Locus in Tumor Suppression and Cell Mortality. *Cell*. 1996;85(1):27-37.
302. Pfeifer G. Mutagenesis at Methylated Cpg Sequences [E-Book]: Current Topics in Microbiology and Immunology; 2006.
303. Di Noia JM, Neuberger MS. Molecular Mechanisms of Antibody Somatic Hypermutation. *Annual Review of Biochemistry*. 2007;76(1):1-22.
304. Wu Y, Chua EHZ, Ng AWT, Boot A, Rozen SG. Accuracy of Mutational Signature Software on Correlated Signatures. *Scientific Reports*. 2022;12(1):390.
305. Li K, Luo H, Huang L, Luo H, Zhu X. Microsatellite Instability: A Review of What the Oncologist Should Know. *Cancer Cell International*. 2020;20:1-13.
306. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AW, Wu Y, et al. The Repertoire of Mutational Signatures in Human Cancer. *BioRxiv*. 2018:322859.
307. Zhou B, Lin W, Long Y, Yang Y, Zhang H, Wu K, et al. Notch Signaling Pathway: Architecture, Disease, and Therapeutics. *Signal transduction and targeted therapy*. 2022;7(1):95.
308. Palumbo Jr A, Meireles Da Costa N, Pontes B, Leite de Oliveira F, Lohan Codeço M, Ribeiro Pinto LF, et al. Esophageal Cancer Development: Crucial Clues Arising from the Extracellular Matrix. *Cells*. 2020;9(2):455.

309. Kai F, Drain AP, Weaver VM. The Extracellular Matrix Modulates the Metastatic Journey. *Developmental Cell*. 2019;49(3):332-346.
310. Rosińczuk J, Taradaj J, Dymarek R, Sopel M. Mechanoregulation of Wound Healing and Skin Homeostasis. *BioMed Research International*. 2016;2016(1):3943481.
311. Rivlin N, Brosh R, Oren M, Rotter V. Mutations in the P53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis. *Genes & Cancer*. 2011;2(4):466-474.
312. Austenaa L, Barozzi I, Chronowska A, Termanini A, Ostuni R, Prosperini E, et al. The Histone Methyltransferase Wbp7 Controls Macrophage Function through Gpi Glycolipid Anchor Synthesis. *Immunity*. 2012;36(4):572-585.
313. Lin LH, Allison DH, Feng Y, Jour G, Park K, Zhou F, et al. Comparison of Solid Tissue Sequencing and Liquid Biopsy Accuracy in Identification of Clinically Relevant Gene Mutations and Rearrangements in Lung Adenocarcinomas. *Modern Pathology*. 2021;34(12):2168-2174.
314. Quelle DE, Cheng M, Ashmun RA, Sherr CJ. Cancer-Associated Mutations at the Ink4a Locus Cancel Cell Cycle Arrest by P16ink4a but Not by the Alternative Reading Frame Protein P19arf. *Proceedings of the National Academy of Sciences*. 1997;94(2):669-673.
315. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between Driver and Passenger Mutations in Individual Cancer Genomes by Network Enrichment Analysis. *BMC Bioinformatics*. 2014;15(1):1-21.
316. Serrano M, Lin AW, McCurrach ME, Beach D, Lowe SW. Oncogenic Ras Provokes Premature Cell Senescence Associated with Accumulation of P53 and P16ink4a. *Cell*. 1997;88(5):593-602.
317. Kamijo T, Zindy F, Roussel MF, Quelle DE, Downing JR, Ashmun RA, et al. Tumor Suppression at the Mouse Ink4a Locus Mediated by the Alternative Reading Frame Product P19 Arf. *Cell*. 1997;91(5):649-659.
318. Shamloo B, Usluer S. P21 in Cancer Research. *Cancers*. 2019;11(8):1178.
319. Dotto GP. Crosstalk of Notch with P53 and P63 in Cancer Growth Control. *Nature Reviews Cancer*. 2009;9(8):587-595.
320. Narrandes S, Xu W. Gene Expression Detection Assay for Cancer Clinical Use. *Journal of Cancer*. 2018;9(13):2249-2265.
321. Hu YC, Lam KY, Law S, Wong J, Srivastava G. Identification of Differentially Expressed Genes in Esophageal Squamous Cell Carcinoma (Escc) by Cdna Expression Array: Overexpression of Fra-1, Neogenin, Id-1, and Cdc25b Genes in Escc. *Clinical Cancer Research*. 2001;7(8):2213-2221.
322. Berns A. Gene Expression in Diagnosis. *Nature*. 2000;403(6769):491-492.
323. Su H, Hu N, Yang HH, Wang C, Takikita M, Wang Q-H, et al. Global Gene Expression Profiling and Validation in Esophageal Squamous Cell Carcinoma and Its Association with Clinical Phenotypes. *Clinical Cancer Research*. 2011;17(9):2955-2966.
324. Sawada G, Niida A, Hirata H, Komatsu H, Uchi R, Shimamura T, et al. An Integrative Analysis to Identify Driver Genes in Esophageal Squamous Cell Carcinoma. *PLOS One*. 2015;10(10):e0139808.
325. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer Biomarker Discovery and Validation. *Translational Cancer Research*. 2015;4(3):256.
326. Tomasetti C, Vogelstein B. Variation in Cancer Risk among Tissues Can Be Explained by the Number of Stem Cell Divisions. *Science*. 2015;347(6217):78-81.
327. Yuan H, Xu Y, Luo Y, Wang N-X, Xiao J-H. Role of Nrf2 in Cell Senescence Regulation. *Molecular and Cellular Biochemistry*. 2021;476(1):247-259.

328. Chen W, Sun Z, Wang X-J, Jiang T, Huang Z, Fang D, et al. Direct Interaction between Nrf2 and P21cip1/Waf1 Upregulates the Nrf2-Mediated Antioxidant Response. *Molecular Cell*. 2009;34(6):663-673.
329. Oshimori N, Oristian D, Fuchs E. Tgf-B Promotes Heterogeneity and Drug Resistance in Squamous Cell Carcinoma. *Cell*. 2015;160(5):963-976.
330. Li W-Q, Hu N, Hyland PL, Gao Y, Wang Z-M, Yu K, et al. Genetic Variants in DNA Repair Pathway Genes and Risk of Esophageal Squamous Cell Carcinoma and Gastric Adenocarcinoma in a Chinese Population. *Carcinogenesis*. 2013;34(7):1536-1542.
331. Wang G, Guo S, Zhang W, Li Z, Xu J, Li D, et al. A Comprehensive Analysis of Alterations in DNA Damage Repair Pathways Reveals a Potential Way to Enhance the Radio-Sensitivity of Esophageal Squamous Cell Cancer. *Frontiers in Oncology*. 2020;10:575711.
332. Sherr CJ. The Ink4a/Arf Network in Tumour Suppression. *Nature Reviews Molecular Cell Biology*. 2001;2(10):731-737.
333. Tang KS, Guralnick BJ, Wang WK, Fersht AR, Itzhaki LS. Stability and Folding of the Tumour Suppressor Protein P16. *Journal of Molecular Biology*. 1999;285(4):1869-1886.
334. Nakashima R, Fujita M, Enomoto T, Haba T, Yoshino K, Wada H, et al. Alteration of P16 and P15 Genes in Human Uterine Tumours. *British Journal of Cancer*. 1999;80(3):458-467.
335. Livak K, Schmittgen T. Analysis of Relative Gene Expression Data Using Real-Time Quantitative Pcr and the 2 [− Delta Delta C [T]] Normalized to Glyceraldehyde-3-Phosphate Dehydrogenase Levels. Qrt-Pcr Was Method. *Methods*. 2001;25(4):402-408.
336. Burri N, Shaw P, Bouzourene H, Sordat I, Sordat B, Gillet M, et al. Methylation Silencing and Mutations of the P14arf and P16ink4a Genes in Colon Cancer. *Laboratory Investigation*. 2001;81(2):217-229.
337. Zhang Y, Xiong Y. Mutations in Human Arf Exon 2 Disrupt Its Nucleolar Localization and Impair Its Ability to Block Nuclear Export of Mdm2 and P53. *Molecular Cell*. 1999;3(5):579-591.
338. Bai P, Xiao X, Zou J, Cui L, Bui Nguyen TM, Liu J, et al. Expression of P14arf, P15ink4b, P16ink4a and Skp2 Increases During Esophageal Squamous Cell Cancer Progression. *Experimental and Therapeutic Medicine*. 2012;3(6):1026-1032.
339. Deshmukh P, Unni S, Krishnappa G, Padmanabhan B. The Keap1–Nrf2 Pathway: Promising Therapeutic Target to Counteract Ros-Mediated Damage in Cancers and Neurodegenerative Diseases. *Biophysical Reviews*. 2017;9(1):41-56.
340. Jiang X, Zhou X, Yu X, Chen X, Hu X, Lu J, et al. High Expression of Nuclear Nrf2 Combined with Nfe2l2 Alterations Predicts Poor Prognosis in Esophageal Squamous Cell Carcinoma Patients. *Modern Pathology*. 2022;35(7):929-937.
341. O'Connor MJ. Targeting the DNA Damage Response in Cancer. *Molecular Cell*. 2015;60(4):547-560.
342. Wang CG, Clifford R, Hewitt SM, Shou J-Z, Alisa M, Goldstein MP, et al. Global Gene Expression Profiling and Validation in Esophageal Squamous Cell Carcinoma (Escc) and Its Association with Clinical Phenotypes. *Clinical Cancer Research*. 2011;17(9):2955-2966.
343. Qixing M, Gaochao D, Wenjie X, Anpeng W, Bing C, Weidong M, et al. Microarray Analyses Reveal Genes Related to Progression and Prognosis of Esophageal Squamous Cell Carcinoma. *Oncotarget*. 2017;8(45):78838.
344. Gallina I, Christiansen SK, Pedersen RT, Lisby M, Oestergaard VH. Topbp1-Mediated DNA Processing During Mitosis. *Cell Cycle*. 2016;15(2):176-183.
345. Sancar A, Tang Ms. Nucleotide Excision Repair. *Photochemistry and Photobiology*. 1993;57(5):905-921.

346. Dev H, Chiang T-WW, Lescale C, de Krijger I, Martin AG, Pilger D, et al. Shieldin Complex Promotes DNA End-Joining and Counters Homologous Recombination in Brca1-Null Cells. *Nature Cell Biology*. 2018;20(8):954-965.
347. Gao S, Feng S, Ning S, Liu J, Zhao H, Xu Y, et al. An Ob-Fold Complex Controls the Repair Pathways for DNA Double-Strand Breaks. *Nature Communications*. 2018;9(1):3925.
348. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell*. 2011;144(5):646-674.
349. Hanahan D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery*. 2022;12(1):31-46.
350. Goel MK, Khanna P, Kishore J. Understanding Survival Analysis: Kaplan-Meier Estimate. *International Journal of Ayurveda Research*. 2010;1(4):274-278.
351. Royston P, Choodari-Oskooei B, Parmar MK, Rogers JK. Combined Test Versus Logrank/Cox Test in 50 Randomised Trials. *Trials*. 2019;20(1):1-10.
352. Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, et al. Gene-Specific Correlation of Rna and Protein Levels in Human Cells and Tissues. *Molecular systems biology*. 2016;12(10):883.
353. Franks A, Airoldi E, Slavov N. Post-Transcriptional Regulation across Human Tissues. *PLoS computational biology*. 2017;13(5):e1005535.
354. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between Protein and Mrna Abundance in Yeast. *Molecular and Cellular Biology*. 1999;19(3):1720-1730.
355. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on Mrna Abundance. *Cell*. 2016;165(3):535-550.
356. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between Differentially Expressed Mrna and Mrna-Protein Correlations in a Xenograft Model System. *Scientific Reports*. 2015;5(1):10775.
357. Maier T, Güell M, Serrano L. Correlation of Mrna and Protein in Complex Biological Samples. *FEBS letters*. 2009;583(24):3966-3973.
358. Prabahar A, Zamora R, Barclay D, Yin J, Ramamoorthy M, Bagheri A, et al. Unraveling the Complex Relationship between Mrna and Protein Abundances: A Machine Learning-Based Approach for Imputing Protein Levels from Rna-Seq Data. *NAR Genomics and Bioinformatics*. 2024;6(1):lqae019.
359. Kosti I, Jain N, Aran D, Butte AJ, Sirota M. Cross-Tissue Analysis of Gene and Protein Expression in Normal and Cancer Tissues. *Scientific Reports*. 2016;6(1):24799.
360. McManus J, Cheng Z, Vogel C. Next-Generation Analysis of Gene Expression Regulation–Comparing the Roles of Synthesis and Degradation. *Molecular BioSystems*. 2015;11(10):2680-2689.
361. Harford JB, Morris DR. Mrna Metabolism & Post-Transcriptional Gene Regulation: John Wiley & Sons; 1997.
362. Wethmar K, Smink JJ, Leutz A. Upstream Open Reading Frames: Molecular Switches in (Patho) Physiology. *Bioessays*. 2010;32(10):885-893.
363. Varshavsky A. The N-End Rule: Functions, Mysteries, Uses. *Proceedings of the National Academy of Sciences*. 1996;93(22):12142-12149.
364. Tang Y-C, Amon A. Gene Copy-Number Alterations: A Cost-Benefit Analysis. *Cell*. 2013;152(3):394-405.
365. Urlinger S, Kuchler K, Meyer TH, Uebel S, Tampé R. Intracellular Location, Complex Formation, and Function of the Transporter Associated with Antigen Processing in Yeast. *European journal of biochemistry*. 1997;245(2):266-272.

366. Banerjee S, Chunder N, Roy A, Sengupta A, Roy B, Roychowdhury S, et al. Differential Alterations of the Genes in the Cdkn2a-Ccnd1-Cdk4-Rb1 Pathway Are Associated with the Development of Head and Neck Squamous Cell Carcinoma in Indian Patients. *Journal of Cancer Research and Clinical Oncology*. 2003;129(11):642-650.
367. Takeuchi H, Ozawa S, Shih CH, Ando N, Kitagawa Y, Ueda M, et al. Loss of P16ink4a Expression Is Associated with Vascular Endothelial Growth Factor Expression in Squamous Cell Carcinoma of the Esophagus. *International Journal of Cancer*. 2004;109(4):483-490.
368. Takeuchi H, Ozawa S, Ando N, Shih C-H, Koyanagi K, Ueda M, et al. Altered P16/Mts1/Cdkn2 and Cyclin D1/Prad-1 Gene Expression Is Associated with the Prognosis of Squamous Cell Carcinoma of the Esophagus. *Clinical Cancer Research*. 1997;3(12):2229-2236.
369. Lin J, Albers AE, Qin J, Kaufmann AM. Prognostic Significance of Overexpressed P16ink4a in Patients with Cervical Cancer: A Meta-Analysis. *PLOS One*. 2014;9(9):e106384.
370. Romagosa C, Simonetti S, Lopez-Vicente L, Mazo A, Lleonart M, Castellvi J, et al. P16ink4a Overexpression in Cancer: A Tumor Suppressor Gene Associated with Senescence and High-Grade Tumors. *Oncogene*. 2011;30(18):2087-2097.
371. Wang Z, Zhang J, Li M, Kong L, Yu J. The Expression of P-P62 and Nuclear Nrf2 in Esophageal Squamous Cell Carcinoma and Association with Radioresistance. *Thoracic Cancer*. 2020;11(1):130-139.
372. Umesh RM, Lahiri M. Overexpression of Topbp1 Leads to Transformation with a Tp53 Mutation of Non-Tumorigenic Breast Epithelial Cells. *BioRxiv*. 2022:2022-2004.
373. Forma E, Krzeslak A, Bernaciak M, Romanowicz-Makowska H, Brys M. Expression of Topbp1 in Hereditary Breast Cancer. *Molecular Biology Reports*. 2012;39(7):7795-7804.
374. Going J, Nixon C, Dornan E, Boner W, Donaldson M, Morgan I. Aberrant Expression of Topbp1 in Breast Cancer. *Histopathology*. 2007;50(4):418-424.
375. Liu K, Bellam N, Lin H-Y, Wang B, Stockard CR, Grizzle WE, et al. Regulation of P53 by Topbp1: A Potential Mechanism for P53 Inactivation in Cancer. *Molecular and Cellular Biology*. 2009;29(10):2673-2693.
376. Michiels S, Laplanche A, Boulet T, Dessen P, Guillonneau B, Méjean A, et al. Genetic Polymorphisms in 85 DNA Repair Genes and Bladder Cancer Risk. *Carcinogenesis*. 2009;30(5):763-768.
377. Moslehi R, Tsao H-S, Zeinomar N, Stagnar C, Fitzpatrick S, Dzutsev A. Integrative Genomic Analysis Implicates Ercc6 and Its Interaction with Ercc8 in Susceptibility to Breast Cancer. *Scientific Reports*. 2020;10(1):21276.
378. Findlay S, Heath J, Luo VM, Malina A, Morin T, Coulombe Y, et al. Shld 2/Fam 35a Co-Operates with Rev 7 to Coordinate DNA Double-Strand Break Repair Pathway Choice. *The EMBO Journal*. 2018;37(18):e100158.
379. Medema RH, Herrera RE, Lam F, Weinberg RA. Growth Suppression by P16ink4 Requires Functional Retinoblastoma Protein. *Proceedings of the National Academy of Sciences*. 1995;92(14):6289-6293.
380. Diehl JA. Cycling to Cancer with Cyclin D1. *Cancer Biology & Therapy*. 2002;1(3):226-231.
381. Clark AS, Karasic TB, DeMichele A, Vaughn DJ, O'Hara M, Perini R, et al. Palbociclib (Pd0332991)—a Selective and Potent Cyclin-Dependent Kinase Inhibitor: A Review of Pharmacodynamics and Clinical Development. *JAMA Oncology*. 2016;2(2):253-260.
382. Lloyd JP. The Evolution and Diversity of the Nonsense-Mediated Mrna Decay Pathway. *F1000Research*. 2018;7.

383. Bond J, Jones C, Haughton M, DeMicco C, Kipling D, Wynford-Thomas D. Direct Evidence from Sirna-Directed “Knock Down” That P16ink4a Is Required for Human Fibroblast Senescence and for Limiting Ras-Induced Epithelial Cell Proliferation. *Experimental Cell Research*. 2004;292(1):151-156.
384. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. Duplexes of 21-Nucleotide Rnas Mediate Rna Interference in Cultured Mammalian Cells. *Nature*. 2001;411(6836):494-498.
385. Zhang HS, Postigo AA, Dean DC. Active Transcriptional Repression by the Rb–E2f Complex Mediates G1 Arrest Triggered by P16ink4a, Tgfβ, and Contact Inhibition. *Cell*. 1999;97(1):53-61.
386. Müer A, Overkamp T, Gillissen B, Richter A, Pretzsch T, Milojkovic A, et al. P14arf-Induced Apoptosis in P53 Protein-Deficient Cells Is Mediated by Bh3-Only Protein-Independent Derepression of Bak Protein through Down-Regulation of Mcl-1 and Bcl-Xl Proteins. *Journal of Biological Chemistry*. 2012;287(21):17343-17352.
387. Suzuki H, Kurita M, Mizumoto K, Nishimoto I, Ogata E, Matsuoka M. P19arf-Induced P53-Independent Apoptosis Largely Occurs through Bax. *Biochemical and Biophysical Research Communications*. 2003;312(4):1273-1277.
388. Milojkovic A, Hemmati PG, Müer A, Overkamp T, Chumduri C, Jänicke RU, et al. P14arf Induces Apoptosis Via an Entirely Caspase-3-Dependent Mitochondrial Amplification Loop. *International Journal of Cancer*. 2013;133(11):2551-2562.
389. Kroemer G. The Proto-Oncogene Bcl-2 and Its Role in Regulating Apoptosis. *Nature Medicine*. 1997;3(6):614-620.
390. Obexer P, Hagenbuchner J, Rupp M, Salvador C, Holzner M, Deutsch M, et al. P16ink4a Sensitizes Human Leukemia Cells to Fas-and Glucocorticoid-Induced Apoptosis Via Induction of Bbc3/Puma and Repression of Mcl1 and Bcl2. *Journal of Biological Chemistry*. 2009;284(45):30933-30940.
391. Shimizu S, Eguchi Y, Kamiike W, Matsuda H, Tsujimoto Y. Bcl-2 Expression Prevents Activation of the Ice Protease Cascade. *Oncogene*. 1996;12(11):2251-2257.
392. Chinnaiyan AM, Orth K, O'Rourke K, Duan H, Poirier GG, Dixit VM. Molecular Ordering of the Cell Death Pathway: Bcl-2 and Bcl-Xl Function Upstream of the Ced-3-Like Apoptotic Proteases (∗). *Journal of Biological Chemistry*. 1996;271(9):4573-4576.
393. Chen D, Tavana O, Chu B, Erber L, Chen Y, Baer R, et al. Nrf2 Is a Major Target of Arf in P53-Independent Tumor Suppression. *Molecular Cell*. 2017;68(1):224-232. e224.
394. Hashimoto-Hachiya A, Tsuji G, Furue M. Antioxidants Cinnamaldehyde and Galactomyces Fermentation Filtrate Downregulate Senescence Marker Cdkn2a/P16ink4a Via Nrf2 Activation in Keratinocytes. *Journal of Dermatological Science*. 2019;96(1):53-56.
395. Furfaro A, Traverso N, Domenicotti C, Piras S, Moretta L, Marinari U, et al. The Nrf2/Ho-1 Axis in Cancer Cell Growth and Chemoresistance. *Oxidative Medicine and Cellular Longevity*. 2016;2016(1):1958174.
396. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000;28(1):235-242.
397. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. Colabfold: Making Protein Folding Accessible to All. *Nature Methods*. 2022;19(6):679-682.
398. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The Universal Protein Resource (Uniprot). *Nucleic Acids Research*. 2005;33(suppl_1):D154-D159.

399. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. Ucsf Chimera—a Visualization System for Exploratory Research and Analysis. *Journal of Computational Chemistry*. 2004;25(13):1605-1612.
400. Byeon I-JL, Li J, Ericson K, Selby TL, Tevelev A, Kim H-J, et al. Tumor Suppressor P16ink4a: Determination of Solution Structure and Analyses of Its Interaction with Cyclin-Dependent Kinase 4. *Molecular Cell*. 1998;1(3):421-431.
401. Russo AA, Tong L, Lee J-O, Jeffrey PD, Pavletich NP. Structural Basis for Inhibition of the Cyclin-Dependent Kinase Cdk6 by the Tumour Suppressor P16ink4a. *Nature*. 1998;395(6699):237-243.
402. Hossain MS, Roy AS, Islam MS. In Silico Analysis Predicting Effects of Deleterious Snps of Human Rassf5 Gene on Its Structure and Functions. *Scientific Reports*. 2020;10(1):14542.
403. Singh A, Thakur M, Singh SK, Sharma LK, Chandra K. Exploring the Effect of Nssnps in Human Ypel3 Gene in Cellular Senescence. *Scientific Reports*. 2020;10(1):15301.
404. Wang J, He X, Luo Y, Yarbrough WG. A Novel Arf-Binding Protein (Lzap) Alters Arf Regulation of Hdm2. *Biochemical Journal*. 2006;393(2):489-501.
405. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods*. 2010;7(4):248-249.
406. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using Polyphen-2. *Current Protocols in Human Genetics*. 2013;76(1):7-20.
407. Flanagan SE, Patch A-M, Ellard S. Using Sift and Polyphen to Predict Loss-of-Function and Gain-of-Function Mutations. *Genetic Testing and Molecular Biomarkers* 2010;14(4):533-537.
408. Capriotti E, Fariselli P, Casadio R. I-Mutant2. 0: Predicting Stability Changes Upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Research*. 2005;33(suppl_2):W306-W310.
409. Ng PC, Henikoff S. Sift: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Research*. 2003;31(13):3812-3814.
410. Li X, Tian D, Guo Y, Qiu S, Xu Z, Deng W, et al. Genomic Characterization of a Newly Established Esophageal Squamous Cell Carcinoma Cell Line from China and Published Esophageal Squamous Cell Carcinoma Cell Lines. *Cancer Cell International*. 2020;20(1):1-13.
411. Michieli P, Chedid M, Lin D, Pierce JH, Mercer WE, Givol D. Induction of Waf1/Cip1 by a P53-Independent Pathway. *Cancer Research*. 1994;54(13):3391-3395.
412. Al-Khalaf HH, Aboussekhra A. P16 Controls P53 Protein Expression through Mir-Dependent Destabilization of Mdm2. *Molecular Cancer Research*. 2018;16(8):1299-1308.
413. McIlwain DR, Berger T, Mak TW. Caspase Functions in Cell Death and Disease. *Cold Spring Harbor perspectives in biology*. 2013;5(4):a008656.
414. Gustafson S, Proper JA, Bowie EW, Sommer SS. Parameters Affecting the Yield of DNA from Human Blood. *Analytical Biochemistry*. 1987;165(2):294-299.
415. Li H, Durbin R. Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform. *Bioinformatics*. 2010;26(5):589-595.
416. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*. 2016;44(D1):D481-D487.
417. Veale R, Thornley A. Increased Single Class Low-Affinity Egf Receptors Expressed by Human Esophageal Squamous Carcinoma Cell-Lines. *South African Journal of Science*. 1989;85(6):375-379.

418. Karakasheva TA, Lin EW, Tang Q, Qiao E, Waldron TJ, Soni M, et al. Il-6 Mediates Cross-Talk between Tumor Cells and Activated Fibroblasts in the Tumor Microenvironment. *Cancer research*. 2018;78(17):4957-4970.
419. Strauss WM. Preparation of Genomic DNA from Mammalian Tissue. *Current Protocols in Molecular Biology*. 1998(1):2.2. 1-2.2. 3.
420. Kumar A, Chordia N. In Silico Pcr Primer Designing and Validation. *PCR Primer Design*. 2015:143-151.
421. Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, et al. Idt Scitools: A Suite for Analysis and Design of Nucleic Acid Oligomers. *Nucleic Acids Research*. 2008;36(suppl_2):W163-W169.
422. Lam C-w, Mak CM. Allele Dropout Caused by a Non-Primer-Site Snv Affecting Pcr Amplification—a Call for Next-Generation Primer Design Algorithm. *Clinica Chimica Acta*. 2013;421:208-212.
423. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. Sift Web Server: Predicting Effects of Amino Acid Substitutions on Proteins. *Nucleic Acids Research*. 2012;40(W1):W452-W457.
424. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly Accurate Protein Structure Prediction with Alphafold. *Nature*. 2021;596(7873):583-589.