The Analysis of Genetic Aberrations in South African Oesophageal Squamous Cell Carcinoma Patients.

By

Victoria Alexandra Patten

Submitted to the University of Cape Town

in fulfilment of the requirements for the degree:

Doctor of Philosophy (Medical Biochemistry)

Department of Integrative Biomedical Sciences

Faculty of Health Sciences

University of Cape Town

Supervisor: Professor M.I. Parker[1]

Co-Supervisors: Associate Professor D. Hendricks[1] and Doctor H. Bendou[2]

1. Division of Medical Biochemistry and Structural Biology, University of Cape Town
2. Division of Computational Biology, University of Cape Town

April 2023

Declaration

I, **Victoria Alexandra Patten**, 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.

Date: 30 April 2023

Abstract

Estimates for 2017 indicate that 20% of cancers globally are gastrointestinal tract (GIT) cancers, with oesophageal cancer being the 8th most common cancer. Oesophageal squamous cell carcinoma (OSCC) occurs in the upper to mid oesophagus and is present at high incidence in developing countries including South Africa. There are no early symptoms, resulting in late diagnosis and poor prognosis.

In this study, tumour and blood DNA was obtained from 35 OSCC patients and subjected to whole genome sequencing (WGS). Bioinformatics analysis pipelines were designed to identify the possibility of novel viral insertions, investigating Human Endogenous Retroviruses (HERV's) insertions alongside the presence of somatic mutations in patient samples. The aims being to identify integration of any foreign DNA , to investigate if there is any linkage between HERV insertion and somatic mutations, and to identify any somatic mutations of potential interest in the OSCC cohort.

The novel virus investigations however, proved to be inconclusive and there appeared to be no link between HERV insertions and somatic mutations present in the patients. Very significantly, it was determined that numerous somatic mutations were present in the MUC3A gene of the patient cohort, an interesting observation as no such previous association with OSCC has been recorded. *MUC3A* is a membrane-bound glycoprotein component of mucous gels and its aberrant expression has been correlated with invasion and metastasis in a variety of other cancers. However, due to the complexity of the particular gene sequence and the known inconsistencies of variant calling performed on complex data sets, these mutations should be viewed with extreme caution as they are likely to be false positives. Analysis of RNA-seq data showed a 4.6 log2 fold increase in *MUC3A* expression in the tumour samples of these OSCC patients, with a *P*-adjusted value of 7.05e^{-06}, suggesting highly significant differential gene expression. Functional enrichment analysis further showed that *MUC3A* was significantly associated with one of the top 5 gene ontologies (extracellular matrix structural constituent) for molecular function ontology class together with a number of collagen (COL) and MMP genes known to play a role in oncogenic progression and membrane stiffness. GSEA and KEGG analysis indicated predominantly chemokine/cytokine pro-inflammatory enriched pathways. Immunohistochemistry staining showed 10 out of 13 of the samples had no detectable levels of MUC3A protein, suggesting that the production of a non-functional truncated protein may lead to the upregulation of *MUC3A* expression that could possibly play a role in downstream pro-oncogenic signalling.

Acknowledgements

- My deepest gratitude to my supervisor, Prof M.I. Parker, for his expert guidance, supervision and constant encouragement in helping me to achieve my research goals.
- My co-supervisor, Prof D. Hendricks, my sincere gratitude for the ideas and advice through the course of this project. He has been a valued support and source of guidance in the field of cancer research.
- My co-supervisor, Dr H. Bendou, my most sincere thanks and appreciation for the tutelage and guidance in the bioinformatics training and analysis component of this project. His patience and wealth of knowledge in the field have been invaluable and essential to my understanding and execution of the data mining processes implemented.
- All the brave patients who were willing to participate in this Doctoral research project and so generously donated biopsies and blood samples.
- My beloved parents and sister who have stood by me and supported me through every decision that has led me to this point. Their love and encouragement have been the pillars of strength in my life and it is through them that I have had the courage to pursue my goals.
- Angus Comrie, for the unwavering support, patience and encouragement, He has been a source of strength and comfort.
- Jeremy Smith and Dane Kennedy from Ilifu, for going above and beyond in helping me with the bioinformatics pipelines and Ilifu cloud computing. Without your help I would have been utterly lost.
- Hendrina Shipanga and Humaira Lambarey, who have always provided help and assistance whenever I've asked. I am very grateful for their kindness and constant willingness to help.
- The Division of Medical Biochemistry at UCT, my deepest gratitude to all who are part of this wonderful working division for all the advice, interesting conversations and laughs along the way. It has been a privilege completing my research with you.
- The University of Cape Town who invested in me and this project, without whom, none of this research would be possible.
- All funders and collaborators involved in this project. The Newton Fund, The South African Medical Research Council, The Wellcome Sanger Institute, The South African National Bioinformatics Institute, and The Centre for Proteomic and Genomic Research.

Table of Contents

List of Abbreviations

AID	: Activation-induced Cytidine Deaminase
BAM	: Binary Alignment Map
BLAST	: Basic Local Alignment Search Tool
bp	: Base Pairs
BSA	: Bovine Serum Albumin
BWA	: Burrows-Wheeler Aligner
CAF	: Central Analytical Facility
cDNA	: Complementary DNA
cGAMP	: Cyclic GMP-AMP
cGAS	: Cyclic GMP-AMP Synthase
ChIP-seq	: Chromatin Immunoprecipitation followed by Sequencing
CMV	: Cytomegalovirus
CNA	: Copy Number Alteration
CNV	: Copy Number Variant
CPGR	: Centre for Proteomic and Genomic Research
dbSNP	: Single Nucleotide Polymorphism Database
DMEM	: Dulbecco’s Modified Eagle Medium
DMSO	: Dimethyl Sulfoxide
DNA	: Deoxyribonucleic Acid
dsDNA	: Double Stranded DNA
EBNA1	: Epstein-Barr Nuclear 1
EBV	: Epstein-Barr Virus
EDTA	: Ethylenediaminetetraacetic acid
EGF	: Epidermal Growth Factor
EGTA	: Egtazic acid
ERK	: Extracellular Regulated Kinase
ERV	: Endogenous Retrovirus
FBS	: Foetal Bovine Serum
FCS	: Foetal Calf Serum
FDR	: False Discovery Rate

FFPE : Formalin Fixed Paraffin Embedded

GATK : Genome Analysis Toolkit

GIT : Gastrointestinal Tract

GORD : Gastro-oesophageal Reflux Disease

GWAS : Genome-wide Association Studies

HERV : Human Endogenous Retrovirus

HBV : Hepatitis B Virus

HGD : High grade Dysplasia

HTLV-1 : Human T-cell Leukaemia Virus

HPV : Human Papilloma Virus

HREC : Health Research Ethics Committee

HSV : Herpes Simplex Virus

ICA : Inter-Correlation Analysis

IHC : Immunohistochemistry

IKK : Inhibitor of NF-κB Kinase

Indel : Insertion/deletion

KSHV : Kaposi Sarcoma-Associated Virus

LFC : Log2 Fold Change

LGD : Low Grade Dysplasia

LTR : Long Terminal Repeat

MAPK : Mitogen-Activated Protein Kinase

MCPyV : Merkel Cell Polyomavirus

MEK : Mitogen-Activated Protein Kinase Kinase

mTOR : Mechanistic Target of Rapamycin

NF-κB : Nuclear Factor-κB

NGS : Next Generation sequencing

NTC : No Template Control

OAC : Oesophageal Adenocarcinoma

ORF : Open Reading Frame

OSCC : Oesophageal Squamous Cell Carcinoma

PBS : Phosphate Buffered Saline

PCA : Principal Components Analysis

PCR	: Polymerase Chain Reaction
PI3K	: Phosphatidylinositol 3-kinase
PON	: Panel of Normals
Pro	: Proline
pRB	:Retinoblastoma Protein
RAS	: Retrovirus-associated DNA Sequences
rle	: Relative Log Expression
RNA-seq	: RNA sequencing
ROS	:Reactive Oxygen Species
RT	:Reverse Transcriptase
RT-PCR	: Real Time – Polymerase Chain Reaction
SANBI	: South African National Bioinformatics Institute
SD	: Standard Deviation
SDS-PAGE	: Sodium Dodecyl Sulphate – Polyacrylamide Gel
SEA domain	: Sea urchin, enterokinase, agrin domain
SEM	: Standard Error of the Mean
Ser	: Serine
SMG	: Significantly Mutated Gene
SNP	: Single Nucleotide Polymorphism
SNV	: Single Nucleotide Variant
SQL	: Structured Query Language
ssDNA	: Single Stranded DNA
SV	: Structural Variant
TE	: Transposable Elements
TBS	: Tris Buffered Saline
TFF	: Trefoil Family of Receptors
Thr	: Threonine
Tm	: Melting Temperature
tmm	: Trimmed Mean of M-Values of edge-R
TR	: Tandem Repeats
Tyr	: Tyrosine
UCT	: University of Cape Town

VCF	: Variant Calling File
VEGF	: Vascular Endothelial Growth Factor
vsf	: Variance Stabilising Transformation
WES	: Whole Exome Sequencing
WGS	: Whole Genome Sequencing
WITS	: University of the Witwatersrand

List of Symbols

%	-	Percentage
°C	-	Degrees Celsius
A	-	Amperes
g	-	Grams
kDa	-	Kilo Dalton
kg	-	Kilogram
L	-	Litre
m	-	Meter
M	-	Molar
mg	-	Milligram
min	-	Minute
ml	-	Millilitre
mM	-	Milimolar
mmHg	-	Pressure (Millimetres of Mercury)
mW	-	Miliwatt
ng	-	Nanogram
nM	-	Nanomolar
v	-	Volume
V	-	Volts
α	-	Alpha
β	-	Beta
µg	-	Microgram
µl	-	Microliter
µM	-	Micromolar

List of Figures

Chapter 1: Literature Review and Introduction

Chapter 3: Preliminary Bioinformatics Analysis of Novel Viral Insertions

Chapter 4: Analysis of Human Endogenous Retrovirus (HERV) Integration in the Human Genome

Chapter 5: Identification of Somatic Mutations

Chapter 6: Analysis of Gene Expression

List of Tables

Chapter 1: Literature Review and Introduction

Chapter 2: Methods and Techniques

Chapter 3: Preliminary Bioinformatics Analysis of Novel Viral Insertions

Chapter 4: Analysis of Human Endogenous Retrovirus (HERV) Integration in the Human Genome

Chapter 5: Identification of Somatic Mutations

Chapter 6: Analysis of Gene Expression

Appendix 1

Appendix 2

Appendix 3

Appendix 4

Appendix 5

Appendix 6

Appendix 7

Chapter 1: Literature Review and Introduction

1.1 The Burden of Oesophageal Cancer

Cancer is a leading cause of death globally, and is widely regarded as a formidable barrier in the pursuit of increasing life expectancy in all countries of the world [1]. Cancer of the oesophagus is one of the most frequently reported malignancies, generally described as the seventh most common type of cancer globally in terms of incidence, and constitutes the sixth leading cause of cancer-related deaths. It accounts for more than 500 000 cases of mortality each year, and in 2020 was responsible for one in every 18 cancer deaths [2]. Diagnosis often only occurs during advanced stages of the disease due to a lack of early clinical symptoms [3].

Oesophageal cancer comprises two distinct epidemiologically, histologically and pathologically diverse subtypes; oesophageal adenocarcinoma (OAC) and oesophageal squamous cell carcinoma (OSCC) [4–6]. OAC typically occurs in the distal third of the oesophagus, close to the gastroesophageal junction while OSCC occurs higher up in the proximal two thirds of the oesophagus, originating from the squamous epithelial lining of the oesophagus [7–9].

1.1.1 Risk Factors

Both subtypes of oesophageal cancer are frequently preceded by chronic inflammation in the oesophagus, leading to the disruption of normal cell signalling and growth [10]. Risk factors commonly associated with OSCC include heavy alcohol use, tobacco smoking, certain vitamin and iron deficiencies (A, C and zinc), consumption of excessively hot beverages, infection (such as human papillomavirus), poor oral health and low socioeconomic status [11–14] Conversely, OAC is associated with distinctly different risk factors including Barett's oesophagus, obesity and gastro-oesophageal reflux disease (GORD) [4,5,10,14].

1.1.2 Development of Oesophageal Cancer

The development and mechanism of carcinogenesis of OSCC progresses from normal epithelium to basal cell hyperplasia and dysplasia (low to high grade), followed by carcinoma *in situ* and then invasive OSCC involving either direct mechanical injury or exposure of the

oesophageal mucosa to carcinogens, or a combination of both [6,8,9]. OAC development and progression follows from metaplastic changes in the oesophageal epithelium stemming from chronic GORD and progresses from normal squamous cell epithelium, to columnar epithelium (known as Barrett's oesophagus), to low then high grade dysplasia and in due course, invasive adenocarcinoma [6]. Biologically, OSCC is reported to closely resemble head and neck squamous cell carcinomas, sharing many of the distinguishing characteristics. OAC on the other hand more closely resembles chromosomally unstable gastric adenocarcinoma [6,9].

1.1.3 Incidence

The two types of oesophageal cancer OAC and OSCC vary geographically and between populations, with OSCC being the most common subtype present in a high incidence belt prevalent in developing countries of Eastern Asia, and Eastern and Southern Africa. Approximately 90% of oesophageal cancer patients in these regions present with OSCC [2,4,5,10]. Globally, approximately 87% of all oesophageal cancers are defined as OSCC, while OAC constitutes only 11% of cases [15]. However, over the past four decades, the incidence of OSCC has been on the decrease, while OAC incidence, primarily observed in developed western countries such as North America, Europe and Australia, has been rising sharply and is expected to sustain this trend [5,15]. In Africa however, incidence rates of OSCC appear to be increasing with some of the highest incidence rates globally reported in East Africa [16]. Ferlay *et al* reported that globally, we can expect to see a 35% increase in all oesophageal cancers between 2018 and 2030, with an accompanied increase in number of deaths during the same time [17].

As with other digestive system cancers, a higher incidence is reported in men than in women, with men accounting for up to 70% of all oesophageal cancers worldwide. Men have a three to four times higher risk of developing OSCC than women, and seven to ten times higher risk of developing OAC [2,10]. These increased rates in men can possibly be explained by a higher prevalence of abdominal obesity and GORD, along with the associations of higher testosterone levels where oestrogen is reported to have a protective effect in females [18]. Figure 1.1 shows a graphical representation of regional incidence rates for both men and women. Incidence rate for both sexes is reported to increase generally with age predominantly affecting individuals older than 65 years [2,10].

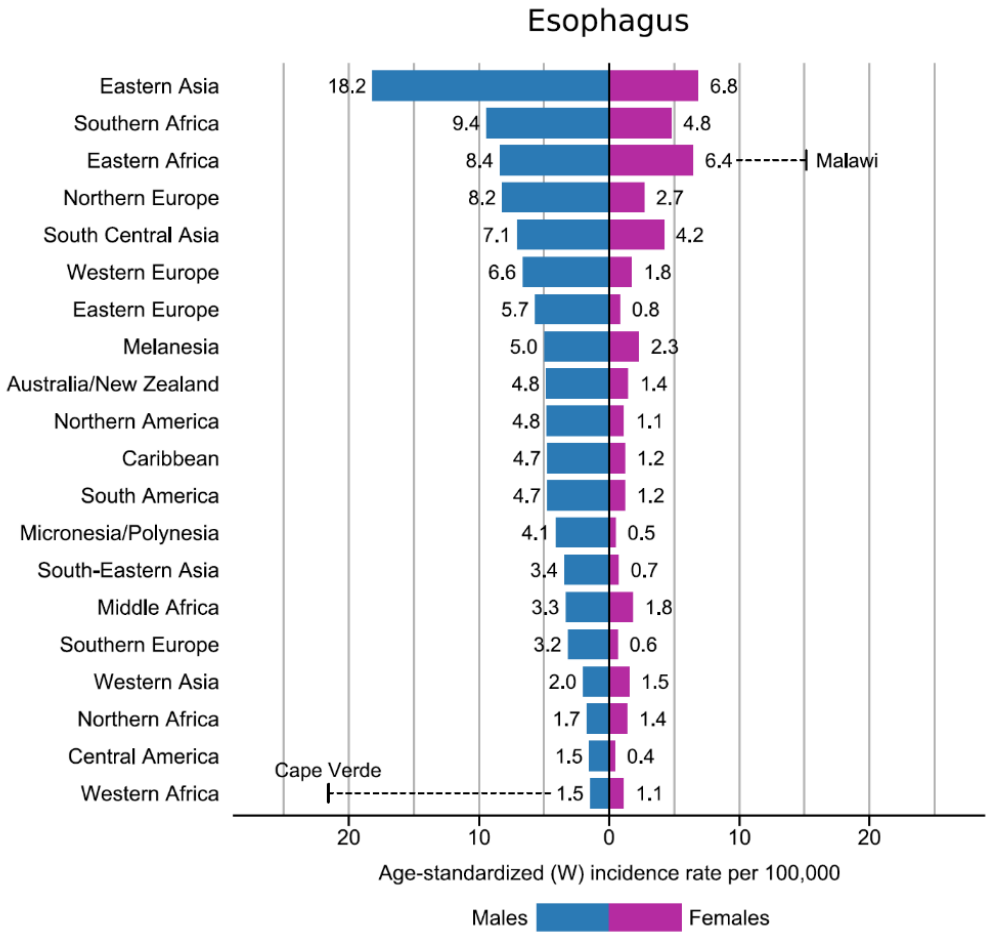


Figure 1.1: Regional-specific incidence age-standardised rate by sex for oesophageal cancer in 2020 taken from GLOBOCAN statistics reported for 2020 [2].

1.1.4 Prognosis

Prognosis of oesophageal cancer varies greatly between geographical regions but is generally considered poor due to late diagnosis, with an overall five year survival rate of <20% [5,6,10]. OAC has typically shows a better overall median survival when compared with OSCC, particularly in the early stages of the disease [19,20]. In 2020 it was reported that age-standardised mortality rates for both sexes globally (per 100 000) are highest in Eastern Asia, Eastern African, and Southern Africa [10]. This is plausibly due to these being high incidence regions, as well as socio-economic factors and the majority of the affected populations coming from rural areas with limited access to health care systems and later diagnosis.

1.1.5 Genetics

Analyses of comprehensive mutational changes using high throughput sequencing have detected widespread multiple genetic mutations, identified as driver mutations, associated with oesophageal cancers [21–23]. OAC is commonly associated with GORD and complications arising therefrom. OSCC however, is thought to have a more exposure driven causality where direct and prolonged exposure to various carcinogenic compounds in the upper digestive tract (i.e. tobacco smoke, hot beverages, alcohol etc) is likely to have a modulatory effect on the genetic and epigenetic makeup of the exposed squamous epithelial cells, facilitating neoplastic transformation [24].

Compelling evidence shows that the genomic alteration profiles of OAC and OSCC vary considerably [25] with genes commonly altered in OAC including *ERBB2, KRAS, EGFR, SMAD4, FGF3/4/19, VEGFA, CCNE1* and *GATA4/6* [9], and in OSCC, *TP53, AJUBA, CDKN2A, KMT2D, ZNF750, FAT1, NOTCH1*, *NOTCH 3, PIK3CA, NFE2L2, RB1, KDM6A, FBXW7, CREBBP* and *TGFBR2* more frequently altered [26].

The dysregulation of *TP53* and other genes is reported as a prominent characteristic of OSCC where changes in such genes can already be detected in precursor lesions [27]. Somatic mutations in *TP53* are present in more than 83% of OSCC patients, with the over expression of *EGFR* shown in up to 76% of cases, *CCND1* in 46% of cases and *CDK4/CDK6* in 24% of cases common in OSCC, associated with poor prognosis [3,28,29]. Genetic mutations detected in OSCC are most commonly associated with the cell cycle, DNA repair mechanisms, growth factor signalling and apoptosis [30]. Mutations in cell cycle control genes (*CKN2A, RB1, NFE2L2, CHEK1 and CHEK2*), and in genes affecting differentiation (*NOTCH1* and *NOTCH3*) have been associated with 2-10% of OSCC cases[21,22,29].

Conversely, the amplification of *CCNE1* and *Cyclin E*, along with mutations in *MGST1* have been implicated in the development of OAC [3]. Xie *et al* suggested that identifying genetic variants influencing the risks of oesophageal cancer might provide a better understanding of the biology of the disease and how these genetic changes interact with environmental factors in the aetiology of the disease, especially OSCC, and may also assist in risk stratification for future tailored and targeted prevention of this cancer [14].

The vast and significant clinical, histological and etiological differences between OSCC and OAC give rise to major variances in incidence and characteristics further dependent on geographical and socioeconomic factors, making full understanding and prevention of

oesophageal cancer very difficult [10]. Table 1.1 shows the defining differences between OSCC and OAC.

Furthermore, studies predominantly focusing on a Kenyan population have also demonstrated the importance and contribution of genetic factors to the development and progression of OSCC, and assisting in the predisposition of individuals with a family history to oesophageal cancer [31]. Kenya, like South Africa, is an area of very high risk for OSCC.

Table 1.1: Key differences between OSCC and OAC as adapted from Smyth *et al.* (2017) [6]

Differences	OSCC	OAC
Geographical Distribution	Most common in East Asia, Eastern and Southern Africa	Most common in developed regions in western Europe, North America and Australia
Main Risk Factors	Tobacco smoking, alcohol use, thermal injury and regional micronutrient deficiency	Acid or bile reflux, Barrett's oesophagus, and central or visceral obesity
Tumour Location	Throughout the oesophagus, more common in the upper and middle third.	More common in the distal oesophagus.
Frequent comorbidities	Liver cirrhosis, chronic obstructive pulmonary disease, synchronous and metachronous cancer of the aero-digestive tract, and atherosclerosis	Obesity and atherosclerosis
Diagnosis and symptoms	Late diagnosis with no early warning symptoms	Early diagnosis. GORD and Barrett's oesophagus
Curative treatment	Definitive chemoradiotherapy, or chemoradiotherapy followed by surgery	Neoadjuvant or perioperative chemotherapy followed by surgery, or neoadjuvant chemoradiotherapy followed by surgery
Palliative treatment	Chemotherapy, radiotherapy or stenting	Chemotherapy (plus trastuzumab if HER2-positive radiotherapy or stenting

1.1.6 Incidence and Risk in South Africa

Recorded Incidences of oesophageal cancer weren't always high in South Africa, and it was only in the later decades of the 1900's into the 2000's where a rapid rise in the number of diagnosed cases was detected and reported [32,33]. However, this does not take into account that these statistics reflect an era where access to sufficient health care for black South Africans, particularly in rural areas, was severely lacking which could have skewed the statistics. Oesophageal cancer has become the third most commonly diagnosed cancer

among Black South African's and is further reported as the fourth highest cause of death in males of a mixed-race population [34]. As with the incidence trends reported for other high incidence regions, rates in South Africa increase noticeably with age, specifically between the ages of 60 and 70 years [35]. The former Transkei region in the Eastern Cape Province is largely considered as the most affected area in the country and continually rising rates over the past recent decades is largely due to changes in environmental factors as well as possibly increased exposure to carcinogens [36,37].

The two biggest contributing risk factors in the development of this cancer are in keeping with the rest of the high incidence belt, being high tobacco usage and high alcohol intake [38,39]. However, there are other contributing factors that are more unique to the South African population and these include the exposure to environmental smoke from extensive utilisation of wood and charcoal for cooking and heating [40,41], the use of wild herbs such as *Solanum nigrum*, increased incidence of human papilloma virus (HPV) infection [36,42], and the high consumption of home-made beer brewed with maize instead of sorghum, frequently found to be contaminated with fungal mycotoxins [38]. These factors combined with the now frequent adoption of a more western diet including fats and animal proteins, are seen as important risk factors for oesophageal cancer. Analyses of populations in the Transkei and the Kwa-Zulu Natal Province have further indicated that poorer people are at a higher risk due to their nutritionally deficient diets and fungal contamination of staple maize and traditionally cooked wild plants, combined with high volumes of alcohol and tobacco use [36,37,43].

1.2 Viruses at Large

1.2.1 Viruses and Integration

Viruses are the most numerous and most diverse organisms on modern earth, evoking a profound influence on the evolution of life [44–46]. Fossil and genomic data demonstrates the presence of common viral genes in the genomes of organism known to have diverged tens of millions of years ago, together with common antiviral genes strongly indicate that specific viruses and viral defences are millions of years old [47,48].

Viral particles consistently outnumber other organisms by one to two orders of magnitude in marine, soil and animal-associated environments [49], and approximately 10^{30} viruses are found in oceans alone, with an abundance that exceeds that of archaea and bacteria by at least 15-fold [50]. It is often postulated that evolution of life would not have been possible without viruses [51], and two driving mechanisms of evolution occurred through the provision

of strong selective pressure for resistance to viral infection, alongside the movement of exogenous genetic materials into cells [46]. Exceptionally strong evolutionary pressure is derived through the evolution of host resistance to viral infections resulting in host death, and since viruses need hosts for replication, there also exists very strong evolutionary pressure for viruses to overcome host cell defences [46,52].

Accordingly, humans, like all other organisms on earth, are constantly bombarded with viruses leading to numerous disease states [53]. At least 30 families of viruses have been identified to be infectious to humans and other mammals. These require the manufacture of progeny virions within a host cell to facilitate horizontal viral transmission with subsequent attachment and disassembly in a host for the next infectious cycle [54]. Many viruses have the ability to integrate their DNA into the genetic material of their host species [55], and a vast degree of invasion of the human genome by RNA and DNA viral sequences are reported to have become integrated into the vertebrate genome. It is speculated that viral DNA could account for between 8% and 40% of the human genome, and may have the potential to trigger tumorigenesis [53,56,57].

In a review by Feschotte and Gilbert (2012)[57], it was reported that recent studies have suggested that all major types of eukaryotic viruses can be integrated, leading to vertical transmission and potential fixation within a host population. Furthermore, it has been reported that viruses are likely largely responsible for the generation of much of the variation driving evolution and the selection of new traits, including gene duplication, exon shuffling and the diversification of non-coding regulatory elements [58].

1.2.2 Viral Modification of Host Genomes

Described as arduous pathogens, viruses present a unique challenge to host recognition systems [58]. Because all organisms are susceptible to infection by viruses, all organisms therefore have dedicated detection and prevention systems against nucleic acid parasitism with varying responses dictated by the needs of the organism [58,59]. The detection of viral genomes composed of nucleic acids has been favoured through the evolution of host cell receptors designated for this purpose [60,61].

Viruses frequently promote a parasitic and destructive lifestyle. This perpetual conflict with host cells allows for hijacking and manipulation of the host machinery through the continuous selection of genetic adaptations within the virus. Furthermore, antiviral immune defences of the host are often counteracted favouring the expression of viral genes[57,58]. A review by

Rapicetta *et al* (2002) echoed this observation, stating that multiple viral mechanisms and viral proteins profoundly impact infected host cells and their immune responses to ensure successful viral infection[62]. Conversely, recent discoveries have shown a more symbiotic coexistence of some (if not most) viruses with their hosts (discussed in reviews [63–65]). In fact, host-virus interactions can range from parasitism to clear mutualism [66] where viral infection is shown to benefit hosts, mostly through the contribution of genes allowing for survival under otherwise detrimental conditions (reviewed in [63,67]). Upon infection, genes specific for viral replication are expressed, but, frequently viral genomes also encode genes whose products lead to the alteration of host metabolism allowing hosts to adapt to new environments [46].

1.2.3 Viruses in Cancer

1.2.3.1 Human Oncogenic Viruses

Cancers, including those of the gastrointestinal tract (GIT) result from intricate interactions between environmental factors and host genetic factors in response to chronic inflammation by cigarette smoke, alcohol, ultra-violet light, asbestos and x-rays [68,69]. Several viruses have also been reported as playing a significant role in the multistage development of human cancers, evoking such responses through the expression of certain proteins, nuclear antigens and RNA's to manipulate host immune responses and promoting chronic inflammation [70,71]. In fact, reports suggest that 15-20% of all cancers are associated with viral infections [72,73]. Oncogenic viruses can contribute to different steps in the process of carcinogenesis, and the association and involvement of certain viruses in different cancers can be anywhere between 15% and 100% [73].

1.2.3.2 Mechanisms of Viral Oncogenesis

Over time, viruses have evolved ways to exploit and subvert their host cell machinery to allow for their own propagation, while at the same time, hosts have evolved mechanisms to allow maintenance of the integrity of their cellular environment for life-sustaining functions. In this way, the fate of both the host and the invading pathogen is decided by the extent to which either one is in control of the critical homeostatic pathways [74]. It is only when cumulative changes or external forces causing DNA damage disrupt the fragile equilibrium, that the virus then shifts and drive uncontrolled cellular proliferation, accumulation of mutations and the evasion of antitumour immunity [74].

Four main mechanisms by which viruses lead to oncogenesis in humans has been described. For a detailed review of these mechanisms, see Krump and You, (2018) [74]. The mechanisms and pathways are listed below in Table 1.2.

Table 1.2: Mechanisms of viral oncogenesis and the signalling pathways frequently affected.

1	**Targeting Tumour Suppressor Pathways**
2	**Targeting Host Signalling Pathways** • PI3K-AKT-mTOR pathways • MAPK Signalling • Notch Signalling • WNT/B-catenin signalling • NFκB signalling
3	**Exploiting Host DNA Damage Response**
4	**Manipulation of Host immune System**

1.2.3.3 Endogenous Retroviruses and Transposable Elements

Some viruses not only infect host cells, but are capable of inserting their genomes into that of their hosts. The best-studied of these are retroviruses whose replication cycles are dependent on inserting their own genomes into their host's genome. In this way, a copy of the virus genome is generated within the host in a form referred to as a 'provirus' which can be triggered to replicate and produce new virions [46]. Through replication of proviruses, retroviruses are able to acquire cellular genes of the host [75].

Over recent decades a number of viral families have been shown to transmit vertically through the human genome, retaining some of the hallmarks of exogenous retroviruses, as with the human immunodeficiency virus (HIV) and human T-cell leukaemia virus (HTLV) and their genomic structure. [76,77]. Between 8%-9% of the modern human genome is made up of retroviruses that became integrated through exogenous retroviral infection over the course of millions of years, becoming stable and part of what is now our inherited genetic material [78–81]. These are referred to as Human endogenous retroviruses (HERV's) that are not infectious but over the course of time have been subjected to repeated amplification and transposition events that have led to the presence of multiple copies of proviruses within the DNA [54]. As HERV's have been present in our genomes for around 30 million years, it is plausible that some transposition events have modified and modulated host gene expression, conferring a selective advantage to the host. However, the potential for negative

impact does exist as the opportunity for mutation accompanies transposition, playing a possible role in cancer development [82,83].

Transposable elements (TE) are frequently described as discrete DNA sequences that maintain the ability to translocate within the human genome from one position to another, leading to the development of interspersed repeats [84–86]. Classification of these TE's is dependent on their ability to encode genes for transposition, and can either be autonomous, such as Long Interspersed Nuclear Elements (LINE's), or non-autonomous, such as Short interspersed Nuclear Elements (SINE's). LINE's effectively encode all sequences that move in the genome, while SINE's are more structurally deficient and dependent on LINE's for movement. SINE's are reported to have evolved independently to LINE's, but in parallel, with the most common SINE family being Alu elements. There also exist some TE's where the distinction between autonomous and non-autonomous remains unclear, such as HERV's. These however, have been particularly well characterised and it is commonly believed that their integration leads to phenotypic alterations in the human genome. Thus variations arising in host genomes can often be attributed to these potent, broad-spectrum mutator elements [84].

Over time, the accumulation of multiple mutations acquired in these viral elements (including deletions, insertions, truncations and frameshifts) has led to the assumption that they are commonly inactive and unable to replicate [87,88]. However, studies in live patients rather than immortalised cell lines have suggested that some HERV's may in actual fact still have open reading frames (ORF) with a potential for protein expression and capabilities for replication in modern humans [79,89,90]. These include the HERV-K family viruses, reportedly some of the most recently biologically active retroviral elements in the human genome with the least number of mutations with the ability to encode functional retroviral proteins, producing retrovirus-like particles [88,91,92]. Findings have further suggested a significant evolutionary role as well as potential roles in various diseases [79,93,94].

HERV's are classified into 22 independently acquired families, some of which (such as HERV-K) are associated with the development autoimmune diseases or a number of different cancers [95], such as breast cancer [96], lung cancer [97], prostate cancer [98], hepatocellular carcinoma [99], melanomas [100], germ cell tumour [101], leukaemia [102], and lymphoma [103]. The disruption of host gene regulation through hijacking and manipulation by these retroviral elements can influence the expression of the hosts own genes, posing a long-lasting burden on the genome [57], while it is speculated that HERV's might transform benign cells and induce cancer though several mechanisms including:

(1) Hypermethylation leading to the activation of HERV sequences

(2) expression of HERV-encoded oncogenes Rec and NP9 which interact with transcriptions factors and activate immune-suppressor pathways

(3) mutational insertions causing the inactivation of tumour suppressor genes

(4) homologous recombination

(5) recruitment of transcription factors, oncogenes or growth factors via LTR's for retroviral gene transcription

(6) the induction of syncytial formation by Env protein aiding in the propagation and progression of cancer cells [95].

The discovery of significantly elevated expression of HERV elements in cancer cells has driven research to investigate using HERVs as biomarkers for malignant transformation, staging and prognosis of cancers [104–106]. However, even though many studies have demonstrated the expression of HERV in multiple cancer tissues, the exact causal role in cancer development remains controversial [95]. Several factors have been proposed to trigger HERV protein expression in cancer, and these include UV radiation, oestrogen hormone, and smoking, while the intrinsic activation of the MAPK and p16INK4A-CDK4 pathways lead to HERV expression in melanoma [107]. It has also been reported that the transposable ability of HERV elements might influence tumorigenesis due to retroviral movement causing disruption and instability of the host genome [108].

The HERV-K family is the most recently acquired HERV subgroup by humans and so far, are the only known human endogenous proviruses to have retained open reading frames for all viral proteins [56,95]. HERV-K (HML-2) is the most studied subgroup, and the only one with a full length ORF [56]. Moyes *et al* reported that HERV-K113 and HERV-K115 are far more prevalent in Africa than in other regions with between 30-40% frequency, and that racial origin is an important factor for integration [88]. It was further suggested that the initial infection and incorporation of HERV-K113 likely occurred in Africa either during or after the migration of *Homo sapiens* north and eastward between 150 000 – 200 000 years ago [109–111].

1.2.3.4 Viruses Associated with Oesophageal Squamous Cell Carcinoma

As discussed above, multiple risk factors have been associated with oesophageal malignancies, including carcinogenic pathogens such as oncogenic viruses. Such viruses include HPV, EBV as well as Cytomegalovirus (CMV) and Herpes Simplex Virus (HSV) [112–114].

Human Papilloma Virus

HPV is a small, non-enveloped double-stranded circular DNA virus belonging to the *Papillomaviridae* family with more than 200 genotypes, classified into high risk (e.g. HPV-16 and 18), and low risk (e.g. HPV-6 and 11) based on their propensity for carcinogenesis [115–117]. Most commonly, HPV carcinogenesis has been characterised in cervical squamous cancer, but some research has also shown that persistent infection with high risk HPV can lead to head and neck cancers and oesophageal cancer [118] as the upper gastrointestinal tract may be exposed to HPV through oral transmission [119].

Some studies have suggested that the integration of HPV genomic DNA into the host genome could lead to the upregulation of viral oncogenes E6 and E7, facilitating the progression of oncogenesis [120]. This is brought about by mechanisms that disrupt the expression of the repressor E2 gene, and promote the abnormal expression of E6 and E7 [121]. HPV viral oncoprotein E6 targets p53 degradation and upregulation of telomerase expression leading to immortality of transformed cells, while oncoprotein E7 inhibits pRb suppressor protein leading to proteosome-dependent degradation [122].

In 1982, HPV was first suggested as having an association and contributing role in OSCC by Syrjanen *at al*, when morphological similarities were observed between HVP induced lesions in the genital tract and OSCC [123]. To date however, contention still exists over the exact role of HPV in the development of OSCC, if any, as no firm evidence has been established [124]. The frequency of HPV infection in oesophageal cancer is seen to vary greatly between 0%-88% [125,126]. However, this range may not accurately reflect the true situation as discrepancies in the findings are potentially due to variations in sampling methods, demographic and ethnic factors, anatomic sites and methods used for viral detection [21].

Many studies investigating the association of HPV with OSCC, have shown that the occurrence of HPV in OSCC fluctuates geographically, as HPV associated OSCC rates were frequently found to be twice as high in the common ‘high incidence OSCC belt’ across Asia and Eastern and Southern Africa when compared to western countries in Europe as

well as America [21,119]. Some studies showed the prevalence of HPV in OSCC to range from 11.7% to 39.9% in high OSCC incidence areas, where the most common genotypes detected were HPV16 (11.4%), HPV18 (2.9%), and HPV6 (2.1%) [21,124]. Demographic analysis showed that 20% of men and 18.4% of women were positive for HPV genotypes in oesophageal cancer [21].

During persistent infection with high risk HPV16, increased expression of viral oncoproteins E6 and E7 inactivate p53 and pRB tumour-suppressor proteins respectively, interfering with cell cycle control and causing genome instability as with other cancers [127,128]. Interestingly, it has been shown that in OSCC cell lines, high risk HPV18 oncoproteins do not promote cancer progression by the same mechanisms as seen in cervical cancer. Instead, oncoprotein E7 targets p130, a transcriptional regulator of cell cycle genes [129], while E6 deregulates miR-125b microRNA tumour suppressor [130] leading to activation of the Wnt/β-catenin signalling pathway and affecting cell migration, proliferation and apoptosis, and promoting the progression of cancer [131,132].

However, the exact role of HPV in OSCC remains unclear and a point of contention among researchers. Continued investigations with well-designed, case controlled studies using optimal detection methods are required.

Epstein-Barr Virus

EBV, also known as human herpes 4, is a dsDNA virus belonging to the *Herpesviridae* family affecting approximately 90% of the global population causing mostly asymptomatic infections [133–135]. Associated with 1.5% of all cancers globally, it is most commonly implicated in Burkitt's lymphoma, Hodgkin's disease, nasopharyngeal carcinoma, gastric carcinoma and leiomyosarcoma [136]. Similar to HPV, the HBV detection rate in oesophageal cancer has varied substantially ranging from 1.8% to 35.5% in several different studies [137–140].

The first reports of HBV detection in OSCC showed that 8.33% oesophageal tumour samples and 6.25% OSCC cell lines to contain a highly conserved EBV genome [141]. EBV DNA was then subsequently detected in 35-36% of samples in Taiwan [137] and later in Germany [142]. A most compelling study by Wu *et al* (2005) further showed a positive outcome where 164 oesophageal cancers (151 OSCC and 13 undifferentiated) showed clear associations with EBV [140].

However, these observations are a continual point of contention among researchers as a number of others have found no correlation between EBV with OSCC in studies from Russia [143], Malaysia [138], China [144], Iran [145], Greece [146] and Japan [147]. Therefore, whether EBV in fact is associated with OSCC remains unclear. These contradictory data could be due to a combination of geographical, racial and differences in detection techniques, and continued investigations with larger cohorts from different regions is required to illuminate these findings and uncover correlation relationship [114,148].

Cytomegalovirus and Herpes Simplex Virus

CMV and HSV both belong to the *Herpesviridae* family along with EBV. CMV is reported to possess the largest genome of all herpes viruses, and infections are commonly acquired from either sexual contact during adulthood or perinatally [149]. Multiple organs can be affected by CMV infection including the gastrointestinal tract and the oesophagus. CMV-associated oesophagitis was been described in oesophageal cancer patients, however, this association has not been thoroughly described [71,150].

HSV is categorised into two types, namely HSV-1 and HSV-2. HSV-1 is primarily transmitted through oral contact causing herpes lesions in the head and neck region, especially in the oral cavity (commonly referred to as cold sores). HSV-2 is a sexually transmitted infection more frequently associated with genital lesions and acting as a co-factor in the development and progression of cervical cancer [140,151]. Studies have reported a prevalence of 30% for HSV-associated oesophageal cancer in a high OSCC incidence region of China, where HSV-1 and HSV-2 DNA and protein were detected in 31.7% of well differentiated OSCC samples, proposing a plausible aetiological role of HSV in OSCC [151]. Furthermore, it has also been shown that mixed herpes viruses and HPV infection are highly associated (71.4%) with oesophageal cancer, although the mechanism of action for promotion of tumorigenesis remains to be clarified [71].

The aetiological role of oncogenic viruses in OSCC has remained contentious and debatable for many years as studies from a number of regions have reported high prevalence while other studies have found none. These varying observations may be dependent on geographical locations where countries in the high incidence areas appear to show higher prevalence of oncogenic viruses. Among the oncogenic viruses reported, HPV has the

strongest consensus of association while the implication of other viruses remain debatable [114].

1.3 Somatic Mutations in Cancer

Somatic mutations are a set of mutations acquired throughout the lifetime of an individual, distinguishable from germline mutations that are inherited from parents and transmitted to offspring [152]. They occur initially in healthy cells and most frequently do not cause alterations to cell behaviour [153]. However, occasionally key genes become altered in a manner that provides a competitive advantage to the mutated cell, promoting the formation of persistent mutant clones and initiating the process of tumour cell transformation [154,155]. For many decades it has widely been described that cancers are abnormal clones of cells characterised by and caused by abnormalities of the genome, further supported by the determination that agents that cause DNA damage and genetic mutations, also cause cancer [156]. Stratton *et al*. has suggested that two constituent processes form the basis of cancer development, being the continuous acquisition of heritable genetic variation and mutations in individual cells, and the natural selection that acts upon the resultant diverse phenotypes [152]. This natural selection frequently leads to the elimination of cells that have acquired deleterious mutations, yet in some instances may foster the progression of cells carrying alterations that confer the ability for more efficient and continued proliferation, and occasionally cells may develop the propensity for autonomous proliferation, invasion and metastasis [152].

Several distinct classes of DNA sequence changes are categorised as somatic mutations in the genome. These include single nucleotide polymorphisms (SNP's), insertions or deletions (indels) of small or large fragments of DNA, rearrangements where DNA is broken and re-joined to different segments, copy number variations (CNV's) leading to gene amplification or alternatively the complete absence of a DNA sequence from the genome [152]. Epigenetic changes can also lead to genomic mutations altering chromatin structure and gene expression, manifesting through changes in gene methylation states [152]. Mutagens originating both internally and externally to cells continually damage and affect cells, leading to DNA instability. In normal, healthy cells, this damage is routinely repaired, however, some mutations evade the DNA repair mechanisms and become fixed in the genome [152]. The rate of mutation increases with the exposure to exogenous factors such as carcinogens found in cigarette smoke, aflatoxins and ultra-violet radiation, which have all been implicated in a number of different cancers [157].

Within the distinct classes of DNA mutations mentioned above, each mutation can further be classified according to the tendency for carcinogenesis. Mutations that are found to confer a growth advantage within the cell are frequently referred to as ‘driver’ mutations, while those that do not confer this advantage but are present in the ancestor of the cancer cell, are termed ‘passenger’ mutations [152]. Driver mutations reside in genes typically known as ‘cancer genes’ and have been positively selected for during the evolutionary process of the cancer development [152]. It is likely that most cancer types carry multiple drivers, and on the basis of age-incidence statistics, it has been reported that breast, prostate, colorectal and other epithelial cancers require 5-7 rate-limiting events [158]. Further studies by Schinzel and Hahn supported this claim with their experimental research showing that in normal primary human cells, it is necessary for at least five to six critical changes to transform them into cancerous cells [159].

Recurring somatic mutations occurred in at least 350 out of the approximately 22,000 known protein-coding genes in the human genome with evidence that these contribute to the process of carcinogenesis [160]. Meanwhile, mouse model studies have shown that more than 2000 genes may have the potential to contribute to carcinogenesis if and when appropriately altered [161].

While mutation patterns differ between dominant and recessive cancer genes, approximately 90% of the known somatic mutations occur in dominantly acting genes where the mutation in only one allele is sufficient to contribute to tumorigenesis. The remaining 10% occur in recessively acting genes and require an alteration to be present in both alleles of the gene leading to the abolition of the function of the encoded protein [152]. In each dominantly acting cancer gene, missense mutations, in-frame indels, gene amplifications and genomic rearrangements are considered to be common mutational mechanisms involved in activation of these genes [152].

Among cancer genes, certain gene families, particularly the protein kinases, appear to be more prominently affected that others, with most cancer genes clustering on particular regulatory signalling pathways such as the classical MAPK-ERK pathway [162] where upstream mutations are detected in cell-membrane-bound receptor tyrosine kinases such as *EGFR, ERBB2, FGFR1, FGFR2, FGFR3, PDGFRA* and *PDGFRB*, as well as in downstream cytoplasmic components *NF1, PTPN11, HRAS, KRAS, NRAS* and *BRAF* [152].

1.3.1 Genetic Alterations in Oesophageal Squamous Cell Carcinoma

In recent years, the focus of OSCC research in both Western countries and regions of the high incidence OSCC corridor (Asia, through the east and southern Africa) have shifted their investigations from candidate gene studies to genome-wide association studies (GWAS) and whole exome sequencing (WES) with the purpose of identifying genomic variants associated with OSCC [163].

Multiple studies over the past decade have performed analyses of OSCC tumour-normal paired exomes and have collectively identified 22 driver mutations in genes commonly referred to as significantly mutated genes (SMG's) [9,21–23,29,164,165]. This significance is denoted by the functional relevance of these somatic variants[166]. These 22 SMG's are reported in Table 1.3.

Table 1.3: Significantly mutated genes detected in OSCC. Table adapted from Lin et al (2018)[166].

	Song *et al.* (n = 88)	**Lin *et al*. (n=139)**	**Gao *et* al. (n=113)**	**Sawada *et al*. (n=144)**	**TCGA (n=97)**	**Prevalence (%)**	**Additional Aberrations**	**Validation in OSCC cells**
TP53	1	1	1	1	1	80.5		Tumour-suppressor
NOTCH1	1	1	1	1	1	13.2		Tumour-suppressor
NFE2L2	1	1	1	1	1	9.3	Amplification	Oncogene
KMT2D		1	1	1	1	15.2		
CDKN2A	1	1	1	1		5.8	Deletion, hypermethylation	Tumour-suppressor
ZNF750		1		1	1	9.0	Deletion	Tumour-suppressor
PIK3CA	1	1		1		8.2	Amplification	Oncogene
RB1	1	1	1			4.8	Deletion	Tumour-suppressor
FAT1		1		1		8.8	Deletion	Tumour-suppressor
EP300		1		1		6.5		Tumour-suppressor
FBXW7			1	1		4.2	Deletion	Tumour-suppressor
TGFBR2				1	1	3.2		
AJUBA			1	1		2.8		Tumour-suppressor
CREBBP				1		5.5	Loss of heterozygosity	
FAT2		1				5.0		Tumour-suppressor
NOTCH3				1		4.8		Tumour-suppressor
PTCH1			1			4.5		
KDM6A		1				4.2	Deletion	
FAM135B	1					4.0		Oncogene
TET2				1		3.2		Tumour-suppressor
PTEN		1				2.2	Deletion	Tumour-suppressor
AADAM29	1					1.3		

* All genes have been reported in GWAS catalogue with detected mutations.

OSCC shares a number of these SMG's (such as *NOTCH1* and *ZNF750*) exclusively with squamous cell carcinomas of the lung [167] and head and neck [168], closely resembling these cancer types and quite distinctly different from OAC which is associated with GORD. This observation suggests that cells sharing similar origins acquire genetic aberrations of similar

oncogenic potential, with similar gene expression and cell differentiation profiles. Neoplasms emerging from cell linages with similar development patterns are more closely alike than those of different lineages located within the same tissue suggesting that a lineage-specific mutation process likely contributes to the different subtypes of oesophageal cancer [166].

The 22 SMG's consistently identified across the different study cohorts have also been linked to the development of multiple tumour types while functional studies have confirmed their biological effects in OSCC, confirming their 'driver mutation' status [166]. Many genes that have previously been linked to the development of oesophageal cancer are associated with DNA maintenance and repair, alcohol folate and carcinogen metabolism, cell cycle regulation and apoptosis [169–171]. As with most other cancer types, OSCC has mutations in the most frequently mutated driver gene, TP53 with other driver genes being less recurrent in OSCC [166]. Thus there is growing evidence supporting the notion that genomic alterations and epigenetic modifications contribute to the development and progression of carcinogenesis [172].

In Africa, the aetiology of OSCC still remains inadequately defined since there have been very few studies on OSCC in African populations. However, a systematic review of the genetic factors in the aetiology of OSCC identified 44 somatic changes in 22 genes in an African population. These included *AR*, *CCND1*, *CDKN2A*, *COL1A2*, *EFGR*, *EP300*, *FAT1*, *FAT2*, *FAT3*, *FAT4*, *FBXW7*, *JAG1*, *KMT2C (MLL3)*, *KMT2D (MLL2)*, *MUC2*, *NFE2L2*, *NOTCH1*, *NOTCH3*, *PIK3CA*, *SERPINB4*, *TP53*, and *TP63,* with 20 variants reported in South Africa, 3 variants in Kenya, and 21 variants in a Malawian population [163]. When comparing these genes with those shown in Table 1.4, it can be seen that there is significant overlap with most of the genes common to both.

The most common type of somatic variants that have been reported in OSCC are missense mutations that were detected in 14 out of the 22 genes (64%) [173,174]. Other somatic variants include insertions, deletions, frameshift mutations, loss of heterozygosity (LOH), and copy number gains all accounting for 14% each, while copy number losses accounted for 5% [175–177].

While a number of genes have clearly been highlighted as playing a role in OSCC in Africa and other high incidence regions, there still remains the need for further large scale investigations for reproducibility purposes. The identification of genetic variants and markers of OSCC susceptibility has very clear translational benefits especially to African

populations with regards to understanding and identifying disease risk and progression, and providing insights into potential targeted therapies and screening programs [163].

1.3.2 Mucins

Evidence in the literature suggests that mucins are also involved in the development of some cancers [178–184]. An extensive overview of mucins, and specifically *MUC3A* has been included in this review based on the later research findings reported in this thesis.

Mucosal surfaces commonly line body cavities in contact with the external environment. These include the gastrointestinal tract, the respiratory tract, reproductive tract and ocular surfaces, protecting them against pathogens and preventing dehydration [178,179,184]. Goblet cells of the vascular epithelium express a class of highly glycosylated, gel-forming proteins called mucins that form a viscous mucus layer lubricating mucosal surfaces, maintaining hydration and acting as a barrier against bacteria and other pathogens [178,184]. This complex viscoelastic mucus plays an important role in the immune response and is predominantly comprised of glycoproteins [185].

Contributing to the major structural component of this mucus is the mucin family of large mucins encoded for by the mucin genes. Mucins are heavily glycosylated proteins frequently categorised into two classes depending on their structure and localisation. Seven secreted mucins are categorised (*MUC2, MUC5AC, MUC5B, MUC6, MUC7, MUC8*, and *MUC19*), while eleven membrane bound mucins have been described, also referred to as transmembrane mucins (*MUC1, MUC3A, MUC3B, MUC4, MUC12, MUC13, MUC15, MUC16, MUC17, MUC20* and *MUC21*) [184,186,187]. Secreted mucins are reported to play a major role in barrier formation and surface protection against pathogens, while transmembrane mucins are retained within the plasma membrane and act as cell surface receptors involved in various signalling pathways affecting cell growth, differentiation, cell proliferation and apoptosis [178,185].

Mucin genes are translated into rod-shaped polypeptides which are then extensively glycosylated. The amino acid sequences of these glycoproteins are found to be considerably strewn with tandem repeat (TR) structures with a very high proportion of proline, threonine and serine residues forming what is known as the PTS domain characteristic of the mucin core. These TR's are highly polymorphic in length and sequence variability, are poorly conserved and are repeated multiple times [181,184]. TR's may be considered as the informational content of mucins and thus comprise what is frequently termed the

mucinome[188]. All mucins contain PTS rich repetitive domains where extensive *O*-linkage glycosylation occurs at the threonine and serine residues, forming major binding sites for many bacterial adhesins and thereby functioning to stimulate the immune response. The carbohydrate residues on the protein structure present in a 'bottle brush' formation around the protein core [178]. This structure further acts in an antimicrobial activity where bound pathogens are not killed but rather trapped and cleared through mucociliary transport in the viscoelastic mucin gel [189].

Mucins exert an extensive range of physiological functions including protection and lubrication of epithelial surfaces, maintenance of epithelial integrity, cell adhesion and protection against foreign particles, pathogens and toxins [178]. The extracellular protein backbone of the mucins is protected from proteolytic attack of host proteases by the extensive degree of glycosylation contributing to barrier function [184]. Different types of mucins present in various locations throughout the body perform specific functions, maintaining homeostasis and overall cell survival [178].

1.3.2.1 Transmembrane Mucins: Structure and Function

The transmembrane mucins are large glycoproteins of differing lengths, compositions and cytoplasmic signalling domains [184]. They are anchored to the apical surfaces of the plasma membrane of epithelial cells that come into contact with external environments. Structurally, transmembrane mucins are receptor-like and reminiscent of classical innate immune receptors. They are composed of a heavily glycosylated amino terminal extracellular region performing a barrier function, a single membrane-spanning domain anchoring the protein to the membrane, and a carboxy terminal intracellular cytoplasmic tail prone to phosphorylation [184]. The extracellular domain of these mucins is mainly composed of a varying number of TR's, a SEA domain (sea urchin sperm protein, enterokinase, and agrin), and an epidermal growth factor (EGF)-like domain [190,191].

Sea Domain

The SEA domain is a highly conserved domain of approximately 100 amino acid residues situated close to the luminal side of the membrane in the extracellular region of the glycoprotein. First identified in *MUC1* as playing a role in protein glycosylation, this domain is autoproteolytically cleaved in the endoplasmic reticulum [192], separating the mucin into two subunits: a large extracellular α-chain with numerous TR's, and a shorter β-chain containing

a portion of the extracellular domain, the SEA domain and EGF-like domain, the single transmembrane domain and the cytoplasmic tail. These subunits are non-covalently connected by parallel β-sheets [193]. *MUC1, MUC3, MUC12* and *MUC17* all contain the SEA domain with conserved G-S[V/I]VV cleavage motifs [194]. One of the functions of the SEA domain could be protection against mechanical force, alternatively it may be to shed the extracellular domain from the mucin structure to function as decoy receptors for pathogens [195]. Cells are able to sense the disruption of the SEA domain or release of the extracellular mucin domain, and activate signalling pathways emanating from the cytoplasmic tail [184,196].

EGF-like Domain

The extracellular domains of most transmembrane mucins contain EGF-like domains flanking the SEA domain (if present) that show homology to EGF and other related growth factors [178]. The function of this domain is to interact with EGF receptors to activate receptor signalling as has been reported for *MUC4* [197–199]. Through the release of the extracellular mucin domain, the EGF-like domain in both the α- and β-chain is free to interact with EGF receptors [200], and in this way the α-chain might have a biologically active role similar to that of cytokines, at a more distant location [181].

Cytoplasmic Tail

Cytoplasmic tails of transmembrane mucins vary greatly in length among different mucins. *MUC1* has a long cytoplasmic tail with numerous sites for phosphorylation of tyrosine residues [201]. Other large transmembrane mucins such *as MUC3, MUC12* and *MUC17* contain PDZ binding motifs (post synaptic density protein, Drosophilia disc large tumour suppressor, and zonula occludens-1 protein) in their far C-terminal region that are crucial for the trafficking and anchoring of receptor proteins, and for interacting with the cytoskeleton [202,203].

These intracellular cytoplasmic mucin tails are comparably similar to classical immune receptors in that they link to signalling pathways within the cell. MUC1 is the most extensively researched transmembrane mucin with its cytoplasmic tail associated with several intracellular signalling pathways [184]. Putative phosphorylation sites have been found on the cytoplasmic tails of all transmembrane mucins yet these tails are largely dissimilar in sequence and length and do not contain conserved domains among them. This suggests

that it is highly likely that a substantial degree of functional divergence is present among transmembrane mucins and signalling specificity between different mucins is expected [184].

Upon binding to bacteria or pathogens this extracellular region can be shed from the epithelial surface, triggering the phosphorylation of the intracellular cytoplasmic tail and activating signalling pathways that influence the inflammatory response, cell adhesion, differentiation and apoptosis [178,184,185]. Thus it is highly likely that transmembrane mucins act as signalling receptors able to activate specific signal transduction pathways involved in mucosal maintenance [184].

The transmembrane mucins may play a critical role in the maintenance of cellular mucosa, and the high degree of glycosylation of the extracellular domain provides a barrier function against invading bacteria, pathogens and contaminating agents. In addition, damage to this barrier in turn leads to shedding of the extracellular domain with subsequent signalling pathway activation linked to proliferation and apoptosis. Thus it is plausible that mucins are able to sense both pathogenic threat as well as epithelial damage [184].

1.3.2.2 Mucins in Cancer

To date, each specialised epithelium studied has been shown to display unique patterns of *MUC* gene expression that is frequently altered under numerous pathological conditions including diseases of chronic inflammation and many cancers [178,180]. As mucins commonly function as a protective barrier to pathogens entering the epithelium, it has been suggested that bacteria and circulating pathogens might affect mucin expression and production initiating a role for potential malignancies through the control of bacteria, regulation of inflammation and influence on signal transduction pathway [113].

Mucins commonly display altered expression patterns in human malignancies, both in type and quantity of mucin produced [181,182]. This is used as a disease diagnostic marker and has the potential to become a target for anti-cancer therapies [204]. This increased expression of mucins is favourable to cancer cells in a multitude of different ways through their characteristic patterns of glycosylation providing binding platforms for various growth factors and in turn promoting cell proliferation and metastasis through the downstream activation of pro-cancerous signalling pathways [183]. The chronic inflammation present in many cancerous states can alter mucin expression, for example, increasing *MUC2* expression in gastric cancers [205].

Glycosylation patterns and density of mucins differ quite substantially between healthy normal tissue and those expressed in tumour cells derived from the same tissue [206–208]. Commonly, the differentially glycosylated mucins expressed in tumours are used as markers for cancer. Transmembrane mucins contribute to carcinogenesis via their glycosylated extracellular domain that acts to protect cells against harmful conditions, as well as through their intracellular domains linking to pathways that regulate cell differentiation, apoptosis and inflammation [184]. Changes such as these may reflect the more direct role of mucins in tumour progression achieved for example, through their interactions with Trefoil family of receptors (TFF) which are known to have antiapoptotic functions [181]. Similarly, the EGF-like domains of membrane-bound mucins are capable of interacting with the EGFR family, including the HER2 receptor thereby affecting cell growth, proliferation and differentiation [181,199]. Shedding of the extracellular mucin domain releases and exposes the EGF-like motif to bind receptors, initiating the activation of signalling cascades like ERK1 and ERK2, enhancing proliferation [209]. Thus shedding may act as an activation signal leading to the phosphorylation and continuous activation of the intracellular cytoplasmic tail and subsequent activation of signalling pathways influencing inflammatory responses, epithelial adhesion, differentiation and apoptosis [184]. While this functional link between shedding of the extracellular domain and activation of the intracellular domain has not been conclusively established, it is possible that transmembrane mucins are able to activate intracellular signal transduction pathways by acting as signalling receptors for the external environment [184]. Cancer cells have been found to hijack and exploit this role for sustained pathway activation for survival, metastasis and tumour growth [178,184]. Mucin extracellular domains are detected in biological fluids such as serum and in the lumen of the intestinal tract. Excess shedding of these domains is frequently observed in cases of metastatic carcinomas through continuous activation of the cytoplasmic tails [210]. Furthermore, in cancer cells, the mucin cytoplasmic tail is frequently phosphorylated and can modulate immune responses such as IL-8 production and NF-κB pathways, as well as trafficking to the nucleus to influence transcription [184].

Membrane-bound mucins possess extremely long extracellular domains widely considered as sensors and receptors for the external environment. Because of this capacity to interact with pathogens and oncogenic receptors, they are frequently observed as playing a role in signal transduction and modulation of biological cell properties [211].

A number of studies have investigated the relationship between mucin expression in various cancers, and the biological behaviour of these neoplasms [212]. Differences in *MUC1* and *MUC4* expression have most commonly and most frequently been associated with

malignancy and have been identified as useful indicators of carcinogenesis [213–215]. *MUC1* has repeatedly been found to be upregulated in tumours with a poor outcome and significantly lower patient survival rates [212,216–219]. Alterations in the glycosylation patterns of mucins has been described in cancer where aberrant glycosylation results in truncated oligosaccharide side chains frequently observed in many adenocarcinomas [212].

Membrane-bound mucins, in particular *MUC4*, are often overexpressed in numerous cancers, promoting tumour development and progression [181,220]. This aberrant expression has been reported in pancreatic carcinomas, breast [209], cervical [221], head and neck [222] and gastrointestinal carcinomas [223], with some studies indicating that overexpression of *MUC4* in pancreatic cancer promotes tumour growth through the increase of cell proliferation and decrease of apoptosis [224–226]

The upregulation of *MUC1* expression in tumours may contribute an anti-adhesion function in cancer cells, inhibiting cell-cell aggregation and prompting the dissemination of cells from solid tumours and promoting tumour cell metastasis [227,228]. High levels of *MUC1* gene expression have been positively correlated with lymph node metastasis leading to poor patient prognosis [229]

The most common observations among mucin studies in cancer is the relationship between increased mucin expression and the inhibition of apoptosis in cancer cells. *MUC1* and *MUC4* have been the most intensively researched mucins across a number of different cancers, but even mucins such as *MUC2, MUC5, MUC13* and *MUC16* are associated with various neoplasms through their effect on cell cycle regulation and decreased apoptosis [230–233].

During pathogenic conditions, mucins that are restricted to the epithelial surfaces become exposed to circulation and their overexpression is frequently recognised as a potential biomarker for diagnosis of certain disease states [234]. Mucins are overexpressed, or aberrantly expressed in a number of different cancers, and this overexpression is commonly associated with poor patient prognosis, invasion and metastasis [181].

MUC1 and MUC4 are the most widely studied transmembrane mucins and Reynolds *et al*. (2019) have described the cell regulatory signalling pathways influenced by these mucins in cancer. Figure 1.2 adapted from Reynolds *et al*. (2019) shows an overview of the signalling pathways affected. MUC1 has been shown to activate JAK/STAT3 signalling in hepatocellular carcinoma and enhances the nuclear translocation of EGFR in cervical cancer. In colorectal cancer, MUC1 blocks targeting of c-Abl in the apoptotic response to DNA damage and in lung cancer, STAT3 has been shown to regulate the expression of

MUC1, driving cell survival and colony formation [235]. MUC4 is reported to inhibit apoptosis and enhance proliferation, migration and invasion through the PI3K signalling pathway and in breast cancer its overexpression was shown to promote apoptosis through the augmentation of ErbB2 signalling [235].

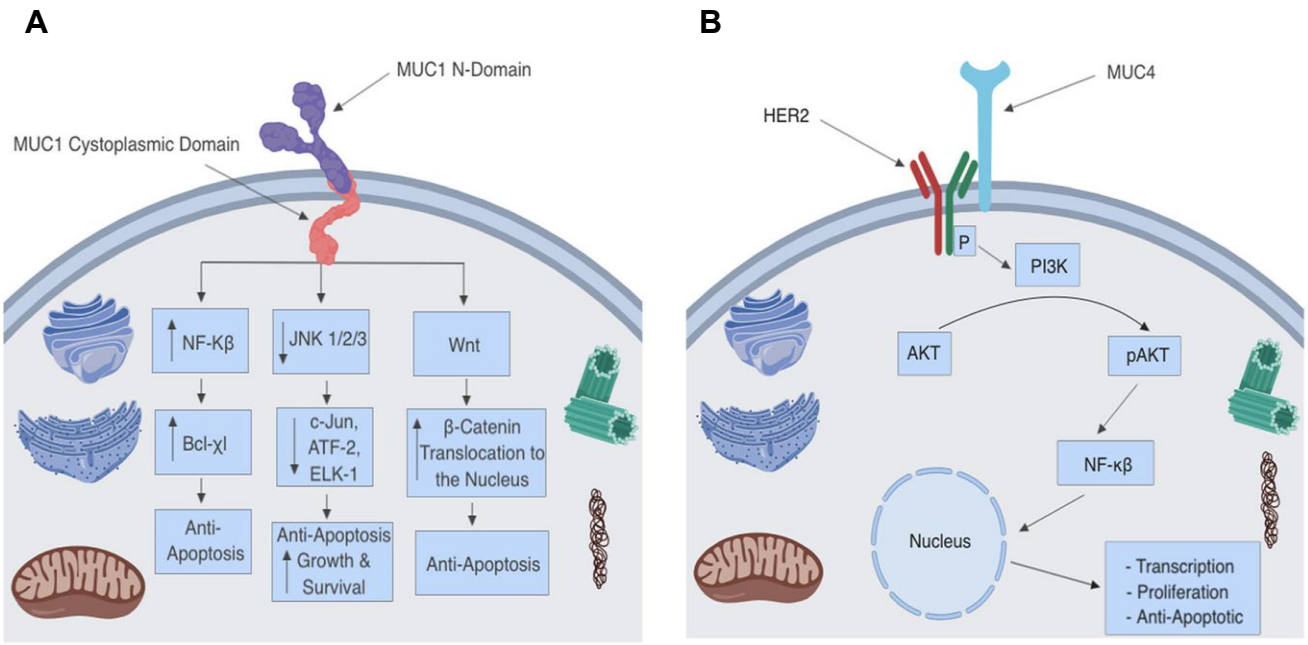


Figure 1.2: Signalling pathways affected by **A**) MUC1 and **B**) MUC4 transmembrane mucins in cancer. Figure adapted from Reynolds *et al*. (2019) [235]

Aside from their involvement in different cell regulatory signalling pathways, mucins are capable of altering cell-cell and cell-matrix interactions [236], promoting metastasis [237], conferring tumour cell resistance to therapeutics [238,239], and promoting immune evasion by tumour cells [240]. Apoptosis can be prevented through both the alteration of intracellular signalling as well as the formation of a physical barrier to limit chemotherapeutics reaching the cell [235].

In conclusion, many mucins have a role in the development of epithelium malignancies through their receptor function and regulation of inflammation, yet they also have the potential to initiate tumorigenesis through their influence on signalling pathways controlling many cell proliferation and regulating cascades [113].

1.3.2.3 Mucins in Oesophageal cancer

Like other epithelial cancers, mucins have also been implicated in the progression of cancers of the oesophagus. Mucin gene expression patterns in both normal and malignant oesophageal tissues have been investigated and *MUC1* and *MUC4* are commonly expressed in the normal oesophageal squamous epithelium and *MUC5B* is expressed in the submucosal glands [180]. In oesophageal malignancies, both MUC1 and *MUC4* expression are upregulated, and increases in *MUC4* expression have been correlated with all stages of squamous cell differentiation [180]. Well differentiated OAC was further reported to display the strongest *MUC* expression accompanying cell differentiation from metaplastic Barrett's oesophagus (specifically *MUC5AC* and *MUC6* in this case) [180].

In 1994, Fegelman *et al*. described a condition where OSCC primary tumours were found to contain a mucin-secreting component [241]. These tumours displayed both OAC and OSCC components and were believed to have arisen from the submucosal salivary glands of the oesophagus [242]. Sagara *et al*. (1999) demonstrated that *MUC1* expression is more frequent and intensive in OSCC, and that expression of MUC1 was significantly more prevalent in advanced stage tumours and is associated with poor patient survival [212]. Upregulation of *MUC1* in OSCC also acts as an effective tumour marker for early stages of tumour growth since it is not expressed in normal oesophageal tissue [212,243]. High expression of *MUC1* is often observed in advanced stages of OSCC, or with lymph node metastasis further suggesting the correlation between *MUC1* expression and tumour invasion and metastasis [212].

Over expression of *MUC1* in cultured oesophageal cells inhibits their aggregation, possibly due to the large, extended and rigid structure of the glycoprotein [227]. *MUC1* is weakly expressed by normal epithelial cells of the oesophagus, and its expression dramatically increases when cells become malignant [244,245], particularly in OSCC [246], and the upregulation of *MUC1* expression might play an essential role in the invasion and metastasis of OSCC cells, and was associated with a poor prognosis in patients [229].

An aberrant pattern of *MUC4* expression has frequently been reported [224], and this has also been shown to correlate with poor patient prognosis [247]. In normal oesophageal epithelium, the transmembrane MUC4 is expressed at very low levels. In comparison, *MUC4* expression is significantly increased in oesophageal cancers, progressively increasing from early metaplasia to invasive OAC [180,248]. Bruyere *et al*. further confirmed the link between upregulated MUC4 expression and OAC with their data indicating a major role for MUC4 in epithelial tumorigenesis [249]. Aberrant *MUC* gene expression has been closely correlated

with oesophageal cell differentiation, and most frequently, MUC1 and *MUC4* genes have been associated with squamous cell differentiation [180].

In an oesophageal meta-analysis, mucin expression was examined in Barrett's mucosa, OSCC, OAC as well as low grade dysplasia (LGD) and high grade dysplasia (HGD) [250]. *MUC2, MUC3, MUC5AC* and *MUC6* displayed a gradient of mucin expression progressing from oesophageal premalignant to malignant lesions. Increasing from lower levels in LGD and Barrett's mucosa, to higher levels in HGD with the highest expression detected in OAC [250]. It was suggested that the increased expression of these mucins in Barrett's epithelium could prove to be early events in the development of OAC [250].

Taking these observations into account, it is suggested that the increased expression of certain mucins in oesophageal cancer could serve as biomarkers for detection as well diagnostic indicators

1.3.2.4 MUC3A

MUC3A is a transmembrane glycoprotein encoded by the *MUC3A* gene in the mucin cluster of chromosome 7q22, expressed in various epithelial cells [251,252]. Structurally, this glycoprotein includes two tandem repeat domains of 375 and 17 amino acid residues respectively at the amino terminus of the extracellular domain, along with two cysteine-rich EGF-like domains flanking a SEA domain. A transmembrane domain follows with a 72 amino acid long cytoplasmic tail at the carboxy terminal (Figure 1.3) [252,253].The cytoplasmic tail contains a YVAL sequence similar to motifs recognised by SH-2 domain containing proteins [254]. This suggests that as with MUC1, MUC3A might be involved with intracellular signal transduction. Furthermore, the extracellular EGF-like domains have been suggested as playing a role in cell proliferation through protein-protein interactions and ligand binding [251,255]. Although the role of MUC3A in cancer development has not been fully elucidated, evidence has suggested that upregulated MUC3A expression is associated with poor prognosis and decreased overall survival in many tumour types [256–258].

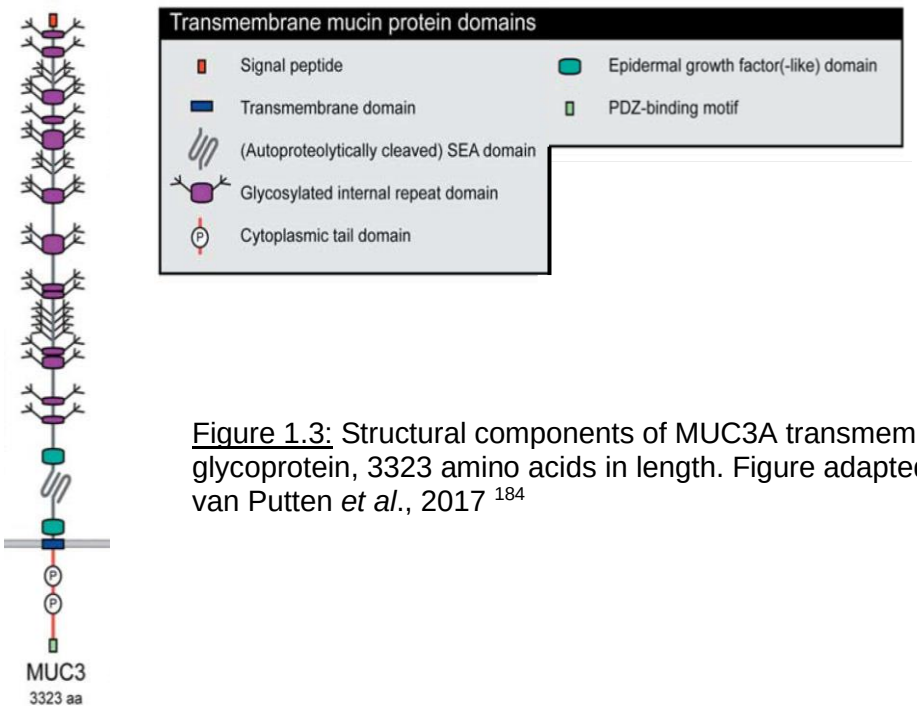


Figure 1.3: Structural components of MUC3A transmembrane glycoprotein, 3323 amino acids in length. Figure adapted from van Putten *et al*., 2017 [184]

Aberrant and upregulated expression of *MUC3A* has been implicated in a variety of different cancers including breast, pancreatic, gastric, colorectal, renal and prostate cancers [257–261]. The mechanism of *MUC3A* upregulation remains somewhat unclear, although some reports have indicated that a tetrameric-branched peptide through a conserved TFLK motif and an hypoxic tumour microenvironment were attributed to *MUC3A* expression [262,263]. Epigenetically, the expression of *MUC3A* is impacted by promoter hypomethylation [256], where *MUC3A* influenced cell migration and inhibited apoptosis through phosphorylation of tyrosine on the two EGF-like extracellular domains [264]. Similar to MUC1, the SEA domain of MUC3A is critical for autoproteolysis, impacting cell migration and invasion through the phosphorylation of HER/ErbB2 affecting the PI3K-Akt signalling pathway [260,265]. Additionally, MUC3A expression may also be regulated through PKC signalling pathways, thereby modulating the invasive and metastatic properties of multiple types of cancers [186]. MUC3A promotes tumorigenesis through the activation of the NF-κB pathway, interrupting DNA damage repair [266].

Patients with high MUC3A expression have a higher likelihood of developing lymph node metastasis and peripheral infiltration, correlated with late stage tumours and thereby corroborating the hypothesis that increased levels of MUC3A are linked to poor prognosis in cancer patients [258]. MUC3A expression may have an adverse independent prognostic

factor and may play an important role in oncogenesis, particularly in metastasis, invasion and the progression of cancer [267].

These observations suggests that *MUC3A* could act as a potential therapeutic target as inhibition of *MUC3A* may limit the rate of tumour evolvement and improve effectiveness of therapeutics already in practice [267]. Expression of MUC3A may also be seen as promising tumour biomarker, especially in cancerous tissues where expression in the normal tissue does not occur [266,268].

1.4 Bioinformatics Tools and Whole Genome Sequence Analysis

1.4.1 Bioinformatics in Cancer Research

Cancer is a progressive disease characterised by several genetic and epigenetic abnormalities and aberrations in gene expression [269–271]. These alterations of the genome differ between cancer types and include mutations, translocations, insertions, deletions and replications leading to the distortion of normal gene expression and function characteristic of cancer phenotypes [272,273]. Thus for any cancer-specific therapies to be developed and initiated, the exact molecular and genomic alterations must be identified in the cancer type of interest [274]. The application of bioinformatics approaches have widely been used to identify indicative phenotypic profiles of numerous cancer types [272]. For example, in 2007 Kim *et al*. was able to produce a list of differentially expressed genes in lung cancer through bioinformatics analysis of data [275]. An increasing body of evidence suggests that the interactions and networks between genes and proteins plays a pivotal role elucidating cancer mechanisms, thus it has become necessary to integrate systems biology, clinical science, omics technologies and computational science to improve research outcomes and understandings, essentially leading to improved diagnoses and targeted therapies [276]. Bioinformatics methods include high-throughput analyses comprised of mathematical and computational systems to facilitate the deciphering of oncogenic genomic data [277–279].

Designed with the intention to integrate data and results from multiple applications, bioinformatics platforms such as the open source development software Bioconductor [280], were developed specifically to address numerous biological questions at the intersection of genomics, genetics, proteomics and biomarker screening in cancer studies, while also providing comprehensive statistical analysis, data mining and visualisation tools [272]. With the development of technologies for large scale data generation and analysis, the approaches to cancer research are rapidly and continually changing [281]. The use of

genomics, transcriptomics and proteomics tools within bioinformatics pipelines has allowed for many new hypotheses to be tested and has prompted the rapid advancements in cancer research [282]. Large scale analyses has seen the proliferation of the number of genetic variants known to be associated with tumorigenic risk and cancer progression [272]. Mapping of amplicons associated with regions of high copy number is another common method used for searching for new oncogenes or for determining which known oncogenes contribute to a specific type of cancer [273]. In this way, bioinformatics approaches have enabled the identification of these oncogenic amplicons in several tumours [283].

With the completion of the sequencing of the human genome and the availability of pair-wise alignment programs such as the Basic Local Alignment Search Tool (Blast)[284], the identification of homologous sequences of known cancer genes has become increasingly relevant in cancer studies [273]. Implementing bioinformatics methods to exploit high throughput genomic data sources such as gene expression assays and mutational studies allows for an efficient and powerful approach for downstream analysis [273]. Target candidates can therefore be defined through gene clusters and desirable expression characteristics when comparing tumour and matched normal cells [273].

The most straightforward inference of differential gene expression can be deduced from a fold-change measurement where interesting gene clusters would exhibit a high fold-change ratio between tumour and normal cells. Although care should be taken to note that fold-change does not take into account measurement noise or sample heterogeneity, it has become commonplace to include as many tumour samples in an analysis as possible to allow statistical methods such as ANOVA and t-tests to be used on a gene-by-gene basis, assigning statistical significance where appropriate [273].

Bioinformatics tools are also widely used to identify, detect and validate cancer biomarkers associated with different stages of tumour development including initiation, progression and advanced stage diagnosis [272]. Alterations of biomarkers can be monitored and evaluated throughout the different cancer stages [276]. It has been suggested that genes with similar expression patterns are likely to share related functions [285]. Identification of new genes with similar expression profiles to tumour markers can be carried out by comparing expression patterns in a pair-wise fashion using appropriate distance metrics, such as Pearson's correlation coefficient and selecting the highest ranking observations [273].

As cancer is largely a genetic disease of altered gene expression patterns, genome instability and somatic mutations, it leaves a trail of genetic markers accompanying

tumorigenesis and cancer progression that can be used to discriminate cancer cells from normal ones. Such distinctions can aid in the identification and detection of bioinformatics strategies to develop strategies targeting cancer cells [272,273]. In this context, bioinformatics is an enabling and essential tool for the identification if biological markers and is continually evolving to enable the identification of influencers of cancer signalling pathways furthering our comprehension of the process of tumour development [272,273].

The applicability, specificity and integration of bioinformatics methodologies software and computational tools are continually being developed, refined and applied to exploit large datasets of genomic information [273,276]. The exploration of molecular mechanisms of cancer along with the identification and validation of novel cancer biomarkers is of growing importance, and the classical statistical and bioinformatics techniques for analysis form an indispensable 'backbone' for computational cancer research [286]. Furthermore, the identification of potential genes and proteins of interest in cancer signalling pathways is critical in the development of more effective cancer therapeutics and strategies [272].

1.4.2 Next Generation Sequencing (NGS) Technologies

WGS provides a comprehensive collection of genomic variants and is now becoming one of the most widely used sequencing applications, providing massive quantities of genome sequences in comparison to previous sequencing methods [287,288]. The most comprehensive assemblies of NGS are commonly derived when paired-end reads are used due to the very repetitive content of the human genome sequence [288]. Currently, most human genome sequencing efforts rely on a reference based assembly method where DNA fragments (reads) are mapped to the reference genome through the use of a multitude of specific alignment tools to build a consensus sequence similar to that of the reference [289–292]. The quality of the resulting genome assembly is of utmost importance for subsequent analyses of sequence variations, thus in addition to the specialised toolsets, additional parameters are necessary to evaluate quality, including Mate-pair information, unassembled reads, and read coverage [288].

The many NGS applications are immediately relevant in multiple fields, especially the medical field and cancer genomics. These methods frequently facilitate the discovery of genomic mutations and enhance the quantification of gene expression and discovery of regulatory RNA regularly classified in cancer [293]. When analysing WGS through bioinformatics pipelines, researchers can expect to discover molecular biomarkers and

genomic variants that can lead to the development of targeted therapies, acceleration of precision medicine and advancements in predicting, diagnosis and treating different diseases [287,294].

1.5 Introduction to the Project

This study analyses the Whole Genome Sequence (WGS) data obtained from 35 paired tumour-normal oesophageal squamous cell carcinoma patients in South Africa.

Despite the fact that several viruses and endogenous retroviruses have been shown to be present in the human genome, not much is known about their origins, their stability, their effect on normal gene function, or their role in disease initiation and progression. However, tentative evidence exists for a possible viral aetiology in oesophageal cancer. In this study, we investigated whether there exist any links between viral integration and cancer, initially making use of oesophageal cancer as a model for this investigation. We aimed to establish a catalogue of integrated non-human DNA sequences in the human genome, and to gain a better understanding of these viruses and their effects in tumorigenesis.

1.5.1 Aims and Objectives

The aims of the study were to:

1. Investigate the presence of novel viral insertions in the patient cohort
2. Investigate the locations and translocations of Human Endogenous Retroviruses, and to identify any links that might exist between their insertion and somatic mutations.
3. Investigate the presence of high impact somatic mutations in the cohort and identify possible genes of interest.
4. Use the analysis of RNA-sequence data to confirm the findings from (3) above.

These aims were achieved through bioinformatics analysis of WGS and RNA-sequence data combined with wet lab confirmation experiments.

While the novel viral insertion investigations were inconclusive, it was found that HERV's were not linked to somatic mutations in this patient cohort. In the investigations examining high impact somatic mutations, numerous mutations in the mucin gene, *MUC3A* were detected, however, these mutations are likely false positives. These findings were further

explored through RNA-sequence analysis for differential gene expression as well as functional enrichment analysis and immunohistochemistry. Thus the main focus of the study shifted heavily to explore the involvement of *MUC3A* in the OSCC samples investigated in this study, but there is still no definite consensus on these findings as they could not be confirmed in laboratory experiments.

Chapter 2: Laboratory Methods and Materials

2.1 MATERIALS

2.1.1 Sample Collection and DNA Sequencing

2.1.1.1 Sample Cohort

Patient recruitment has been ongoing since 2000 as part of a larger Oesophageal Squamous Cell Carcinoma (OSCC) research endeavour. For this doctoral study, patients presenting with histologically confirmed OSCC were recruited through informed consent at Groote Schuur Hospital in Cape Town and Charlotte Maxeke Hospital in Johannesburg, South Africa. Tumour biopsies with matched normal and blood samples were collected from patients by a registered nurse and samples were transported back to the laboratories at the University of Cape Town (UCT) and the University of the Witwatersrand (WITS) respectively, for processing and storage. Since there are no early symptoms, all patients presented with advanced stage 4 cancer, typically with lymph node metastases. No early stage cancers were present in the recruited patients.

The inclusion and exclusion criteria applied for patient recruitment were as follows:

Inclusion Criteria:

- All Patients with histologically confirmed OSSC
- All patients with histologically diagnosed oesophageal cancer

Exclusion Criteria:

- Patients who had received any form of radio or chemotherapy
- Patients who were too ill to go through the procedure

Once patient biopsies and blood samples had been processed, extracted DNA was subjected to Whole Genome Sequencing (WGS). Prior to the onset of this particular study, WGS was carried out on five tumour-normal pairs from Groote Schuur Hospital, at the New York Genome Centre in New York, United States of America. Of these five pairs, two were found to be cross contaminated but three were used in an exploratory pilot bioinformatics analysis for the investigation of viral integration using the facilities and expertise at the Centre for Proteomic and Genomic Research (CPGR) in Cape Town, South Africa.

The main sample cohort for this study comprised of a further thirty-five tumour-normal pairs which were subjected to WGS at the Wellcome Sanger Institute in Cambridge in the United Kingdom, a major global centre for high throughput sequencing. The total patient cohort of thirty-five patients comprised of twenty females and fifteen males with a mean patient age of 62 years for females and 54 years for males (Table 2.1). Twenty-one of the thirty-five patients were recruited at UCT/Groote Schuur, with the remaining 14 recruited from WITS/Charlotte Maxeke hospital. A breakdown of the individual patients' age, gender, % tumour cells, sequencing coverage and sequencing duplication factor is shown in Table 2.2. The entire patient cohort was made up of South African patients seeking medical care at the two government hospitals mentioned above.

Table 2.1: Number of patients and mean age per gender of the main 35-patient cohort subjected to WGS by the Wellcome Sanger Institute

	Number of Patients	Mean Age
Male	15	54
Female	20	62

Table 2.2: Patient cohort age and gender, where F represents females and M represents males. Pilot UCT patients were sequenced at the New York Genome Centre, while WGS of the remaining patients making up the main patient cohort of the study, were sequenced at the Wellcome Sanger Institute, UK. Histological testing was performed on these main thirty-five patient DNA samples and the % tumour cells in the biopsy is indicated. WGS coverage for each tumour and normal sample is shown, as well as the duplication factor for samples in the main patient cohort. Duplication factor represents the fraction of mapped reads where any two reads share the same 5' and 3' co-ordinates.

Pilot UCT Patients

Patient Number	Age	Sex	Coverage T/N
OC547	30	F	80x/40x
OC569	79	F	80x/40x
OC607	48	F	80x/40x

UCT Patients

Patient Number	Age	Sex	% Tumour Cells	WGS Coverage T/N	Duplication Factor T/N
PD39445	57	F	n.d.	42.64/55.65	0.09/0.08
PD39446	45	M	n.d.	42.42/51.14	0.09/0.07
PD39447	41	M	28	50.64/49.49	0.14/0.07
PD39448	52	M	44	46.73/48.17	0.14/0.07
PD39449	79	F	57	47.33/50.99	0.13/0.07
PD39450	50	F	64	51.99/46.1	0.14/0.07
PD39451	71	M	47	55.27/47.23	0.14/0.11
PD39452	53	F	64	53.17/49.06	0.14/0.11

WITS Patients

Patient Number	Age	Sex	% Tumour Cells	WGS Coverage T/N	Duplication Factor T/N
PD44691	70	F	39	33.77/41.83	0.06/0.08
PD44692	54	M	47	31.48/39.89	0.06/0.08
PD44693	59	M	54	42.6/38.02	0.07/0.08
PD44694	54	F	21	40.01/38.1	0.07/0.08
PD44695	63	F	64	41.5/42.37	0.07/0.09
PD44696	54	F	69	34.86/39.44	0.06/0.08
PD44697	38	M	29	36.02/37.06	0.07/0.07
PD44698	45	F	70	38.59/37.61	0.07/0.08

PD39453	37	M	22	51.67/43.32	0.16/0.11	PD44699	81	F	61	34.13/42.24	0.07/0.09
PD39454	67	F	n.d.	51.68/51.82	0.16/0.11	PD44700	71	F	43	34.85/35.76	0.06/0.07
PD39455	48	F	91	47.11/53.46	0.16/0.12	PD44701	69	F	13	35.49/36.71	0.06/0.07
PD39456	41	M	51	50.71/45.18	0.17/0.11	PD44702	65	F	46	36.33/40.62	0.06/0.07
PD39457	57	M	62	48.83/45.3	0.16/0.08	PD44703	78	M	38	37.12/36.23	0.06/0.07
PD39458	60	F	30	48.58/44.17	0.17/0.08	PD44704	56	M	30	34.21/39.34	0.06/0.07
PD39459	64	F	22	62.77/51.09	0.12/0.09						
PD39460	56	M	66	48.05/47.45	0.10/0.09						
PD50649	66	F	55	34.13/31.14	0.10/0.09						
PD50650	60	F	22	37.02/37.83	0.09/0.09						
PD50651	70	M	56	34.68/40.05	0.09/0.09						
PD50653	57	F	48	29.09/33.14	0.09/0.09						
PD51372	60	M	27	36.73/32.16	0.09/0.08	**n.d. = not determined*					

2.1.1.2 Ethics and Consent

Ethical approval for the project was obtained from the UCT/Groote Schuur Hospital Human Research Ethics Committee (Ethics number: 040/2005), and any publications derived from this study will not divulge the identity of the participants. Only a limited set of researchers have access to patient information which is kept confidential and password protected.

2.1.2 Cell lines

Seven OSCC cell lines were used in this study. WHCO 1, WHCO5 and WHCO 6 cells lines were originally established in South Africa from surgical biopsies of primary OSCC [295], while the KYSE 30, Kyse 150, KYSE 180 and KYSE 450 cell lines were Japanese derived [296]. The non-cancerous oesophageal cell line EPC2 a telomerase-immortalised epithelial cell line [297] was used as a control cell line.

2.1.3 PCR Primers

All primers used in Polymerase Chain Reactions (PCR) were purchased with standard desalting from Whitehead Scientific (Pty) Ltd (WhiteSci, Cape Town, South Africa). Primers were prepared as 100 µM stocks and frozen at -20ºC. 10 µM Working solutions of these stocks were used in PCR experiments.

2.1.4 PCR Reagents

PCR reaction reagents were purchased from Roche (Mannheim, Germany) (Table 2.3) and used as per the manufacturer's instructions.

Table 2.3: Reagents used in PCR reactions

Reagent	Supplier	Product Code
PCR Buffer (Excl. Mg^{2+})	Roche	11647687001
$MgCl_2$	Roche	11600770001
dNTP	Roche	11581295001
Taq DNA Polymerase	Roche	11146165001

2.2 METHODS

2.2.1 Extraction of Genomic DNA and RNA

2.2.1.1 Processing of Patient Blood

Blood were 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 clearly labelled 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 clearly labelled 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.

2.2.1.2 Extraction of Genomic DNA from Blood

Tubes containing buffy coat were removed from -80°C storage and thawed at room temperature. Once thawed, the buffy coat was transferred to a sterile 50 ml centrifuge tube, diluted with 2 volumes of 1x PBS and mixed by inverting the tubes, followed by centrifugation at 3000xg for 15 minutes. The resulting pellet was resuspended 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 at 4°C. The supernatant was removed and the pellet was resuspended in 3 ml of T20E5 (20 mM Tris-HCl, 5 mM EDTA), 200 µl of 10% SDS and 75 µl of 10 mg/ml Proteinase K (Roche, 03 115 852 001, Indianapolis, USA), and inverted to mix well and then incubated overnight at 45°C.

The following day, 1 ml of saturated NaCl was added followed by vigorous mixing for 15 seconds and centrifugation at 3200xg for 40 minutes at room temperature. The supernatant

containing the DNA was carefully transferred to a clean tube and 2 volumes of absolute ethanol was added to precipitate the DNA. The tube was agitated gently and centrifuged for 30 minutes at 3000xg at 4°C to precipitate the DNA which was then transferred to a clean microcentrifuge tube. The pellet was then washed with 1 ml of 70% ice-cold ethanol and centrifuged at 7000xg for 5 minutes at 4°C. This step was repeated followed by the removal of the supernatant. The DNA pellet was air dried at room temperature and then dissolved in 400 µl Tris-EDTA buffer 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. The DNA yield was determined using the NanoDrop 2000 spectrophotometer (ThermoFisher Scientific, 2000/2000c, USA), at a wavelength of 260nm. DNA was stored at -20°C until use.

2.2.1.3 Extraction of Genomic DNA 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 2.2.1.4). 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 with 350 µl wash buffer 1 (Buffer AW1) 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 after 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 NanoDrop 2000 spectrophotometer, and samples were stored at -20°C until further use.

2.2.1.4 Extraction of RNA 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 precipitated that may have formed, was then transferred to an 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 1 stock solution was added to 70 µl of digestion buffer (Buffer RDD) and mixed gently by inverting the tube. 80 of this 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-centrifugation 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 minutes 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 260nm using a NanoDrop 2000 spectrophotometer, and samples were stored at -20°C until use.

2.2.1.5 Determination of DNA Integrity

100 ng of DNA was electrophoresed on a 1% agarose gel (SeaKem®, Lonza, 50002, Rockland, ME, USA) together with 1 µl Novel Juice (Bio-Helix, LD001-1000, Taipei, Taiwan) detection dye. A suitable gene-ladder was loaded into the gel (GeneRuler ™ 100bp Plus

DNA Ladder (ThermoFisher, SM0321, Vilnius, Lithuania)). Gels were immersed in TBE buffer (45 mM Tris-borate, 1 mM EDTA) in a gel-electrophoresis system (ADVANCE Mupid®-One 077388, Tokyo, Japan) and run for 35 minutes at 100V. Gels were then examined under ultra violet light using a gel documentation system (Biospectrum ™ 500 Imaging System UVP, Cambridge, UK) in order to visualise the DNA. This standard protocol is described in Lee et al, 2012 [298], with the amendment that Novel Juice fluorescent reagent was added to the samples to provide an environmentally safe, non-hazardous alternative to ethidium bromide for DNA detection.

2.2.2 Whole Genome Sequencing

2.2.2.1 Whole Genome Sequencing at New York Genome Centre:

DNA isolated from five normal and tumour paired biopsies were subjected to WGS at the New York Genome Centre. In the selection criteria, only tumour biopsies containing more than 50% tumour tissue, and normal biopsies with 0% tumour tissue were subjected to sequencing and analysis. Whole genomes of the isolated RNase treated DNA were sequenced on an Illumina HiSeqX using 150 base pair (bp) paired end reads to a depth of 40x and 80x coverage for normal and tumour tissue respectively. Reads were aligned to the NCBI genome build 37 using Burrows-Wheeler Aligner (BWA) [299] and duplicates were removed using Picard [300]. Base quality score was recalibrated and local realignment around indels were performed using GATK [301]. More than 75% of bases were sequenced with a quality score above Q30 (1 in 1000 probability of incorrect base call with 99.9% inferred base call accuracy). Downstream bioinformatics analysis of these sequenced genomes is described in Chapter 3.

2.2.2.2 Wellcome Sanger Institute Data

DNA isolated from thirty-five paired blood samples and tumour biopsies, were subjected to WGS at the Wellcome Sanger Institute in Cambridge, UK. Samples were genotyped for single nucleotide polymorphisms (SNP) using a Fluidigm chip array to confirm the tumour and normal samples matched. Samples were then sequenced on an Illumina HiseqX10 using 150 bp paired end reads to a depth >30x coverage. Paired-end reads were aligned with Burrows-Wheeler Aligner (BWA) [299] and polymerase chain reaction (PCR) duplicates were marked using Picard [300]. The Caveman and Pindel algorithms (github.com/cancerit/CaVEMan, https://github.com/genome/pindel) [302,303] were used to call

SNPs and insertions and deletions (indels) in the whole genome data. Copy number analysis was performed using ASCAT [304]. Downstream bioinformatics analyses of these sequenced genomes are described in Chapters 4 and 5.

2.2.3 RNA Sequencing

RNA sequencing (RNA-seq) was also performed at the Wellcome Sanger Institute on 15 of the WGS patients' RNA. Table 2.4 indicates the RNA sample numbers alongside their matched WGS DNA sample numbers. Eight of the patients were from Groote Schuur/UCT and seven were Charlotte Maxeke/WITS patients. RNA libraries were generated using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (New England Biolabs, E7760, Ipswich, MA, USA) and in-house adaptors were ligated to 100-300 bp fragments. This kit incorporates dUTP during second strand cDNA synthesis for strand-specificity and is then subsequently excised with the USER enzyme. All samples were subjected to 10 PCR cycles using the 'sanger_168 tag' set of primers using Kapa HIFI HotStart ReadyMix PCR kit (Roche, KR0370 – v10.19, UK) and paired-end sequencing was performed on Illumina's HiSeq 2500 with 75 bp read length. GRCh37 reference genome was used for the alignment steps. Downstream analysis of this RNA-seq data is described in Chapter 6.

Table 2.4: RNA patient numbers and their matched WGS DNA patient numbers.

WITS Patients		**UCT Patients**	
RNA Seq	**Matched DNA**	**RNA Seq**	**Matched DNA**
PR44697	PD44697	PR50548	PD50649
PF44699	PD44699	PR50549	PD50650
PD44701	PD44701	PR50550	PD50651
PR44702	PD44702	PR50551	*
PR44703	PD44703	PR50552	PD50653
PR44704	PD44704	PR50553	PD51372
PR44705	*	PR50554	*
		PR50555	*

*DNA not available for these samples

2.2.4 Cell Culture

2.2.4.1 Thawing of Frozen Cells

Cryotubes of frozen cells were removed from liquid nitrogen storage and thawed in a water bath at 37°C. 70% ethanol was used to sterilise the vials, and the contents was then transferred to sterile 12 ml tubes containing 5ml of complete culture media. Complete culture media contains high glucose (4.5 g/L) Dulbecco's Modified Eagle Medium (DMEM) (Gibco, 12800-017, Paisley, UK) supplemented with 10% gamma irradiated foetal bovine serum (FBS) (Gibco, 10493-106, Paisley, UK), 100 U/ml penicillin and 100 µg/ml streptomycin (Biochrom AG, A321-44 and A331-27, Berlin, Germany)

Cell suspensions were centrifuged at 3220xg for 4 minutes at 4°C using an Orto Alresa Consul 21 R centrifuge (Orto Alresa, CE114, Madrid, Spain). The media was removed and the cell pellets were resuspended in 10 ml of DMEM media and re-seeded onto fresh 10ml tissue culture dishes.

2.2.4.2 Culturing and Passaging

The OSCC cell lines were maintained in a proliferative state through culture in complete media, at 37°C in a humidified atmosphere containing 5% CO_2.

The non-cancerous EPC-2 cells were maintained in Keratinocyte Serum-Free Media (KSFM) (Gibco, 17005-034, Paisley, UK) supplemented with 50 µg/ml Bovine Pituitary Extract (BPE) (Gibco, 13028-014, Aukland, New Zealand), 1 ng/ml Epidermal Growth Factor (EGF) (Gibco, 10450-013, New York, USA) and 1% penicillin-streptomycin [297] at 37°C in a humidified atmosphere containing 5% CO_2.

Culture media was refreshed every two to three days using pre-warmed phosphate buffered saline (PBS) for all washing steps. Cells were passaged to ~80% confluence before trypsinisation at 37°C for 4 minutes in 0.25% trypsin-EDTA (BD Difco™ 250, 215240, San Jose, CA, USA) to lift cells from the culture plates. The trypsin was quenched by the addition of complete culture media, centrifuged at 3220xg for 4 minutes at 4°C, resuspended in complete medium and re-seeded onto fresh culture plates. Cells were split in a 1:3 ratio and maintained in this manner until enough plates and cells were obtained for the purpose of DNA and RNA extraction.

2.2.4.3 Freezing Cells for Storage

Cells were washed with pre-warmed PBS followed by trypsinisation and centrifugation as was described above. Following centrifugation, each cell pellet was resuspended in 1.8 ml DMEM complete medium, and 1.8 ml of freeze medium (20% DMSO, in complete medium) was added to obtain a final DMSO concentration of 10%. Cell suspensions were divided into two sterile cryotubes and stored at -80°C for 24-48 hours before they were transferred to liquid nitrogen storage.

2.2.4.4 Extraction of Genomic DNA from Cell Cultures

Extraction of genomic DNA from cell lines was performed as described by Strauss (1998) [305]. Once cultures reached ~80% confluence, cells were harvested by trypsinisation as previously described, quenched and centrifuged at 423xg for 4 minutes at 4°C. The cell pellets were resuspended and washed in 10 ml of ice-cold PBS and re-centrifuged. The supernatant was discarded and the pellets were then resuspended in 600 µl of digestion buffer (100 mM NaCl, 10 mM Tris pH8, 25 mM Na_2EDTA 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 1.5ml microcentrifuge tubes. The tubes were vortexed to mix the contents thoroughly and centrifuged at 1700xg 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 1700xg for 10 minutes. The pellet was washed with 1 ml of 70% ethanol, vortexed gently and centrifuged at 1700xg for 5 minutes. The supernatant was removed, the DNA pellets air-dried for 60 minutes, and resuspended in 50 µl TE-buffer. The DNA yield was measured using the NanoDrop 2000 spectrophotometer at a wavelength of 260nm. DNA was stored at -20°C until use.

2.2.4.5 Extraction of RNA from Cell Cultures

Following cell propagation and culture as described above, RNA isolation was performed once cell cultures reached ~80% confluence. Plates were washed with 10 ml ice-cold PBS and cells were lysed with 1ml of TRIzol reagent (Zymo Research, R2050-1-200, Irvine, CA, USA) added to each plate. Using a combination of pipetting and cell scraping, cell

suspensions were transferred to clean, sterile 1.5 ml microcentrifuge tubes and incubated at room temperature for 5 minutes to allow for the complete dissociation of the nucleoprotein complexes. 200 µl Chloroform was added per 1 ml TRIzol Reagent, the tubes were shaken vigorously by hand for 15 seconds and incubated for a further 3 minutes at room temperature. Samples were then centrifuged at 20 000xg for 15 minutes at 4°C. The clear, upper aqueous layer was transferred to clean 1.5 ml microcentrifuge tubes and the RNA was precipitated by the addition of 500 µl isopropanol followed by shaking thoroughly for 15 seconds. Samples were incubated for 10 minutes at room temperature followed by centrifugation at 20 000xg for 20 minutes at 4°C to pellet the RNA.

Following this centrifugation step, the supernatant was carefully removed from the RNA pellet which was then washed with 1 ml of 75% DEPC-treated-ethanol, mixed by vortexing and re-centrifuged at 15 000x for 5 minutes at 4°C. The RNA pellet was air dried for 10 minutes and dissolved in 50 µl DEPC-treated-H_2O at 55°C for 5 minutes. RNA yield was measured using a NanoDrop 2000 spectrophotometer at a wavelength of 260nm, and samples were stored at -80°C until use.

2.2.5 Polymerase Chain Reaction (PCR)

2.2.5.1 PCR Primers

All PCR primers were designed using standard PCR design protocols [306] and purchased with standard desalting from Whitehead Scientific (Pty) Ltd (WhiteSci, Cape Town, South Africa). Primers were prepared as 100 µM stocks and frozen at -20ºC. 10 µM Working solutions of these stocks were used in PCR experiments.

All primers used in this study are detailed below (Tables 2.5-2.8).

Table 2.5: Forward and Reverse primers for long (611 bp) and short (170 bp) *Autographa Californica* Nucleopolyhedrovirus sequences. The same forward primer was used for each set while the reverse primers differed, providing a long and a short anticipated product length. Primer sequences are provided in 5' to 3' direction.

Primer	5' – 3'	Primer Length	T_A °C	GC%	Product Length
Forward Primer	CAAACGCAACAAGAACATTTGTAG	24 bp	53.5	37.5	
Reverse primer 1 (Long)	GTTTGTGCGTCTCATTACAATGGC	24 bp	57.1	45.8	611 bp
Reverse primer 2 (Short)	GAAGGAATACGAAGAAGAAGACG	23 bp	53.1	43	170 bp

Table 2.6: New *Autographa Californica* Nucleopolyhedrovirus primers designed for some individual insertion fragments detected. Primer sequences are shown in a 5' to 3' direction.

Primer	5' – 3'	Primer Length	T_A °C	GC%	Product Length
Forward Primer 1	GTTTAACGGTTCGTCCAAC	19 bp	51	47.4	134 bp
Reverse Primer 1	CAATGAATTTGGGATCGTCGG	21 bp	54,2	47.6	
Forward Primer 2	CAACAAACAACTGCTCGCAG	20 bp	55	50	142 bp
Reverse Primer 2	GTACGCAGCTTCTTCTAGTTC	21 bp	53	47.6	
Forward Primer 3	CGCGTAGTTATAATCGCTGAGG	22 bp	55.3	50	118 bp
Reverse Primer 3	CGAAGAAGAAGACGACAAGGCG	22 bp	54.5	54.4	

Table 2.7: HERV-K113 forward and reverse primers sourced from literature [307,308]. Primer sequences are shown in a 5' to 3' direction.

Primer	5' - 3'	Primer Length	T_A °C	GC%	Product Length
Forward Primer	GCATGGGGAGATTCAGAACC	20 bp	55.7	55	303 bp
Reverse primer	CATGTTTCCTAGTCAACTTAGC	22 bp	56.4	55	

Table 2.8: Primer sets designed for each of the five clusters where mutations in the *MUC3A* gene were located. Primer sequences are shown in a 5' to 3' direction.

Primer	5' – 3'	Primer Length	T_A °C	GC%	Product Length
Cluster 1					
Forward Primer	TAAGTACACTCAGCACTCCTA	21 bp	52.1	42.9	889 bp
Reverse Primer	GAGATCATGGATGTAGAAGTTACC	24 bp	52.6	41.7	
Cluster 2					
Forward Primer	TTTCTACAGTATCTCTCACAACAGC	25 bp	54.4	40	631 bp
Reverse Primer	TAGGTGATAGTTGTTGGTGTTGTG	24 bp	54.9	41.7	
Cluster 3					
Forward Primer	CTACTTCTCTTACTAGTGCTCTC	23 bp	51.5	43.5	1292 bp
Reverse Primer	GCTGGGAGTATCATGTGA	18 bp	51.2	50	
Midway Forward Primer	TCTTTGATCACCACAACCAC	20 bp	52.9	45	837 bp
Reverse Primer	AAGTGAAGCTGGGAGTACTGT	21 bp	55.7	47.6	810 bp
Cluster 4					
Forward Primer	TCTTTGATCACCACAACCAC	20 bp	52.9	45	1424 bp
Reverse Primer	GTCTCTGAGGTAGTAAAATGTGAGGTGATG	30 bp	58.2	43.3	

Midway Forward Primer	ACTGAGAACGCCACACAC	18 bp	55.3	55.6	1004 bp
Reverse Primer	GCTGGGAGTATCATGTGA	18 bp	51.2	50	837 bp
Cluster 5					
Forward Primer	ACCTCACATGATACTCCC	18 bp	50.6	50	752 bp
Reverse Primer	ATATCAGTGGGTATAGAGGGAAAG	24 bp	53.3	41.7	

2.2.5.2 PCR Methodology

The primers were optimised using cell line or normal blood DNA. PCR reactions were prepared under sterile conditions on ice in 25 µl volumes as per Table 2.9. Thermocycling was carried out using an Applied Biosystems SimpliAmp Thermo cycler machine (Applied Biosystems by Thermo Fisher Scientific, A24B12, Singapore). A standard thermocycling protocol was applied at the start of each primer optimisation as described by Canene-Adams (2013) [306] (Table 2.10). Annealing temperatures were set according to the manufacturer specifications for each primer set, and adjusted to obtain the most optimal conditions for each primer set. 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 in the pursuit of primer optimisation. Optimal primer conditions for each primer set are described in the relevant chapters.

Table 2.9: Standard PCR reaction mix components per single reaction. Reactions were prepared to a final volume of 25 µl.

Reagent	**Stock Concentration**	**Volume Added (µl)**
Distilled H_2O	-	15.8
PCR Buffer	10x	2.5
$MgCl_2$	25mM	2
dNTP	10mM	0.5
Forward Primer	10µM	1.5
Reverse Primer	10µM	1.5
Taq DNA Polymerase	100U	0.2
DNA Template	100ng/ul	1

Table 2.10: Standard PCR thermocycling conditions used as a starting point for all PCRs. [306].

Cycle	Conditions		
Initial Denaturation	94°C:	1 minute	
Denaturation	94°C:	1 minute	25-35 cycles
Annealing	55°C:	1 minute	
Elongation	72°C:	1 minute (per kilobase PCR product)	
Final Extension	72°C:	4 minutes	
Cooling	4°C:	Hold	

2.2.5.3 Post-PCR Visualisation of Product Amplification

After successful amplification, 1-2 µl of the PCR products were electrophoresed through a 1% agarose gel for 35 minutes at 100V. Novel Juice was added to the samples for the purpose of visualisation of the amplified DNA. The GeneRuler™ 100bp Plus DNA Ladder was loaded into the gel as markers for amplicon size identification. Gels were visualised on the Biospectrum™ imaging system under UV light to confirm the quality of the PCR as well as to identify the specific product size.

2.2.5.4 Purification of amplified DNA from Agarose Gel using a Microcentrifuge

DNA purification from agarose gels was performed using the Qiagen QIAquick Gel Extraction Kit (Qiagen, 28704, Hilden, Germany). All centrifugation steps were performed at 17900 xg. Following PCR and visualisation of amplification on a 1% agarose gel under UV light, target bands (DNA fragments) were excised from the gel using a sterile surgical blade and placed into sterile microcentrifuge tubes. Gel slices were weighed and 3 volumes of solubilisation and binding buffer (Buffer QG) was added to 1 volume of gel (i.e. 300 µl buffer adder per 100 mg gel). Tubes were incubated at 50°C for 10 minutes to completely dissolve the agarose, and were vortexed every 2-3 minutes during this incubation to aid in the process. Following incubation, the colour of the tube contents was evaluated, a yellow colour indicating the correct pH (≤ 7.5). 1 gel volume of isopropanol (i.e. 100 µl added for 100 mg of gel) was added to the sample to increase the yield of DNA and mixed by inverting the tubes. The sample was applied to a QIAquick spin column in a 2 ml collection tube (provided) and centrifuged for 1 minute to bind the DNA. The flow through was discarded and the QIAquick column placed back in the same collection tube. 0.5 ml of solubilisation and binding buffer (Buffer QG) was added to the column followed by centrifugation for 1 minute to remove all traces of agarose. 0.75 ml of wash buffer (Buffer PE) was added to the column

and centrifuged for 1 minute. The flow through was discarded and the column was centrifuged for an additional 1 minute. The QIAquick column was placed into a clean 1.5 ml microcentrifuge tube and 30-50 µl of elution buffer (Buffer EB) was added to the centre of the column membrane and incubated at room temperature for 1-4 minutes before centrifuging for 1 minute. DNA eluents were stored at 4°C.

2.2.5.5 Post-PCR DNA Sequencing

PCR's were performed on patient DNA using the optimal conditions established for each primer pair. Resulting amplification products 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 that is free to download.

2.2.6 Immunohistochemistry

Immunohistochemistry (IHC) was performed on formalin fixed paraffin embedded (FFPE) sections of seventeen tumour biopsies collected from patients in the study cohort using a rabbit polyclonal antibody to MUC3A (Life Technologies, PA5-82409, Rockford, IL, USA) according to their standard operating protocol (SOP).

Briefly, 3-4 micron FFPE sections were cut, dewaxed and rehydrated. Heat mediated antigen retrieval was performed using 1M citric acid (pH6) in a pressure cooker for 90 seconds to reverse the effects of the formalin fixing and to unmask the antigenic binding sites. Blocking for endogenous peroxidase activity was performed by treating slides with 1% H_2O_2 solution for 5-10 minutes followed by the application of appropriately diluted (1:100) primary MUC3 antibody. Goat anti-rabbit secondary antibody was then applied followed by rinsing and application of a chromogenic substrate (buffer containing diaminobenzaldehyde) for 3-5 minutes. Slides were rinsed and immersed in 1% copper sulphate for 3-5 minutes followed by counter staining with haematoxylin, and then blued in 'bluing solution' (NH_4 and water), dehydrated through graded ethanol and mounted in Entellan. Staining was viewed using the Optiview detection system (Ventana XT) automated IHC system. Table 2.11 shows the UCT patient FFPE blocks that were stained for MUC3A protein.

Table 2.11: List of UCT patient biopsies fixed in FFPE blocks that underwent IHC staining, including tumour differentiation description.

Patient Number	Tumour Differentiation
PD39445	Undetermined
PD39446	Moderate
PD39447	Moderate
PD39448	Moderate
PD39449	Poor
PD39450	Moderate
PD39451	Poor
PD39452	Moderate
PD39453	Moderate
PD39454	Moderate
PD39455	Moderate
PD39456	Moderate
PD39457	Moderate
PD39458	Moderate
PD39459	Moderate
PD39460	Moderate
PD51372	Moderate

2.2.7 Preparation of Buffers and Reagents

Buffers and reagents prepared for laboratory use are tabulated below. Distilled, autoclaved H_2O was used in all preparations, and where a specified pH was necessary, NaOH or HCl was used to adjust and achieve the suitable pH levels using an OHAUS Starter 3100 (model ST3100, New Jersey, USA) pH meter.

PBS 10X (pH 7.4)

Component	Amount	Final Concentration
NaCl	80g	1.4M
Na_2HPO_4	14.4g	0.1M
KCl	2g	0.03M
KH_2PO_4	2.4g	0.02M
H_2O	1L	

TBE Buffer (10x)

Component	Amount	Final Concentration
Tris	108g	0.9M
Boric Acid	55g	0.9M
0.5M Na_2EDTA	40ml	0.02M
H_2O	Make up to 1L	

Tris-EDTA Buffer

Component	Volume	Final Concentration
1M Tris (pH8)	1ml	0.01M
0.5M EDTA (pH8)	0.2ml	0.001M
H_2O	98.2ml	

Trypsin-EDTA (pH 7.4)

Component	Amount	Final Concentration
Trypsin	0.5g	0.0007M
EDTA	0.2g	0.0007M
PBS (sterile)	1L	

Penicillin/Streptomycin (100x)

Component	Amount
Penicillin	604mg
Streptomycin	1314mg
H_2O	100ml

Digestion buffer preparation used in DNA isolation from cell cultures. Made up to 10 ml.

Reagent	Volume	Final Concentration
5M NaCl	200µl	0.1M
1M Tris (pH 8)	100µl	0.01M
0.5M EDTA (pH 8)	500µl	0.025M
1% SDS	50µl	0.005M
Proteinase K (10mg/ml)	100µl	0.1 mg/ml

DMEM Culture Medium (pH 7.4) preparation for cell culture use. DMEM powder was dissolved in 800ml distilled water and brought to pH 7.4 using $NaHCo_3$. The solution was filtered using a TPP "rapid" 500 0.2 µm PES filter top (Filtermax, 99505), in a sterile tissue culture hood. Bottles of filtered medium were incubated at 37°C for 48 hours to check for bacterial contaminants.

Component	Amount
H_2O	1L
DMEM powder	1 sachet
NaHCo3	3.7g

DEPC-treated-H_2O (autoclaved)

Component	Volume
DEPC	100ul
H_2O	100ml

Autoclave for 30 min and leave to cool in a sterile hood with the cap removed.

Chapter 3: Bioinformatics Analysis of Novel Viral Insertions

3.1 Introduction

Humans, like all other organisms, are constantly bombarded with micro-organisms such as viruses, that may result in many disease states as a consequence of an acute infection [53]. Many viruses have the ability to integrate their DNA into the genome [55]. Many eukaryotic viruses can be integrated, leading to vertical transmission and potential fixation within a host population [57]. The human genome exhibits a vast degree of invasion by pathogens, such as viral DNA. Integrated viral DNA may account for as much as 8% of the human genome as during the process of evolution several RNA and DNA viruses have become integrated into the vertebrate genome, and there is no reason to believe that the process has ended [53,56].

The causal association between viruses and cancers is estimated to be approximately 10-15% of malignant tumours globally [309], while Cantalupo *et al* (2018) have further suggested that the expression of transforming proteins by oncogenic viruses may contribute to the progression of tumorigenesis through their action on key cellular targets, and in so doing, altering the cellular biology of the host. Changes in host cellular gene expression frequently is the result of repression or activation of essential signalling pathways by these transforming proteins, and thus the integration of viral DNA is now viewed as a specific hallmark of tumorigenesis for certain viruses [53]. Several viruses have already been implicated and associated with the development and progression of certain human cancers [309,310]. Such viruses include hepatitis C virus (HCV), Epstein-Barr virus (EBV) and human herpesvirus 8 (HHV8), along with hepatitis B virus (HBV), human papillomavirus (HPV) and Merkel cell polyomavirus which have been found to be integrated within the host genome [57,310–315].

This chapter, investigates the presence of viral integration in the genomes of three OSCC patients from South Africa with the hopes of identifying possible novel viral integrations. This pilot study was intended to be a quick exploratory study to briefly look at the possibility of novel viral insertions in OSCC patients. We recognise that the incidence of viral integration would be low given the rarity of the events, and that using such a small sample set is not ideal. However, this investigation was merely a quick exploration of the three sets of data available to us at the time.

3.2 Results

3.2.1 DNA Extraction from Patient Biopsies

DNA isolation from tumour and normal biopsies for these patients was performed as per the standard operating protocols described in section 2.2.1.3. The patient DNA had been stored at -80°C, but because these samples had been used as part of another study, only DNA from the tumour sample T547 was still available for further analysis.

Before use in PCR experiments, the integrity of the DNA was validated through agarose gel electrophoresis. Sample concentrations were also determined using a NanoDrop 2000 spectrophotometer (ThermoFisher Scientific, 2000/2000c) at a wavelength of 260 nm. Samples were prepared at 100 ng/ul stocks for future use.

3.2.2 Whole Genome Sequencing

DNA isolated from five paired normal and tumour biopsies, were subjected to WGS at the New York Genome Centre in New York, USA. In the selection criteria, only tumour biopsies with more than 50% tumour tissue, and normal biopsies with 0% tumour tissue were subjected to sequencing and analysis. These patient samples had been part of a previous study in the research group and the availability of this WGS data at the start of this study enabled our initial analysis to proceed while the main cohort samples underwent WGS.

3.2.3 Bioinformatics Investigations into Viral Insertions

Bioinformatics Analysis of the WGS data from three of the five sample-pairs sequenced at the New York Genome Centre was performed using the facilities at the Centre for Proteomic and Genomic Research (CPGR) in Cape Town, South Africa. The remaining two samples were excluded due to sample contamination detected during the quality control process.

Raw data files were converted to FASTQ files and aligned to the Human Reference Genome GRCh37 (also known as Hg19) using the Burrows-Wheeler Aligner (BWA) [299] to map reads with high confidence for variant calling. Variant calling was then performed using the established Vy-PER pipeline [316] through two sub-pipelines for 1) classical variant-calling and 2) virus integration detection. To detect virus integration in the unmapped reads that could not align to the human reference genome, the pipeline first extracted paired-end reads where one end aligned to the human reference genome and the second end did not. Low-complexity reads were discarded and the remaining reads were aligned to known virus

genomes to test for viral origins. Here some reads mapped fully or partially to a virus genome. Candidate virus sequences were then tested for low complexity and the remaining candidate sequences were tested for human origin using exact alignment to the human genome. In the interest of speeding up this alignment step, candidate sequences were exactly aligned to their own small reference sequence window around the end of the read-pair that mapped to the human genome. Candidates failing this alignment were then consecutively exactly aligned against the entire human genome using BLAT (Basic Local Alignment Tool) [316]. DNA sequences from normal samples were compared with matching tumour DNA sequences for each patient in order to identify viral integration differences between the sets. Using this approach, a number of putative viral sequences and integration sites were identified.

Figure 3.1 gives an overview of the bioinformatics workflow used.

The analysis identified a number of foreign DNA insertions within the normal and tumour DNA of the three patients' genomes. Table 3.1 shows the putative foreign DNA integrations that were identified in the normal and tumour DNA of each patient. Multiple foreign integrations were identified in the tumour DNA of patient 547, while only the Human Herpesvirus 7 was identified in the patient's normal DNA. Four foreign integrations in the tumour DNA of patient 569 were detected with only one in the corresponding normal DNA, while three foreign integrations were detected in each of the tumour and normal DNA of patient 607.

Most foreign DNA integrated into these three patients appeared to be bacterial or as expected, Human Herpesvirus. The few insertions that stood out as of potential interest were the *Emiliana Huxleyi Virus 86* in sample T547, and the *Gryllus Bimaculatis Nudivirus* and *Trichodisplasia Spinulosa* Associated Polyomavirus in sample T569. However, the most interesting observation was the presence of the *Autographa Californica* Nucleopolyhedrovirus in all three of the patients, T547 (tumour), N569 (normal) and T607 (tumour). In total, 42 individual insertions of this virus were identified across the three patients. Because this particular virus appeared in all three of the patients at some point, it was decided to investigate this integration specifically.

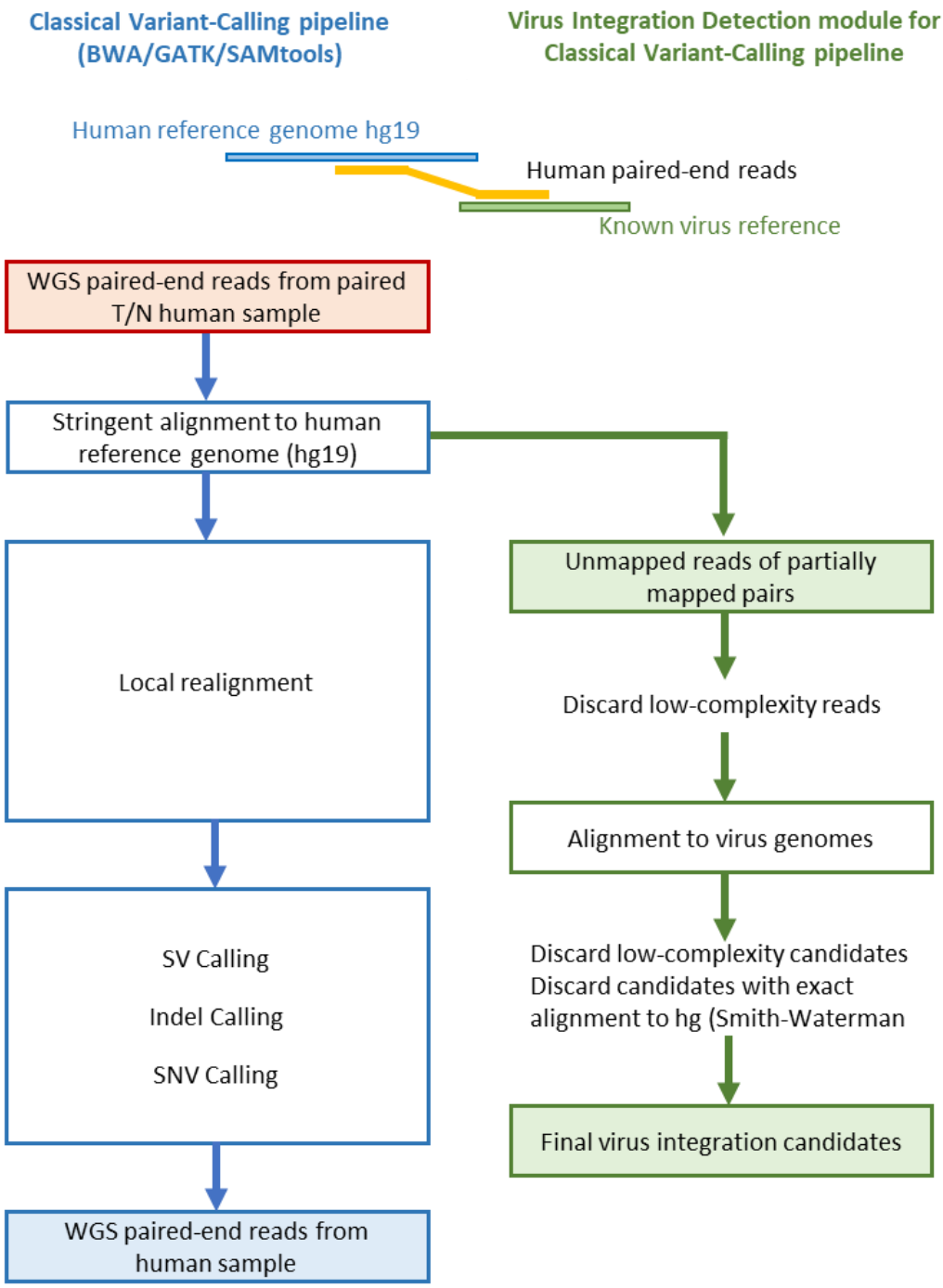


Figure 3.1: Overview of Vy-PER pipeline used for the detection of viral integration. WGS = Whole Genome Sequencing. SV = Structural Variant, SNV = Single Nucleotide Variant, indel = insertions and deletions, T/N = tumour and paired normal.

Table 3.1: Foreign DNA insertions identified for patients 547, 569 and 607 where T represents the tumour genome and N represents the normal genomes.

Accession Number	Name	Number of Insertions
T547		
NC_001623.1	*Autographa Californica* Nucleopolyhedrovirus	22
NC_001716.2	Human Herpesvirus 7	1
NC_007346.1	Emiliana Huxleyi Virus 86	4
NC_019486.1	*Lactobacillus* Phage LF1	1
NC_023503.1	*Streptococcus* Phage 20617	3
NC_031941.1	*Streptococcus* Phage phiARI0131-2	1
N547		
NC_001716.2	Human Herpesvirus 7	3
T569		
NC_001806.2	Human Herpesvirus 1 Strain 17	1
NC_006273.2	Human Herpesvirus 5 Strain Merlin	3
NC_009240.1	*Gryllus Bimacularis Nudivirus*	3
NC_014361.1	*Trichodisplasia Spinulosa* Associated Polyomavirus	17
N569		
NC_001623.1	*Autographa Californic* Nucleopolyhedrovirus	10
T607		
NC_000896.1	*Lactobacillus* Prophage Phiadh	11
NC_001623.1	*Autographa Californica* Nucleopolyhedrovirus	10
NC_005355.1	*Lactobacillus* Prophage Lj 965	18
N607		
NC_0008961.1	*Lactobacillus* Prophage Phiadh	8
NC_001716.2	Human Herpesvirus 7	5
NC_005355.1	*Lactobacillus* Prophage Lj 965	13

The *Autographa Californica* Nucleopolyhedrovirus is a large 130-kb double-stranded DNA virus of the *Baculoviridae* family that commonly infects arthropod insect hosts, more specifically, *Lepidoptera* [317,318]. This virus is known to be able to infect mammalian cells and transport it's genome across the nuclear membrane [319]. The mechanism of DNA transport into the host cell nucleus is proposed to be through docking to nuclear pore complexes in the host nuclear membrane, for translocation of DNA [320]. However, the mechanism of how these patients came to be infected by this virus initially remains unknown.

When assessing the 42 individual *Autographa Californica* Nucleopolyhedrovirus insertions across all three patients, it was observed that the insertion fragments detected on various chromosomes appeared to overlap and it was possible to construct a single 611 base

contiguous viral DNA sequence. This was aligned against a viral sequence database using the National Centre for Biotechnology Information (NCBI) Basic Local Alignment Seach Tool (BLAST) function. This BLAST alignment confirmed that the contiguously aligned sequence detected from the patient samples matched to the *Autographa Californica* Nucleopolyhedrovirus genomic sequence.

3.2.4 PCR Confirmation of Viral Integration

3.2.4.1 Primer Design

Two sets of primers were designed and processed through the IDT Oligo Analyser (https://eu.idtdna.com/calc/analyzer) to identify possible hairpins or self-dimers, and to confirm suitable ΔG values. Primers were prepared as 100 µM stocks and stored at -20°C. 10 µl Working solutions of these stocks were used in PCR reactions.

The first primer set was designed to amplify the bulk of the 611 base sequence, while the second primer set was intended to amplify the initial segment of the sequence (170 bp). Primer sequences for both the long product length (611 bp) and the shorter (170 bp) product length, along with the GC content and annealing temperatures (T_A) are shown in table 2.5, section 2.2.5.1.

3.2.4.2 Primer Optimisation and Patient DNA PCR

The two primer sets were optimised starting with a temperature gradient thermocycling profile (Table 3.2) with the standard reaction mix set up described in section 2.2.5.2. A no-template-control (NTC) was prepared with each reaction as a negative control and to rule out the possibility of reaction mix contamination upon product analysis. GAPDH primers were also used as a positive control for the PCR and were run at an annealing temperature of 60°C as previously optimised. GAPDH product size was expected at 452 bp.

Table 3.2: Temperature gradient PCR profile

Cycle	Conditions		
Initial Denaturation	94°C:	4 minutes	
Denaturation	94°C:	30 seconds	
Annealing	53°C/55°C/56°C:	30 seconds	35 cycles
Elongation	72°C:	45 seconds	
Final Extension	72°C:	7 minutes	
Cooling	4°C:	Hold	

The annealing temperature gradient of 53°C, 55°C and 56°C on tumour DNA extracted from patient 547 is shown in Figure 3.2. There is a long-sequence amplification product at 55°C and a non-specific short product amplification at 53°C, however, there was a strong presence of primer-dimers and GAPDH primers did not appear to amplify, possibly due to primer contamination or degradation.

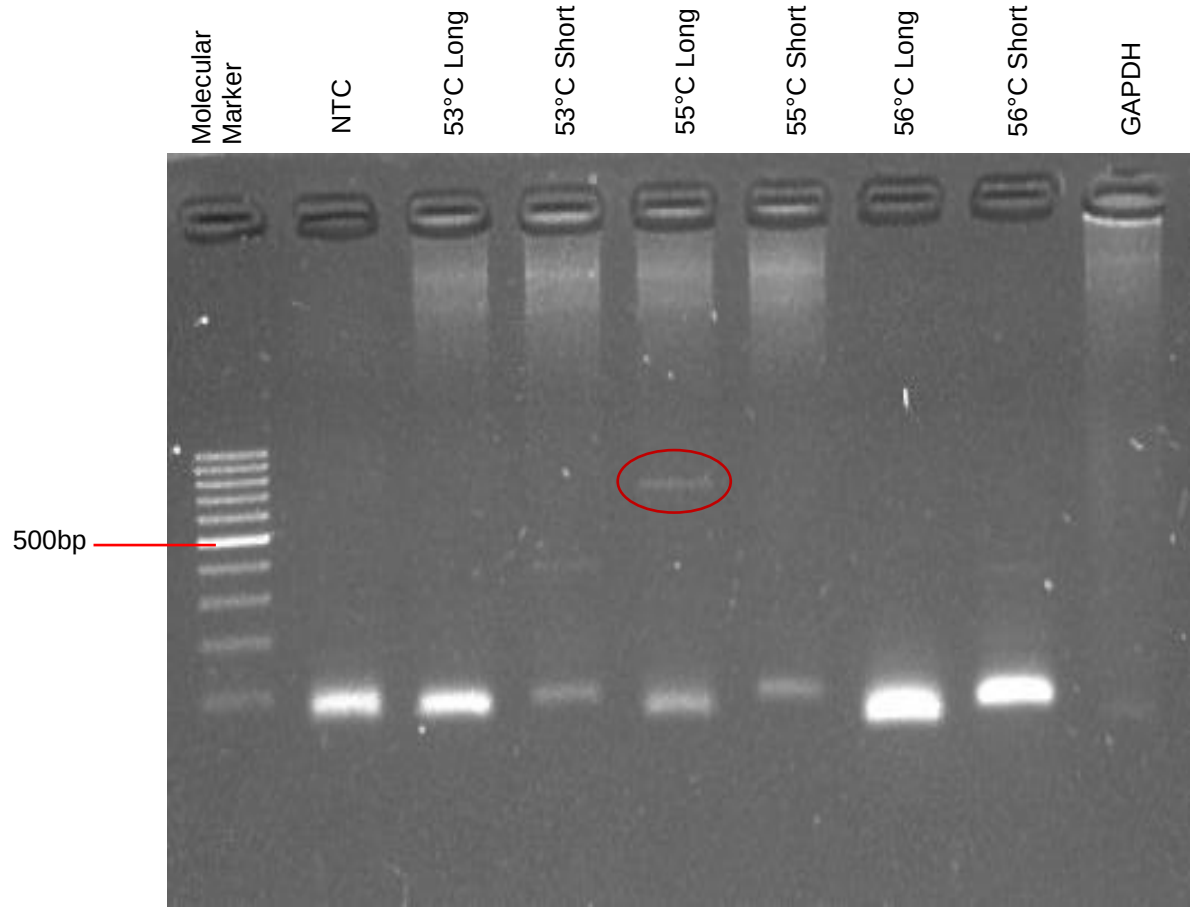


Figure 3.2: Annealing temperature gradient applied to T547 DNA. PCR products separated on a 1% Agarose gel at 100V for 45 minutes. The gel was exposed to UV light for visualisation of the amplification bands for long (611 bp) and short (170 bp) primer sets. The red ring indicate a possible long sequence target band (611 bp) at 55°C. NTC represents the no-template control.

This PCR was repeated using these PCR products as templates for the second PCR. Annealing temperature was set at 53°C and the reaction was prepared in duplicate and run through two separate thermocycler machines. Primer concentration was also decreased from 0.6 µM to 0.25 µM. and the PCR products were separated on a 1.5% agarose gel at 100V for 45 minutes and visualised under UV light.

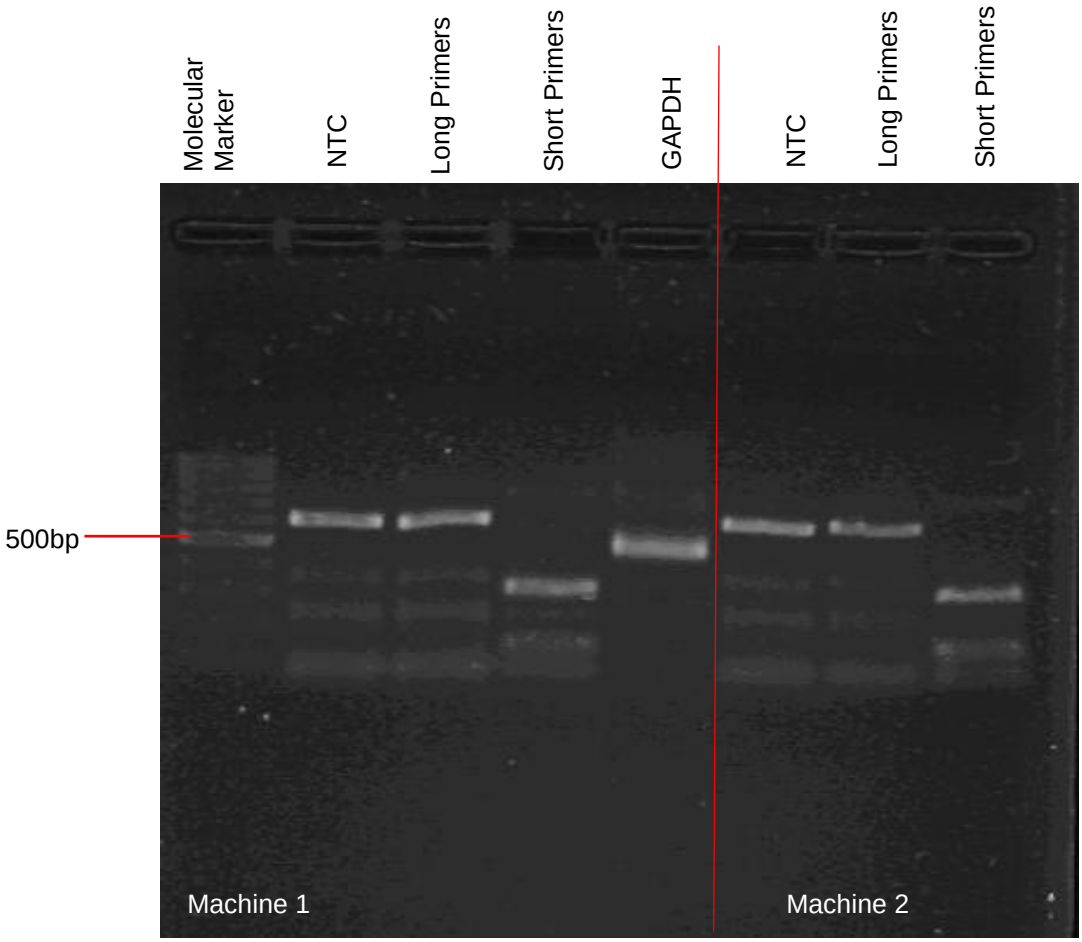


Figure 3.3: Re-amplification of PCR products obtained in Figure 3.2 using both long (611 bp) and short (170 bp) primer sets. PCR run at 53°C. Product amplification was visualised on a 1.5% agarose gel under UV light. NTC represents the no-template control.

Amplification of GAPDH is clearly present in Figure 3.3 and amplification bands can be seen for both long and short primer sets. However, several non-specific bands are present and an amplification band is visible in both NTC lanes. This could be due to accidental pipetting errors when loading the gel for electrophoresis as the bands closely resemble those of the long primer set samples.

A magnesium titration was carried out using the long sequence primers at an annealing temperature of 53°C. The PCR product from the previous reaction (illustrated in Figure 3.2) was again used as a template for the reactions. Figure 3.4 shows the PCR products electrophoresed through a 1.5% agarose gel.

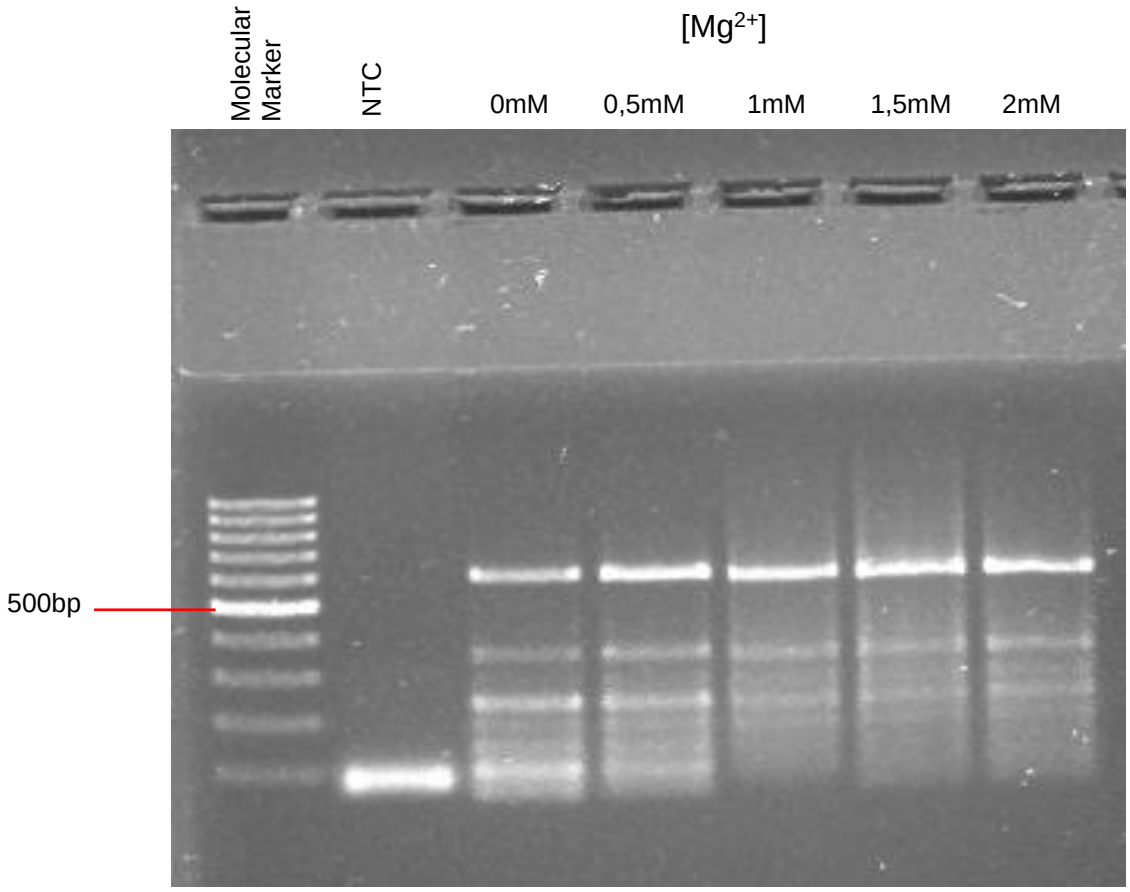


Figure 3.4: Magnesium titration at 53°C with primers for the long 611 base product using PCR products illustrated in Figure 3.2. PCR products were electrophoresed through a 1.5% agarose gel at 100V for 45 minutes and visualised under UV light. NTC represents the no-template control.

Clear amplification of the DNA target at 611bp across all levels of the magnesium concentrations is visible in Figure 3.4. However, a large degree of non-specificity is still present.

After visualisation of the agarose gel indicated the presence of both target sequence bands, electrophoresis was repeated using a 1.5% gel incorporating ethidium bromide into the TBE buffer in the system tank to allow for visualisation of the gel under a manual UV light box for the purpose of target band excision. The target bands were excised from the agarose gel using a sterile surgical scalpel blade and the fragments transferred to clean, sterile micro-centrifuge tubes for weighing and DNA purification using the Qiagen QIAquick Gel Extraction Kit (Qiagen, 28704, Hilden, Germany). The DNA extraction and purification of PCR samples from the gel was carried out as per the manufacturer's instructions as described in section 2.2.5.4. DNA elution through the spin column was repeated two to three times to ensure as

much DNA was eluted as possible. Two eluents were obtained for the longer 611 bp sequence PCR product and three eluents from the shorter 170 bp PCR product.

The eluted DNA was then used as templates in a further PCR. Primer concentration was again decreased back down to 0.25 µM and thermocycling carried out for 30 cycles with the annealing temperature set at 53°C. PCR products were electrophoresed through at 1.5% agarose gel at 100V for 45 minutes and visualised under UV light (Figure 3.5).

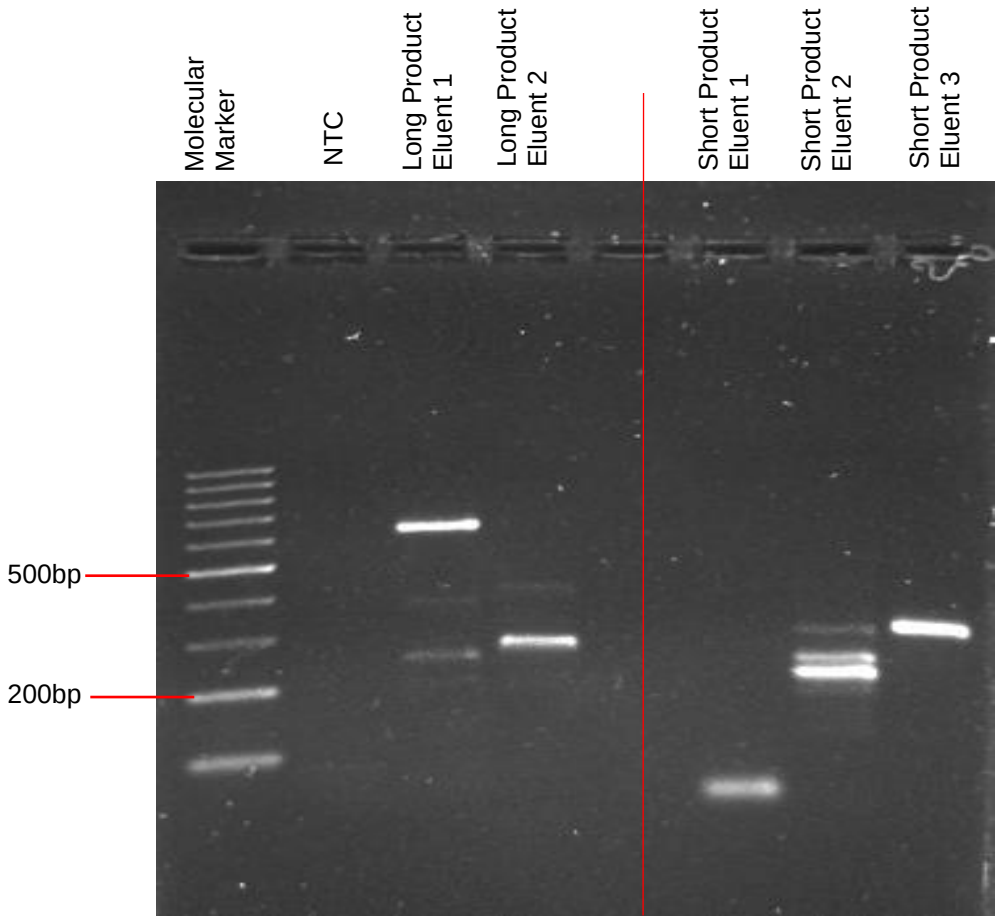


Figure 3.5: Eluents from purified PCR products previously obtained were used as templates for a repeat PCR. Eluent 1 from the long sequence PCR product was reamplified successfully and can be seen as a bright, clear band at 611 bp. No target bands are visible from the re-amplification of the eluents of the short 170 bp sequence. Eluent 1 refers to the first 50 µl aliquot passed through the elution column, eluent 2 refers to the second 50 µl aliquot and eluant 3 refers to the third 50 µl aliquot passed through the column to elute DNA. NTC represents the no-template control.

The PCR product for the long, 611 bp sequence was subjected to post-PCR Sanger sequencing at two separate sequencing facilities, as described in section 2.2.5.4. The resulting chromatogram sequences were blasted against the NCBI BLAST tool selecting for a viral sequence database. No matches to the *Autographa Californica* Nucleopolyhedrovirus reference sequence were found, but rather, sequences mapped to some cloned human gene sequences. Furthermore, the sequences obtained from the separate sequencing facilities did not match each other.

The question of primer specificity was raised and it was decided to re-design new PCR primers for some of the individual virus fragments identified in the bioinformatics analysis. From the 22 fragments that were detected in the tumour sample T547, a further three sets of primers were designed. A PCR was performed using all three sets of primers designed for individual viral fragments of the *Autographa Californica* Nucleopolyhedrovirus present in the DNA sample of tumour biopsy T547. Biopsy T547 DNA was used along with the corresponding normal blood DNA (N547) (previously extracted) from the same patient. Working stocks of the DNA were prepared at 100ng/µl for PCR use.

The first PCR was prepared in 25 µl reactions as before and the thermocycling profile was set with annealing temperatures of 54˚C, 56˚C and 62˚C for 35 cycles. The PCR products were electrophoresed through a 1.5% agarose gel at 100V for 35 minutes together with Novel Juice for visualisation under UV light (Figure 3.6). Primers for fragment 1 were run at an annealing temperature of 54˚C, primers for fragment 2 were run at 54˚C and 56˚C, and primers for fragment 3 were run at 54˚C and 62˚C.

Faint amplification bands were visible only at 54˚C with primers for fragment 2, although the bands were slightly larger than anticipated. We speculated that these visible amplification bands could possibly be our target products and to investigate further, a primer concentration gradient was performed at 54˚C with an incremental increase on 0.5 µM. Primers for fragment 2 were used in the reaction mix and GAPDH primers were run as a positive control. PCR samples were electrophoresed at 70V for 1 hour to allow for optimal and clear migration of products, before visualisation under UV light (Figure 3.7).

Results of this PCR remained ambiguous with non-specific amplification visible on the gel.

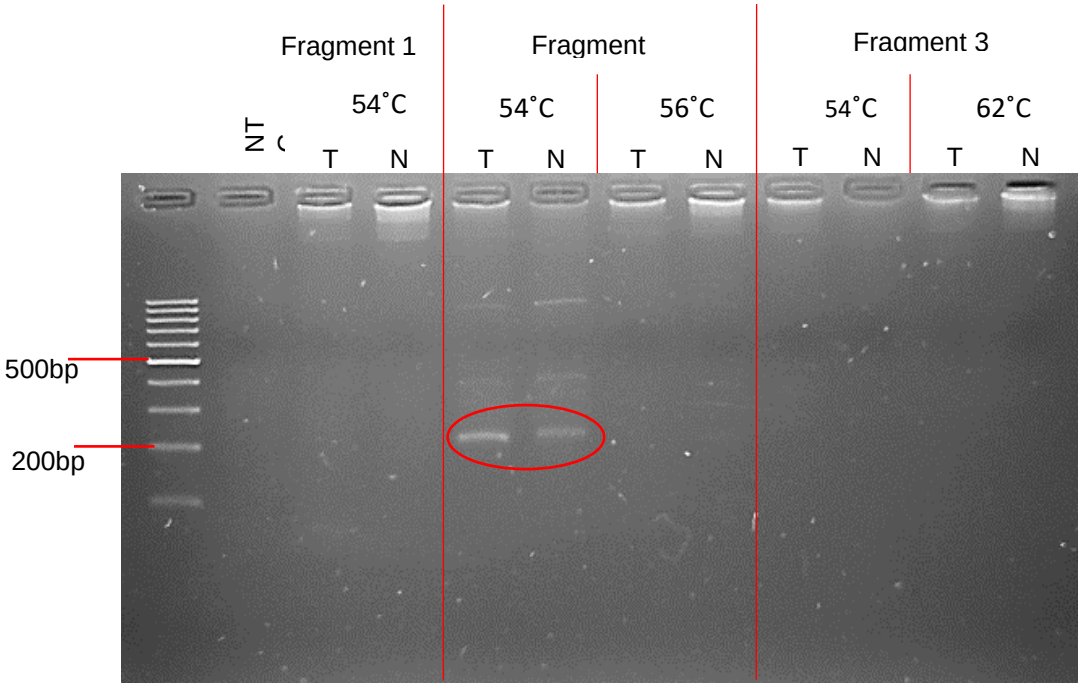


Figure 3.6: Three new sets of primers for three separate *Autographa Californica* Nucleopolyhedrovirus insertion fragments. T represents tumour sample T547, and N represents corresponding normal samples N547. Primers were run at separate annealing temperatures and PCR products visualised on a 1.5% agarose gel under UV light. Fragment 1 target size= 134 bp, fragment 2 target size = 142 bp and fragment 3 target size = 118 bp. NTC represents the no-template control. The red ring indicates the target bands at 54°C for Fragment 2.

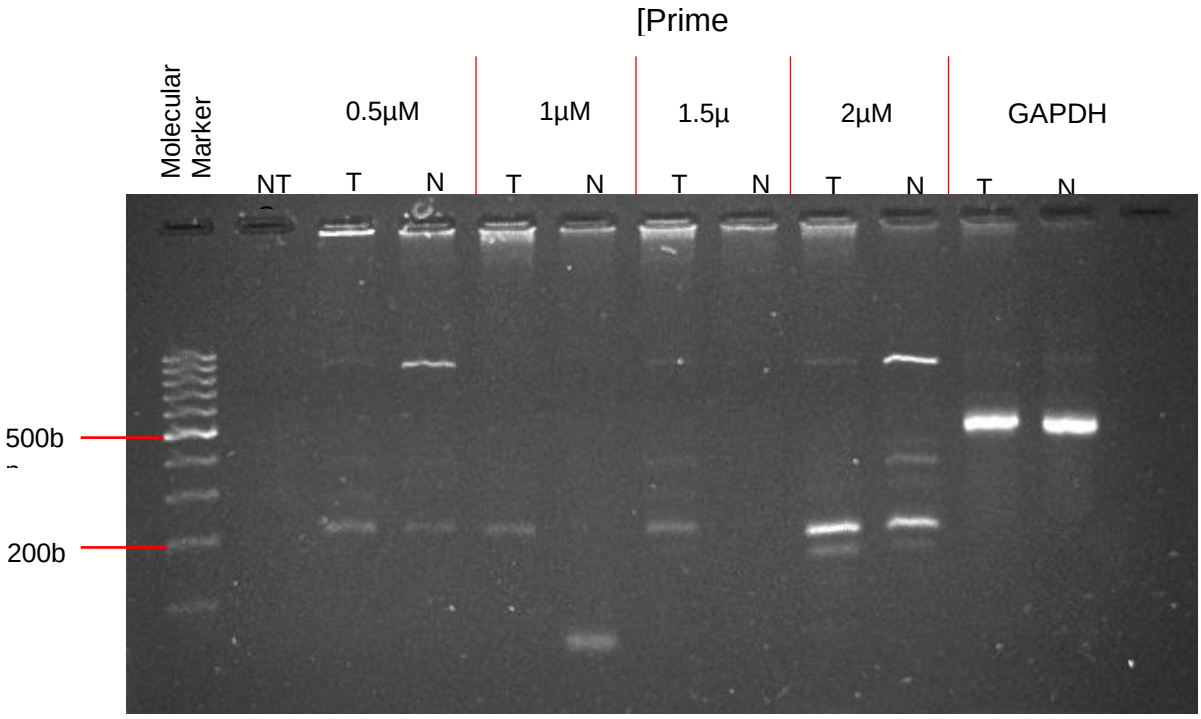


Figure 3.7: Primer concentration gradient with primers for fragment 2 (142 bp) as well as GAPDH control. T represents tumour sample T547, and N represents corresponding normal samples N547. Visualisation of PCR products under UV light post gel electrophoresis at 70V for 1 hour. NTC represents the no-template control.

GAPDH (452 bp) amplified clearly in both the tumour and normal sample, but a large degree of non-specific amplification was visible at all concentrations with primers for fragment 2, although possible target bands might be close to the 200 bp mark, similar to Figure 3.6.

A further PCR was carried out using primers for fragment 1, at 50˚C, 52˚C, and 54˚C for 40 cycles and the resulting products electrophoresed through a 1.5% agarose at 70V for 1 hour. Multiple amplification bands were visualised at 50°C for the tumour DNA (Figure 3.8A) and a possible target product shown at around 134 bp (encircled in red). The PCR product for this sample was used as a template for a repeat PCR at 49˚C, 50˚C, and 51˚C and a probable target band was visible at 134 bp (Figure 3.8B).

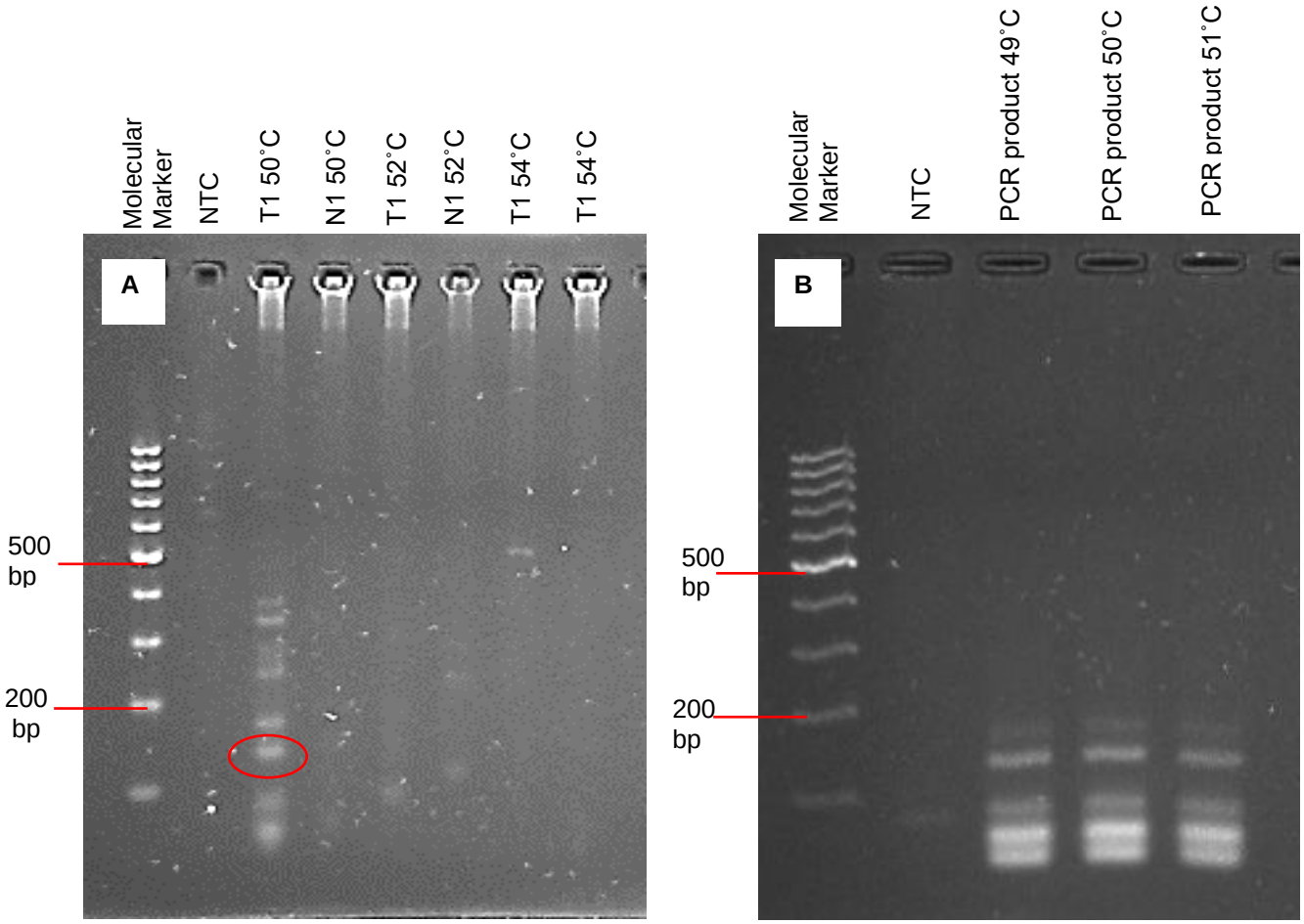


Figure 3.8: **A**) PCR amplification with primers for fragment 1 (134 bp) carried out at 50°C, 52°C and 54°C. Target product size circled in red. **B**) The tumour sample PCR product from (A) was then used as a template for a repeat PCR at 49°C, 50°C and 51°C. The presence of non-specific amplification is still visible but target size bands are clear. NTC represents the no-template control. The red ring indicates the target band at 50°C in the tumour sample for Fragment 1 in (A).

A magnesium titration was then carried out with 0.5 µM incremental increases from 0-2 µM, at both 49°C and 50°C using the same PCR product from Figure 3.8A (gel image not shown). Following gel electrophoresis, target size bands were excised from the gel for purification. Target bands were found at 49°C at a Mg^{2+} concentration of 2 mM, as well as at 54°C at both at 1.5 mM and 2mM concentrations. These bands were excised under UV light with a sterile surgical scalpel blade and transferred to clean microcentrifuge tubes for weighing and DNA purification from the gel. The Qiagen QIAquick Gel Extraction Kit (Qiagen, 28704, Hilden, Germany) was used for this purification step and DNA was eluted in 30-50 µl volumes as described in section 2.2.5.4. These eluents were used as templates in a final PCR carried out at annealing temperatures of 49°C and 54°C with 30 cycles of the PCR profile.

The PCR products were electrophoresed and visualised on a 1.5% agarose gel (Figure 3.9) with ethidium bromide to allow for the further excision of the clear target bands.

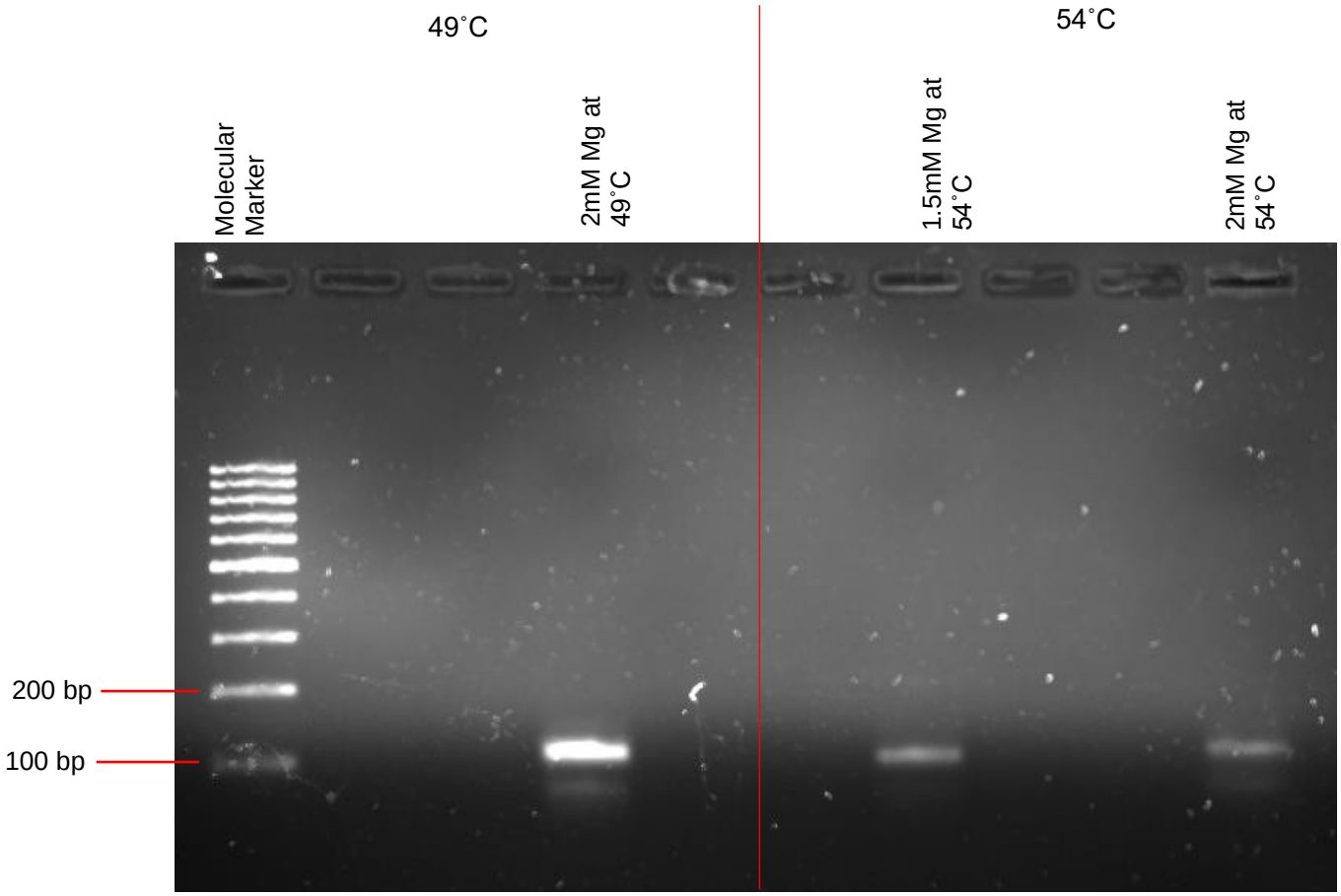


Figure 3.9: PCR amplification using eluted DNA purified from bands excised from agarose gel in the previous Mg^{2+} titration (gel image not shown). Annealing temperatures set at 49°C and 54°C. Target size bands are visible in the anticipated size range.

The bands visualised in Figure 3.9 were excised from the gel under UV light as before. DNA was extracted and purified from the gel as before, and purified products were subjected to Sanger sequencing. Disappointingly, the three sequences did not match the *Autographa Californica* Nucleopolyhedrovirus reference sequence when blasted on NCBI, nor did they match the previously sequenced samples.

These sequences were aligned, together with the original 611 bp long sequence, to a vector database on the NCBI resource to investigate whether amplification of bacteria or vectors had occurred. An NCBI BLAST alignment of the original 611 bp sequence together with the forward sequence matched strongly to the terminal end of a Baculovirus vector pBacPAK-His3, beta-lactamase gene. The reverse sequence obtained had a moderate match to the terminal end to a cloning vector, pBacPAK1.

Of the smaller individual sequence fragments blasted, the forward sequences matched moderately to the same pBacPAK-His3, beta-lactamase gene while reverse sequences had no significant similarity.

3.3 Discussion

This brief exploratory pilot study aimed to determine the feasibility of further extensive investigations of a larger sample cohort to determine the presence of novel viral DNA integration into the genomes of OSCC patients. The investigation was carried out on three sets of matched tumour-normal OSCC WGS data available, and while we acknowledge that the rarity of viral integration events demands a greater sample cohort for accuracy, the results determined in this study corroborated those found by the Wellcome Sanger Institute indicating no obvious viral integrations among the larger cohort.

Bioinformatics analysis of WGS's indicated the presence of a number of viral integrations with the *Autographa Californica* Nucleopolyhedrovirus identified in all three of the patients. Considerable time was spent optimising primers and repeating PCR's, yet sequencing of the amplified DNA showed no match to the anticipated viral reference sequence, but rather mapped to bacterial clone vectors pBacPAK-His3 and pBacPAK1.

This result prompted the re-evaluation of primer specificity and PCR conditions, along with speculations into sequencing and bioinformatics analysis methods. The sequence mapping to bacterial clones and vectors also raised the question of how the vector sequences landed

up in the patient genomes, however this fell beyond the scope of this project but should be carefully considered for future investigations to eliminate artifacts.

It would appear that the bioinformatics pipeline used was not appropriate and provided false positives of viral integrations. WGS data was mapped and aligned to a reference viral genome database to confirm the speculated viral insertions. However, the presence of *Autographa Californica* Nucleopolyhedrovirus was not found to be present in any of the three patient tumour or normal samples, nor were the other *Emiliana Huxleyi Virus 86*, the *Gryllus Bimaculatis Nudivirus* or the *Trichodisplasia Spinulosa* Associated Polyomavirus insertions that has also previously been identified.

These conflicting findings together with the inability to identify the *Autographa Californica* Nucleopolyhedrovirus insertions through PCR suggested that the analysis performed up to this point had been flawed.

This initial short exploratory investigation was a pilot study carried out before the onset of the main study, and as such, WGS data from only three patients was available for use. We recognise that due to the rarity of insertion events, investigations with a larger sample cohort and a more robust and reliable bioinformatics analysis pipeline may allow for the detection of novel viral insertions. However, this was an exploratory investigation showing ambiguous and inconclusive results. Furthermore, foreign insertion investigations that were also performed by the Wellcome Sanger Institute as part of a larger study also revealed no obvious viral integrations among the full cohort of 35 sample pairs. Therefore, after careful consideration the focus of the project was shifted to move on and investigate the presence and possible translocation of Human Endogenous Retroviruses (HERV's) together with the presence of somatic mutations in these pilot samples, and later in a larger sample cohort. These investigations and findings are discussed in Chapters 4 and 5.

Chapter 4: Analysis of Human Endogenous Retrovirus (HERV) integration in the human genome.

4.1 Introduction

This chapter focuses on investigations into the integration of well-researched and well documented Transposable Elements (TE's), Human Endogenous Retroviruses (HERV's).

TE's are best described as discrete DNA sequences that possess the ability to translocate within the genome from one position to another, giving rise to interspersed repeats [84–86]. Classification of TE's is dependent on their ability to encode genes for transposition, and can either be autonomous, such as Long Interspersed Nuclear Elements (LINE's), or non-autonomous, such as Short Interspersed Nuclear Elements (SINE's). LINE's effectively encode all sequences that move in the genome, while SINE's are more structurally deficient and depend on LINE's for movement. SINE's have evolved independently but in parallel to LINE's, with the most common SINE family being Alu elements. There are also some TE's where the distinction between autonomous and non-autonomous transposition remains unclear, such as HERV's. These however, have been particularly well characterised and it is commonly believed that their integration leads to phenotypic alterations in the human genome. Thus variations arising in host genomes can often be attributed to these potent, broad-spectrum mutator elements [84].

Endogenous retroviruses have been detected in every animal species tested, including humans, and constitute approximately 8% of the human genome [111,308]. It is widely accepted that HERVs originated from infection by an ancestral exogenous retroviral element that became integrated into host germ cell DNA during evolution more than a million years ago, resulting in Mendelian vertical transmission [77,88,91]. Most of these incorporated sequences are highly defective and non-protein coding. As a result of their longevity within host genomes, they have accumulated a large degree of mutations including deletions, insertions, truncations and frameshifts [87,88]. However, some integrated viral elements can produce both mRNA transcripts as well as viral proteins in the host germline cells. These include the HERV-K family of viruses, reportedly some of the most recently biologically active retroviral elements in the human genome with the least number of mutations and the ability to encode functional retroviral proteins, and produce retrovirus-like particles [88,91,92].

Two of the most well-known and well researched HERV-K retroviral elements include HERV-K113 and HERV-K115. These are full-length proviruses with open-reading frames and present in only a portion of the population [56,88]. HERV-K113 and HERV-K115 are far more prevalent in Africa than in other regions, with a frequency of between 30-40%, and racial origin appears to be an important factor of integration [88]. It is further suggested that the initial infection and incorporation of HERV-K113 likely occurred in Africa either during or after the migration of *Homo sapiens* north and eastward between 150 000 – 200 000 years ago [109–111].

HERVs have been implicated in the development of some cancers, suggesting links with breast cancer, lymphoma, melanoma, ovarian cancers and prostate cancers. For reviews see [92,321]. The disruption of host gene regulation through hijacking and manipulation by these retroviral elements influences the expression of cellular genes, posing a long-lasting burden on the genome [57]. Furthermore, the discovery of significantly elevated expression of HERV elements in cancer cells has driven research into the investigation of HERVs as biomarkers for malignant transformation, staging and prognosis of cancers [99,105,106]. Two possible pathogenic mechanisms might exist whereby HERV-K elements influence the host. Firstly, the possibility that auto-antibodies may be induced in the host by the HERV-K viral proteins, or that these proteins may function as onco-proteins. Secondly, the functions of HERV-K loci or the long terminal repeats (LTRs) might induce dysregulation of the host genome, including recombination with HERV-K sequences and leading to chromosomal instability and large-scale chromosomal abnormalities. Additionally, the aberrant expression and regulatory function of HERV-K transcripts on proximal host-genes has been identified in numerous diseases and pathologies [92]. It has also been reported that the transposable ability of HERV elements might influence tumorigenesis due to retroviral movement causing disruption and instability of the host genome [108].

Taking into consideration the outcome of the investigations into novel viral insertions, the aim of this study therefore evolved into investigating the presence of HERVs in a larger patient cohort with the objective to identify whether the integration of these viral elements could be linked to, or influence the occurrence of somatic mutations present in the genomes of these OSCC patients.

4.2 Investigations of HERV Insertions

4.2.1 PCR Primers

HERV-K113 primers sourced from the literature [307,308] (shown in section 2.2.5.1, Table 2.7) were used to detect integrated HERV-K113 at the pre-integration/LTR (long terminal repeat) sites in the pilot patient DNA sequence data that was previously used in the novel virus insertion analysis.

4.2.2 PCR for HERV Insertions

As with the PCR investigations described in Chapter 3, only DNA from patient 547 was available for experimentation as the DNA from patients 569 and 607 had been depleted. An exploratory PCR was set up using the HERV-K113 primers on patient tumour DNA (T547) and normal blood DNA (N547), as well as non-patient blood DNA.

The PCR reaction mix was prepared in 25 µl volumes using 100 ng/ul DNA template as described in section 2.2.5.2. An annealing temperature of 56˚C was used and thermocycling was set for 35 cycles (Table 4.1). GAPDH was run as a control at an annealing temperature of 60˚C.

Table 4.1: Thermocycling conditions for HERV-K113 primers.

Cycle	Conditions		
Initial Denaturation	94°C:	4 minute	
Denaturation	94°C:	30 seconds	
Annealing	56°C:	30 seconds	35 cycles
Elongation	72°C:	30 seconds	
Final Extension	72°C:	7 minutes	
Cooling	4°C:	Hold	

After PCR, amplification products were electrophoresed through a 1.5% agarose gel at 100V for 35 minutes and visualised under UV light. Figure 4.1 shows the target bands for the HERV-K113 product (303 bp) in the patient tumour (T547), normal (N547) and the non-patient blood samples. The GAPDH positive control can also be observed at the target size of 452 bp.

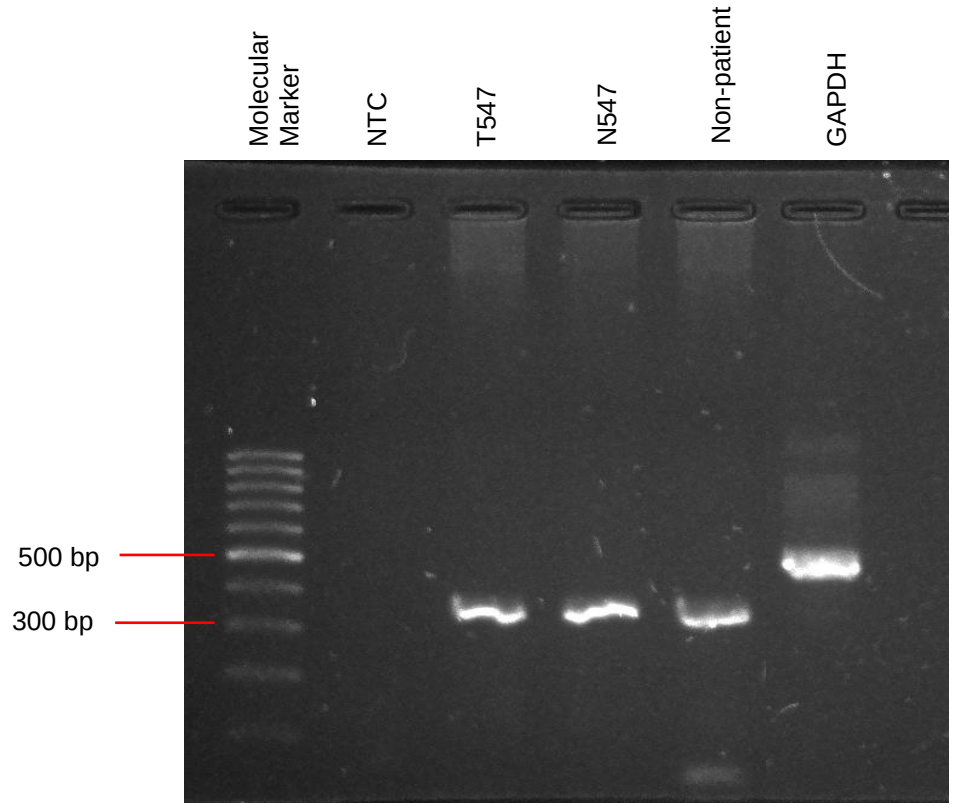


Figure 4.1: Gel visualisation of PCR products (303 bp) using HERV-K113 primers on patient 547 tumour and normal DNA, as well as non-patient blood DNA. GAPDH (452 bp) primers were included as a positive control to confirm the efficacy of the thermocycling conditions. NTC represents no-template control.

The amplification products obtained for samples T547, N547 and the non-patient blood DNA were subjected to bidirectional Sanger sequencing as described in section 2.2.5.4. Figure 4.2 shows the DNA sequence chromatogram that was obtained for each of the three PCR products. The resulting sequence (199 bp) was then blasted on the NCBI BLAST online resource and returned a 73% query cover. That is, 73% of the 199 bp were found to match to HERV-K113 with 146/147 homology (see Figure 4.3).

A 5'TACAAAGCAGAGAATATACTTGCTTTCAGCATTTTTAAGGTTTTTAGTTTTCTAGTACTCATCTTG
TTGAAATGATTTGCCTTGTTCAATATTGTTTTCTTCTGAAAATAGTTACAAATTCATACTTACACAA
ACTCACTTACTCTATAATTTTCTTACACCTAACCTTTATCTTTAGACTAACATATATTGAACTCTAT3'

B

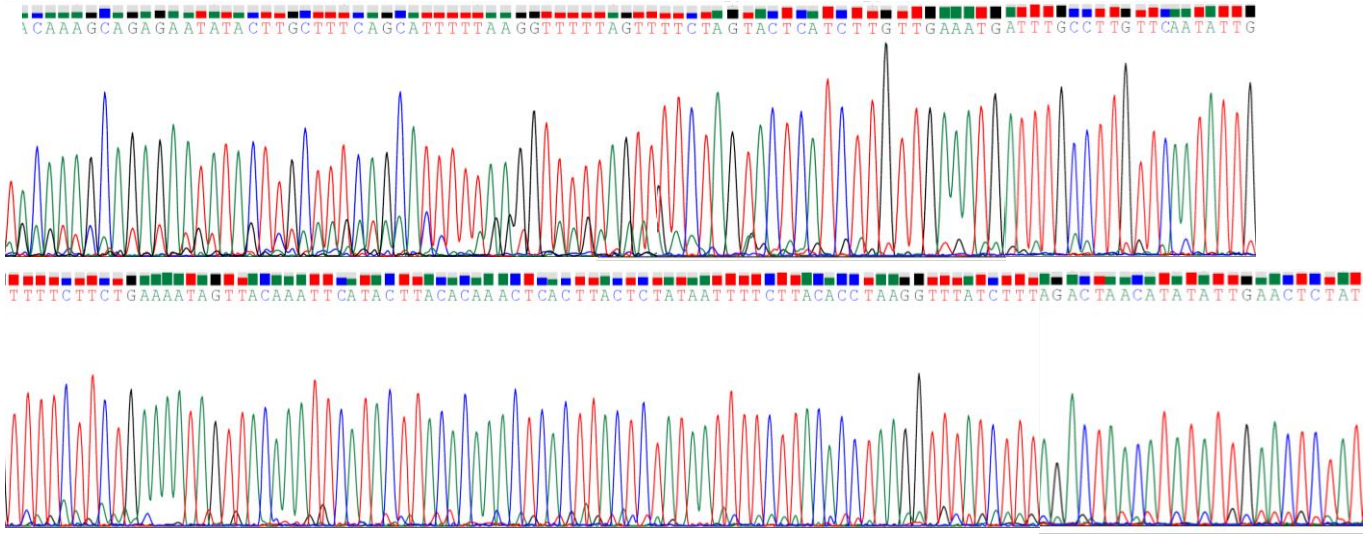


Figure 4.2: **A**) HERV-K113 sequence and **B**) the Chromatogram obtained from Sanger sequencing of PCR products of samples T547, N547 and non-patient blood DNA. All three samples returned the same chromatogram sequence. This sequence was run through the NCBI BLAST online resource.

These sequence results confirm the presence of HERV-K113 in the pilot patient study.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Eukaryotic synthetic construct chromosome 19	357	1158	100%	8e-95	98.99%	CP034497.1
Homo sapiens chromosome 19 clone RP11-420K14, complete sequence	357	703	100%	8e-95	98.99%	AC092364.3
Eukaryotic synthetic construct chromosome 19	351	1152	100%	4e-93	98.49%	CP034522.1
Homo sapiens FOSMID clone ABC9-43816200N12 from chromosome unknown, complete sequence	267	377	100%	1e-67	99.32%	AC226770.1
Homo sapiens endogenous virus HERV-K113 5' long terminal repeat, partial sequence	267	267	73%	1e-67	99.32%	AF387847.1

Download GenBank Graphics

Homo sapiens endogenous virus HERV-K113 5' long terminal repeat, partial sequence

Sequence ID: AF387847.1 Length: 300 Number of Matches: 1

Range 1: 104 to 250 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
267 bits(144)	1e-67	146/147(99%)	0/147(0%)	Plus/Plus

```
Query  1    ACAAAGCAGAGAATATACTTGCTTTCAGCATTTTTAAGGTTTTTAGTTTTCTAGTACTCA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  104  ACAAAGCAGAGAATATACTTGCTTTCAGCATTTTTAAGGTTTTTAGTTTTCTAGTAGTCA  163

Query  61   TCTTGTTGAAATGATTTGCCTTGTTCAATATTGTTTTCTTCTGAAAATAGTTACAAATTC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  164  TCTTGTTGAAATGATTTGCCTTGTTCAATATTGTTTTCTTCTGAAAATAGTTACAAATTC  223

Query  121  ATACTTACACAAACTCACTTACTCTAT  147
            |||||||||||||||||||||||||||
Sbjct  224  ATACTTACACAAACTCACTTACTCTAT  250
```

Figure 4.3: NCBI BLAST results of the HERV-K113 sequence of PCR products. HERV-K113 showed a 73% query cover with a 99% homology (146/147 identities matched). Max (maximum) score indicates the highest alignment score calculated from the sum of rewards for matched nucleotides and penalties for mismatches and gaps. Total Score provides the sum of alignment scores of all segments from the same subject sequence. Query Cover is the percentage of the query length that is included in the aligned segments. E-value denotes the number of alignments expected by chance with the calculated score or better. And, Ident (identity) shows the highest percentage identity for a set of aligned segments to the same subject sequence.

4.3 Analysis of Patient DNA

The next step of the bioinformatics analysis sought to implement pipelines to determine the locations and influences of HERV insertions within the sample cohort, combined with investigations to identify the presence of somatic mutations that might be linked to these HERV insertions. The literature indicates that HERV's, particularly HERV-K, may play a potential role in disease progression, especially cancer (reviewed in Li et al., 2019; and Zhang et al., 2019) thus we hoped to explore any possible connections between HERV insertions, somatic mutations and OSCC development and progression.

The raw WGS data of the pilot three tumour-normal sample pairs sequenced at the New York Genome Centre, along with the WGS data from the thirty tumour-normal sample pairs sequenced at the Wellcome Sanger Institute were transferred to the SANBI servers where decryption and extraction of useable BAM (Binary Alignment Map) files to be run through relevant software.

All the bioinformatics methodologies were performed in-house, by the student using only previously published and well documented software for tumour-normal paired analyses, as described below. Pipelines were established according to software guidelines. Analysis and processing were performed using the Linux command line, and all scripts and code written and used can be found at https://github.com/VictoriaPatten/phd-scripts/tree/main/ERVcaller.

Use was made of Ilifu Cloud Computing and Storage (www.ilifu.ac.za) to expand the computing power and storage space available. Ilifu is a big data infrastructure for data-intensive research in bioinformatics and astronomy and is operated by a consortium of universities and research organisations in the Western and Northern Cape in South Africa. All of the extracted BAM files from the tumour-normal paired samples were transferred across from SANBI for storage on the Ilifu cloud. This was achieved using a Globus account transfer link.

Software containers were built by the systems developers on both the SANBI and Ilifu clusters, and included all necessary software tools needed for the analysis of the sequence data.

After receiving transfer of all of the extracted BAM files, files were sorted, and indices were created using Samtools [324] functions *samtools sort()* and *samtools index()* respectively. Figure 4.4 shows an overview of the bioinformatics workflow used.

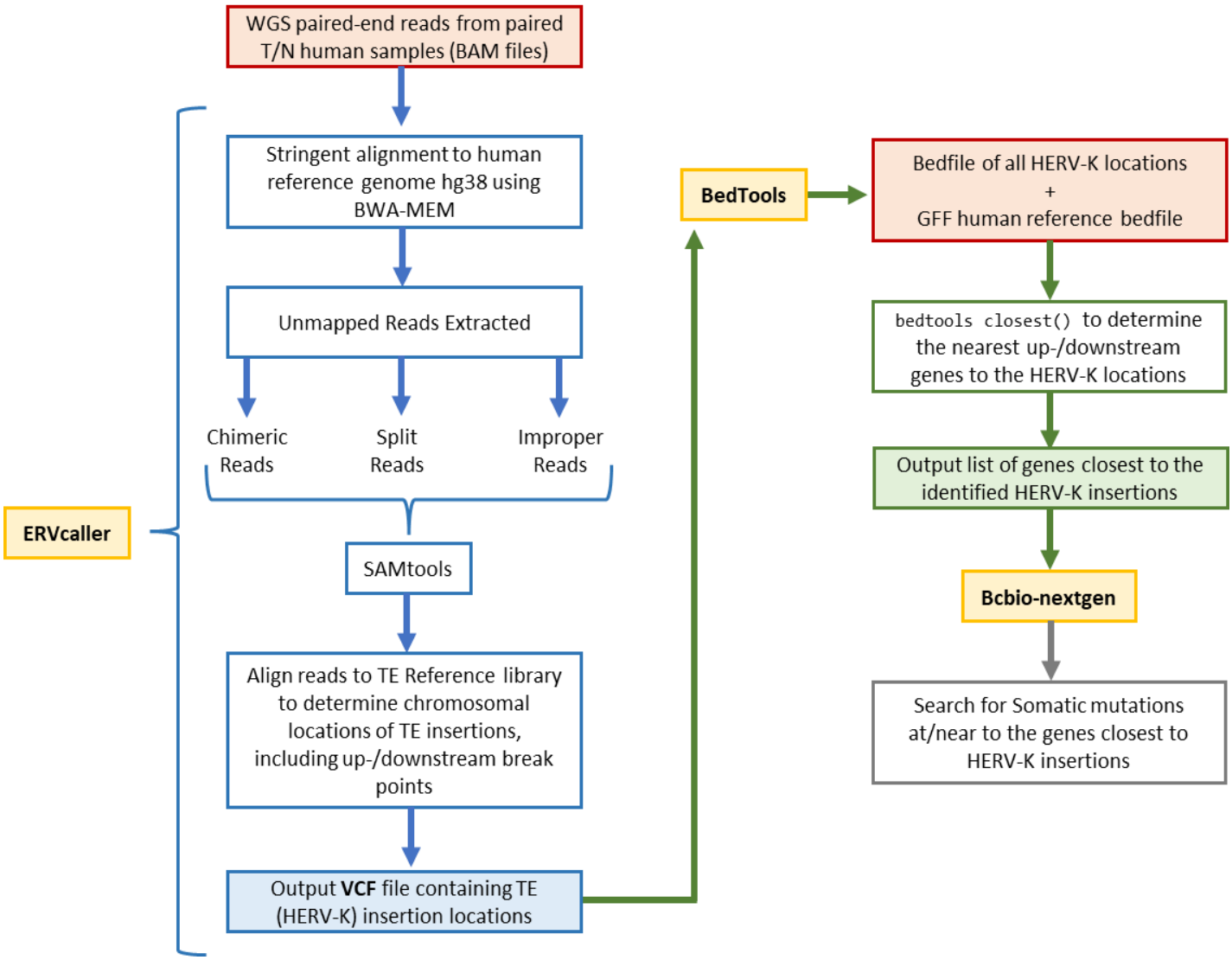


Figure 4.4: Overview of the bioinformatics pipeline used to determine the chromosomal locations of HERV-K insertions as well as the tools used to identify the closest up- and downstream genes and somatic mutations. WGS = Whole Genome Sequencing; T/N = tumour and matched normal; TE = transposable elements; VCF = variant call format, HERV = human endogenous retrovirus

4.4 Wellcome Sanger Institute DNA sequence data

DNA from thirty-five paired blood samples and tumour biopsies were subjected to WGS at the Sanger Institute in Cambridge, UK as described in section 2.2.2.2. OSCC patients were recruited from Groote Schuur Hospital in Cape Town as well as Charlotte Maxeke Hospital in Johannesburg. DNA was isolated from the blood and biopsy samples (as described in section 2.2.1) and was shipped to the Wellcome Sanger Institute in the UK for WGS. Samples were SNP genotyped and sequenced using 150 bp paired-end reads to a depth of >30x coverage. The BWA aligner was used to align the reads while PCR duplicates were

marked using Picard tools. Caveman and Pindel algorithms were used to call SNPs and indels while copy number analysis was performed using ACAT.

Samples from 17 females and 13 males were sequenced and made up the initial sample cohort of thirty OSCC patients. Quality control of the sequencing data was performed by the Wellcome Sanger Institute as part of their sequencing pipeline. These QC metrics are presented in Appendix 1, Table A1.1.

4.5 Bioinformatics Analysis

4.5.1 ERVcaller Pipeline

ERVcaller is a software tool that was developed in 2018 for the identification and accurate genotyping of the allele frequency of non-reference, unfixed transposable element (TE) insertions, particularly endogenous retroviruses (ERVs), in the human genome using Next Generation Sequencing (NGS) short read sequence data [325]. When developing the software and comparing with other existing tools available, the developers identified that ERVcaller showed the most efficient and accurate detection of unfixed TE insertions, detecting the most precise breakpoints, and consistently achieving high sensitivity and precision with the various TE references of differing sequence complexities [325]. The software consists of a three module pipeline process: 1) extraction of unmapped reads, 2) obtaining supporting reads, and 3) detection and genotyping of the unfixed TE insertions identified.

ERVcaller uses either BAM or FASTQ files as input into the pipeline, and aligns the raw reads to a reference genome directly using BWA-MEM [326]. In this pipeline setup, the latest GRCh38 human reference genome was used. All reads that did not fully map to the reference genome were extracted and from these, chimeric reads, split reads and improper reads were obtained using Samtools, a library software package used for parsing and manipulating SAM/BAM format alignments [324]. These supporting reads were aligned to a TE reference library and were then used to determine the chromosomal locations of the TE insertions including upstream and downstream breakpoints. Confident TE insertions were identified only when the following criteria were met: 1) the insertion had at least two supporting reads, 2) one of the supporting reads must either be a chimeric or an improper read, 3) each supporting read must have an average alignment score of more than 30, and 4) a minimum of 50 bp was required for each read length mapping to the human reference genome. With this stringent filtering, high confidence unfixed TE insertions were thus

genotyped based on the reads crossing the breakpoint [325]. The output VCF file provides the chromosome, position, sample ID, reference allele, alternate allele, quality score and the filtering information including the 'status' parameter. The status of each TE event indicates the read type. Where the status of the detected TE is indicated as: 0 = Inconsistent direction for the supporting reads; 1 = One breakpoint detected by only chimeric and/or improper reads without split reads; 2 = Only one breakpoint is detected and covered by split reads; 3 = Two breakpoints detected, and both of them are not covered by split reads; 4 = Two breakpoints detected, and one of them are not covered by split reads; 5 = Two breakpoints detected, and both of them are covered by split reads. In these analyses TE's of all status levels were considered.

ERVcaller v1.3 was installed into singularity containers on both SANBI and Ilifu servers. The initial analysis, setting up and execution of the pipeline was carried out using the three patient tumour-normal pairs previously sequenced at the New York Genome Centre from a previous study (Samples 547, 569 and 607). The WGS data from the thirty pairs of samples sequenced at the Wellcome Sanger Institute were transferred to SANBI and run through the pipeline as well.

Bash scripts with arguments for the human reference genome to be used (GRCh38), the sequencing type (paired-end) and the specific input files to be mapped were prepared for each patient (https://github.com/VictoriaPatten/phd-scripts/tree/main/ERVcaller/erv.sh).

The resulting ERVcaller output was tabulated and converted to VCF (variant caller format) files for subsequent genetic association analysis. These VCF files thus provided information on HERV-K insertion locations on each chromosome per patient.

4.5.2 BEDTools

To be able to associate the detected HERV-K TE insertions with the proximal genes per chromosome, it was necessary to use a specialised software created for genomic arithmetic.

BEDTools v2.30.0 [327] is a fast and flexible suite of utilities for performing genome analysis tasks and allowing for intersecting, merging, counting, complementing and shuffling of genomic intervals from multiple files, in multiple formats.

It is possible to determine whether distinct genomic features (such as aligned sequence reads, gene annotations, polymorphisms and mobile elements) overlap or are associated

with each other allowing for the characterisation of results and the assessment of the biological impact. Browser Extensible Data (BED) and General Feature Format (GFF) file formats are commonly used to represent genomic features and are used for comparison in BEDTools [327].

The HERV-K insertion positions from the VCF output files for all patients obtained from ERVcaller were combined into a single BED file. This BED file was constructed in columns containing the information on chromosome, start and end position and whether the insertion was found on the forward or reverse strand of the DNA. A second BED file was also required for comparison to the human reference library of genes. A reference GFF (general feature format) file was downloaded and converted to BED format for this purpose. This BED file thus contained columns for chromosome, start and end position of genes, gene names and strand direction for all reference genes of the human genome. Using the *bedtools closest()* function these two BED files were searched against each other for overlapping features. In the event that no feature from the query BED file overlapped the current feature in the reference BED file, *bedtools closest()* reports the nearest feature to the given location (the least genomic distance from the start or end of the query feature). Using the arguments *-D* (report closest feature) with *-iu* (ignore upstream, report downstream) and *-id* (ignore downstream, report upstream), we were able to identify the closest upstream and downstream genes to the HERV-K insertion positions respectively [328].

4.6 Results

4.6.1 Preliminary Sample Findings (Patients 547, 569 and 607)

The VCF output files obtained from the three pilot study WGS data processed through the ERVcaller pipeline was analysed to identify the different TE's present and to gain a better understanding of the ratio in which these elements occur. Figure 4.5 demonstrates that HERV-K insertional elements were vastly fewer in both tumour and normal samples of the three patients when compared to other unfixed TE elements (SVA, ALU and LINE1).

Tumour samples consistently displayed higher numbers of all TE insertions when compared with their respective normal samples. Although HERV-K insertions were drastically lower than the other TE elements, this trend was also observed for HERV-K insertions, suggesting possible translocations and/or duplications.

Figure 4.6 indicates the number of HERV-K insertions per chromosome. Tumour sample T547 showed insertions on chromosomes 1, 8, 12, 14 and 18 that were not present in the paired normal sample, N547. Sample T569 showed insertions on chromosomes 4 and 19 that were not detected in the paired normal samples, N569, and, sample T607 showed insertions on chromosomes 12 and 19 that were not present in the paired normal sample, N607.

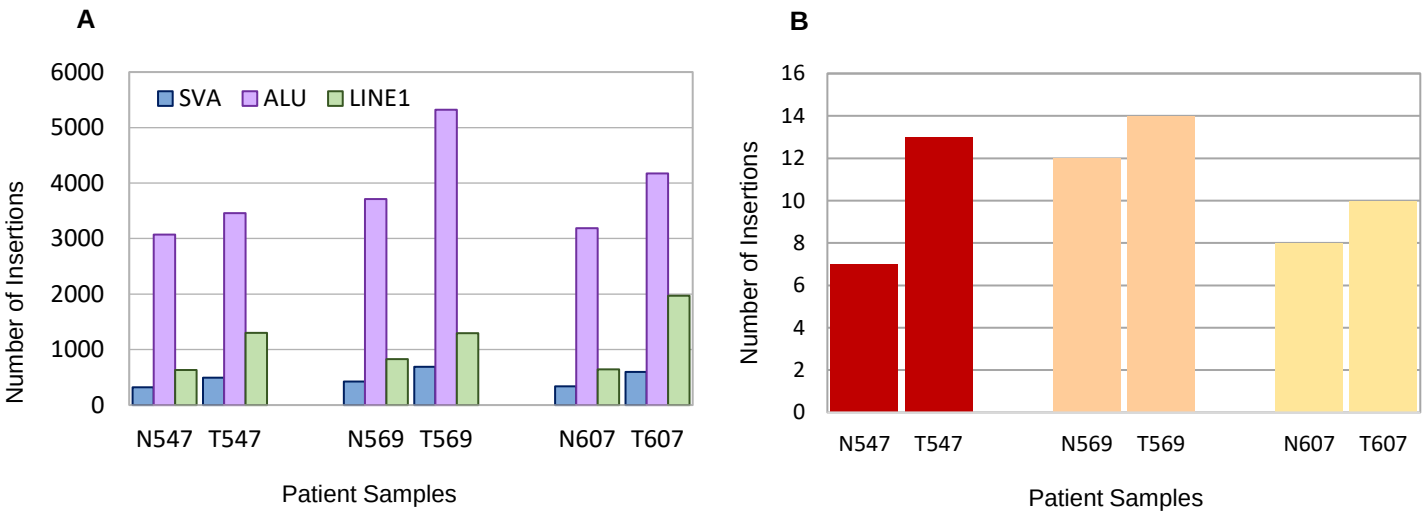


Figure 4.5: **A**) Visual representation of the number of SVA, ALU and LINE1 insertions per patient, illustrating the differences between tumour and normal samples for each patient. **B**) The number of HERV-K insertions per patient, showing the differences between tumour and normal.

A BED file of this output data was then created and run through BEDTools [327] to identify the closest proximity genes to these HERV-K insertions. Table 4.2 shows all the upstream and downstream genes that were detected per insertion per patient, and Table 4.3 provides a condensed list of all the common genes.

Figure 4.6: Number of HERV-K insertions detected per chromosome for **A**) tumour and paired **B**) normal pilot samples. Differences are observed on Chromosomes 1, 8, 12, 14 and 18 for samples T547 compared to N547. Insertions were detected on chromosome 4 and 19 in sample T569 compared to N569, and sample T607 shows more insertions on chromosomes 12 and 19 that are not present in N607.

Table 4.2: Upstream and Downstream genes detected relative to all HERV-K insertions positions for all tumour and normal samples. Entries highlighted in grey appear in the tumour sample only and not in their matched normal sample pair.

T457

Chromosome	Upstream gene	Insertion Position	Downstream gene
1	*CHIAP1*	111259975	*CHIAP2*
4	*SNRPCP13*	9601620	*ENPP7P11*
5	*LINC02063*	4537495	*LOC107986400*
8	*COL22A1*	139463347	*KCNK9*
10	*HPSE2*	99256367	*CNNM1*
12	*ZNF970P*	37739005	*AK6P2*
12	*ZNF970P*	37738034	*AK6P2*
15	*TPM1*	63082400	*LOC107984798*
18	*LOC100128360*	2000824	*LOC105371957*
18	*LOC105372045*	30137898	*MIR302F*

N547

Chromosome	Upstream gene	Insertion Position	Downstream gene
4	*SNRPCP13*	9601620	*ENPP7P11*
5	*LINC02063*	4537495	*LOC107986400*
10	*HPSE2*	99256367	*CNNM1*
12	*ZNF970P*	37738045	*AK6P2*
15	*TPM1*	63082403	*LOC107984798*

T569

Chromosome	Upstream gene	Insertion Position	Downstream gene
1	*NVL*	224340389	*CNIH4*
4	*SNRPCP13*	9601523	*ENPP7P11*
7	*WDR60*	158980700	*LINC00689*
8	*COL22A1*	139463347	*KCNK9*
10	*HPSE2*	99256367	*CNNM1*
12	*NTF3*	5537508	*ANO2*
12	*ZNF970P*	37738973	*AK6P2*
12	*ZNF970P*	37738034	*AK6P2*
19	*LINC00665*	36332606	*ZFP14*

N569

Chromosome	Upstream gene	Insertion Position	Downstream gene
1	*CHIAP1*	111259975	*CHIAP2*
7	*WDR60*	158980700	*LINC00689*
8	*COL22A1*	139463347	*KCNK9*
10	*HPSE2*	99256367	*CNNM1*
12	*NTF3*	5537542	*ANO2*
12	*ZNF970P*	37738978	*AK6P2*
12	*ZNF970P*	37738034	*AK6P2*

T607

Chromosome	Upstream gene	Insertion Position	Downstream gene
1	*CDK4P1*	105473257	*LOC105378881*
4	*SNRPCP13*	9601620	*ENPP7P11*
10	*HPSE2*	99256368	*CNNM1*
12	*ZNF970P*	37738992	*AK6P2*
12	*ZNF970P*	37738043	*AK6P2*
19	*LINC00665*	36332758	*ZFP14*

N607

Chromosome	Upstream gene	Insertion Position	Downstream gene
1	*CDK4P1*	105473257	*LOC105378881*
4	*SNRPCP13*	9601620	*ENPP7P11*
10	*HPSE2*	99256368	*CNNM1*
12	*ZNF970P*	37738992	*AK6P2*

Table 4.3: List of common upstream and downstream genes relative to the HERV-K insertions detected across tumour and normal samples.

Upstream Genes	Downstream Genes
CDK4P1	*AK6P2*
CHIAP1	*ANO2*
COL22A1	*CHIAP2*
HPSE2	*CNIH4*
LINC00665	*CNNM1*
LINC02063	*ENPP7P11*
NTF3	*KCNK9*
NVL	*LINC00689*
SNRPCP13	*MIR302F*
TPM1	*ZFP14*
WDR60	
ZNF970P	

(Respective up- and downstream search code can be found at https://github.com/VictoriaPatten/phd-scripts/tree/main/ERVcaller/ERVcaller_downstream_bedtools.sh and https://github.com/VictoriaPatten/phd-scripts/tree/main/ERVcaller/ERVcaller_upstream_bedtools.sh).

Aside from the upstream and downstream proximity genes, the data also showed that three insertions occurred within genes in both tumour and normal samples (Table 4.4). All three patients presented an HERV-K insertion within the *ABCC2* gene in both their tumour and normal samples. Patients 569 and 607 showed an insertion in *TMED2-DT* in tumour and normal samples, and patient 547 showed an insertion within *WDHD1* in both tumour and normal samples. These genes are all protein coding and *WDHD1* has been reported as having a role in the occurrence of OSCC [329,330].

Table 4.4: HERV-K insertions detected within genes for both tumour and normal samples.

Patients (T and N)	HERV-K Insertion Position	Inside Gene
547/569/607	99827974	*ABCC2*
569/607	123581935	*TMED2-DT*
547	55024644	*WDHD1*

The next step was to investigate the role and/or presence of HERV-K in the study cohort of the paired tumour-normal samples sequenced at the Wellcome Sanger Institute.

4.6.2 Main Patient Cohort Results

Once the WGS data of the main sample cohort had been received at the SANBI servers and transferred to Ilifu for processing, all tumour and blood DNA sequences were run through the ERVcaller pipeline followed by the BEDTools genomic arithmetic software. A similar table of results was obtained indicating the closest proximity upstream and downstream genes to all of the HERV-K insertions, and a list of common genes most likely to be affected was obtained.

Table 4.5 shows this list of upstream and downstream genes relative to the detected HERV-K insertions as well as genes where the insertions were reported to be within the gene. *TPM1, TMEM117* and *LINC00559* are the top three genes featured in most patients with HERV-K insertions. Furthermore, HERV-K insertions were detected within the *TMEM-117* and *LINC00559* genes. . The differences between tumour and normal samples are further visually represented in Figure 4.7.

The number of patients with HERV-K insertions proximal to *TPM1, TMEM117* and *LINC00559* are all higher in the normal sample versus the paired tumour samples. The same is true for *CHIAP1* and *CHIAP2* as well. In fact discrepancies exist between all tumour and normal samples for all genes identified suggesting that HERV-K insertions are not fixed and may present some degree of movement and translocation within the patients.

TPM1 is a member of the tropomyosin family of actin-binding proteins involved in muscular (striated and smooth) contraction and its downregulation has previously been reported to play an influencing role in the OSCC tumorigenesis [331]. While *TMEM-117* and *LINC00559* have not previously been identified as having a role in OSCC, the methylation and downregulation of *TMEM-117* has been associated with tumorigenesis and metastasis of pancreatic cancer, and breast cancer [332]. Mao *et al* recently described that *LINC00559,* a non-protein coding RNA, promotes cancer cell, growth along with colony formation and cell-cycle progression in animal and *in vitro* studies [333].

Table 4.5: Proximal genes detected relative to HERV-K insertion positions.

		Number of patients with HERV-K Insertion	
Gene	**Proximity to HERV-K insertion**	**Tumour**	**Normal**
TPM1	Upstream	24	26
TMEM117	Inside gene	23	24
LINC00559	Inside gene	20	21
CHIAP1	Upstream	15	18
CHIAP2	Downstream	15	18
ANO2	Inside gene	13	12
TMED2-DT	Inside gene	7	8
ZNF419	Downstream	6	8
WDHD1	Inside gene	5	8
ZNF528	Upstream	4	3
DPPA5P1	Downstream	4	3
NVL	Upstream	3	7
CCDC185	Upstream	3	2
SCGB1D1	Upstream	3	1
CNIH4	Downstream	3	7
CAPN8	Downstream	3	2
SCGB2A1	Downstream	3	1
MINCR	Upstream	2	1
ZNF16	Inside gene	2	1
ZNF696	Downstream	2	1
DUX4L2	Upstream	1	2
SUSD2	Upstream	1	2
DUX4	Downstream	1	2
GGT5	Downstream	1	2

This list of genes identified through ERVcaller and BEDTools analysis of WGS data will be further described in Chapter 5 with regards to the presence and potential implications of any somatic mutations that might have arisen.

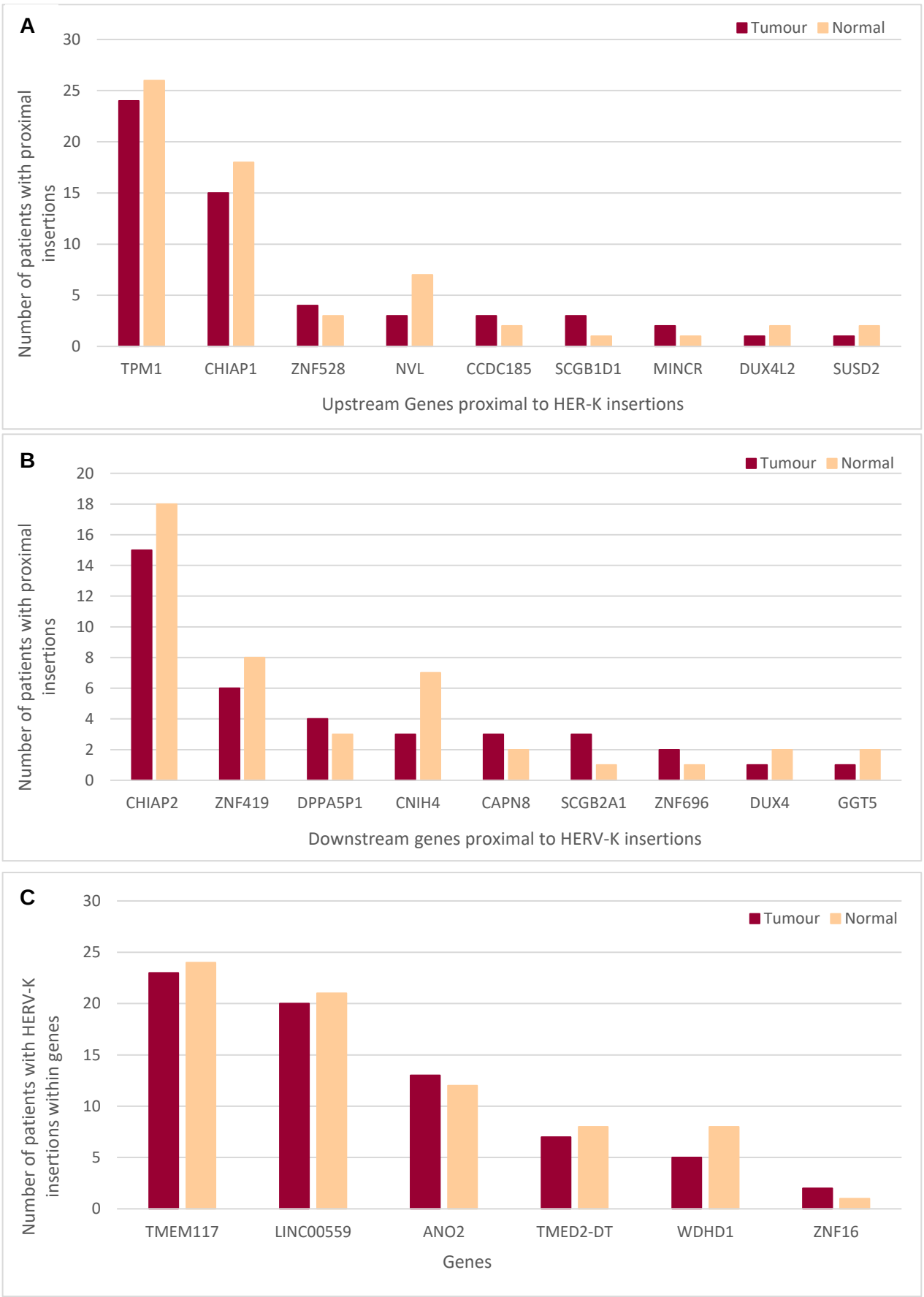


Figure 4.7: Number of patients with HERV-K insertions proximal to these **A**) upstream genes, **B**) downstream genes or **C**) insertions within genes. Tumour samples are represented in red and their paired normal samples are represented in peach.

4.7 Discussion

HERV-K insertion analysis allowed us to identify where HERV-K TE's were integrated in the genomes of the patients in this study, and to compare the matching tumour, normal and blood pairs to illustrate insertion differences indicating the likely transposition of these elements. Both the preliminary analysis and the analysis using the full sample cohort showed a trend whereby tumour samples had increased numbers of insertions possibly suggesting that in the tumour samples, the HERV-K elements are more active leading to regions of duplication and replicates, which is in keeping with literature reports alluding to elevated expression of HERV elements found in various cancers [99,105,106]. It is still unsure whether these insertions that were detected in this study might induce dysregulation of the host genomes and lead to chromosome instability and abnormalities. At this stage, we hypothesise that these HERV-K insertions and translocations might influence the proximal genes and further investigations may elucidate the role they might play. It is well known that TE's play an essential role in the maintenance of genomic stability, chromosomal architecture and transcriptional regulation, and the dysregulation of these elements has been reported in different cancer types [334]. The results obtained in this chapter indicate an increase in TE's (HERV-K) in the some of the tumour samples of the cohort, thus we postulate that this increase might be linked to regulation of the genome through influencing gene expression. Integration of TE's into existing genes can lead to the production of chimeric proteins while increased insertions into intronic regions could possibly interfere with transcription and gene stability [334–336]. It is likely that, due to the fact that TE's frequently translocate within the genome, an increase in the number of TE's could play a role in the development or progression of cancer through their activation and influence on chromosomal architecture and transcriptional regulation. These highly mutagenic elements are commonly associated with multiple steps of cancer development and progression although the exact mechanisms of action remain largely unexplained [334].

From the genes identified by the bioinformatics pipeline, a number of upstream and downstream proximal genes are shown in Table 4.7, along with genes where HERV-K insertions were found to be inserted within the gene. A number of these genes were also detected in the preliminary analysis using the three pilot pairs sequenced by the New York genome Centre. *ANO2, CHIAP1, CHIAP2, CNIH4, NVL, TMED2-DT* and *TPM1* all appeared in both analyses. Discrepancies between tumour and normal observations suggest that these HERV-K integrated sequences are not fixed in the genomes of these patients and

may possibly move around. This is in agreement with literature descriptions of this family of HERV's (discussed in section 1.2.3.3).

Some of the identified genes have also previously been reported to be involved in OSCC. *GGT5* (Y. Wang et al., 2022), *CHIAP2* [338], *SCGB2A1* [339], *SUSD2* [340], *ZNF419* [341], and *WDHD1* [329,330] have all been reported to show aberrant expression levels in OSCC and suggested to play an influencing role in tumorigenesis.

Further investigations of this study focuses on whether any of these genes identified as proximal to HERV-K insertions showed any somatic mutations in the tumour samples that could be linked to the insertions. Discussed in Chapter 5, bioinformatics analysis was performed on the WGS data using specialised variant calling software to examine the presence of somatic mutations.

Chapter 5: Identification of Somatic Mutations

5.1 Introduction

In the early twentieth century, chromosomal abnormalities were observed as one of the first links between cancer and genetic mutations [342–344]. After the discovery of DNA and the description of its structure, it was purported that chemical carcinogens and agents causing DNA damage that results in mutations leading to the mechanistic causation of cancer through the expansion of a single abnormal cell [345,346]. Conclusive evidence of this was reported in studies where fragments of cancer cell DNA were introduced into normal cells leading to transformation and malignancy [347,348]. This important and pivotal discovery launched the extensive research endeavour still ongoing today, to search for and identify abnormal genes promoting the development of human cancers .

A widely accepted understanding of the progression of cancer still remains the description of cancer as a Darwinian evolutionary mechanism whereby phenotypically normal cells acquire the hallmarks of cancer through positive selection for survival along with the development of somatic mutations [271,349]. Throughout the duration of a lifetime, all individuals are prone to the accumulation of spontaneously occurring mutations in somatic cells. Most of these are harmless, but occasionally, some will lead to phenotypic consequences affecting genes or regulatory elements, conferring a selective advantage to the cell, promoting growth and survival [350]. As links between somatic mutations and cancer have been established over the years, studies have led to the discovery of oncogenes whereby genetic mutations result in a gain of function leading to the progression and transformation of cancer cells. Parallel to this, studies on hereditary cancers have identified tumour suppressor genes typically inactivated by mutations [350,351]. The term ‘driver mutation’ has been coined to describe such mutations under positive selection, while ‘passenger mutation’ is used for variants that bring about no advantageous biological effect to the cell [152].

Recently the development and utilisation of high-throughput DNA sequencing has transformed our understanding of cancer genetics by facilitating the complete sequencing of more than 2500 whole cancer genomes and 10 000 cancer exomes. This has resulted in the identification of genes not previously recognised as having a role in various cancers [350].

This chapter describes an investigation into somatic mutations in a cohort of South African patients diagnosed with oesophageal cancer, via bioinformatics analysis of whole genome sequence (WGS) data of matched tumour-blood DNA. Using raw WGS data a variant calling

pipeline was established to align patient genomes to a reference genome (GRCh38) to call for somatic variants in the tumour samples only. It was possible to identify all genes with high impact severity (consequence of mutation) mutations as well as to confirm the mutations previously identified by the Wellcome Sanger Institute. The established pipeline was also utilised to investigate whether ERVcaller genes identified in Chapter 4 were associated with any somatic mutations.

Variant calling for somatic mutations was performed and identified a high level of mutations in the *MUC3A* gene, not previously associated with OSCC. However, we were unable to verify these mutations through PCR in the laboratory and suspected they were false positives. When re-evaluating the bioinformatics data using a panel of normal (PON) approach and a different variant caller, none of the initial mutations were confirmed thus we concluded they were indeed false positives. However, the re-analysis identified similar high numbers of different mutations in the same gene although due to time constrains, no laboratory confirmation was performed on these. We remain concerned that these mutations are also false positives as it is perplexing that no previous OSCC studies have reported mutations in this gene.

5.2 Bioinformatics Pipeline Set-up

5.2.1 Whole Genome Sequencing Data

As described in section 2.2.2, DNA from thirty-five blood samples and their paired tumour biopsies from South African patients underwent WGS and limited analysis at the Wellcome Sanger Institute in the UK. They confirmed a number of known mutated genes that had previously been reported to be associated with OSCC (Table 5.1).

Table 5.1: Known mutated genes associated with OSCC identified by bioinformatics analysis performed at the Wellcome Sanger institute on the WGS data from this study. This analysis was performed using Caveman and Pindel software [302,303]

AJUBA	*CREBBP*	*FAT1*	*KMT2C*	*NOTCH2*	*RB1*
BRCA1	*CUL3*	*FAT2*	*KMT2D*	*NOTCH3*	*TGFBR2*
BRCA2	*EGRF*	*FAT4*	*LRP1B*	*PIK3CA*	*TP53*
CCND1	*EP300*	*FBXW7*	*NFE2L2*	*PREX2*	*TERT*
CDKN2A	*ERBB4*	*KDM6A*	*NOTCH1*	*PTCH1*	

This information was used as a reference point for further analyses performed by the student, described in Figure 5.1 below.

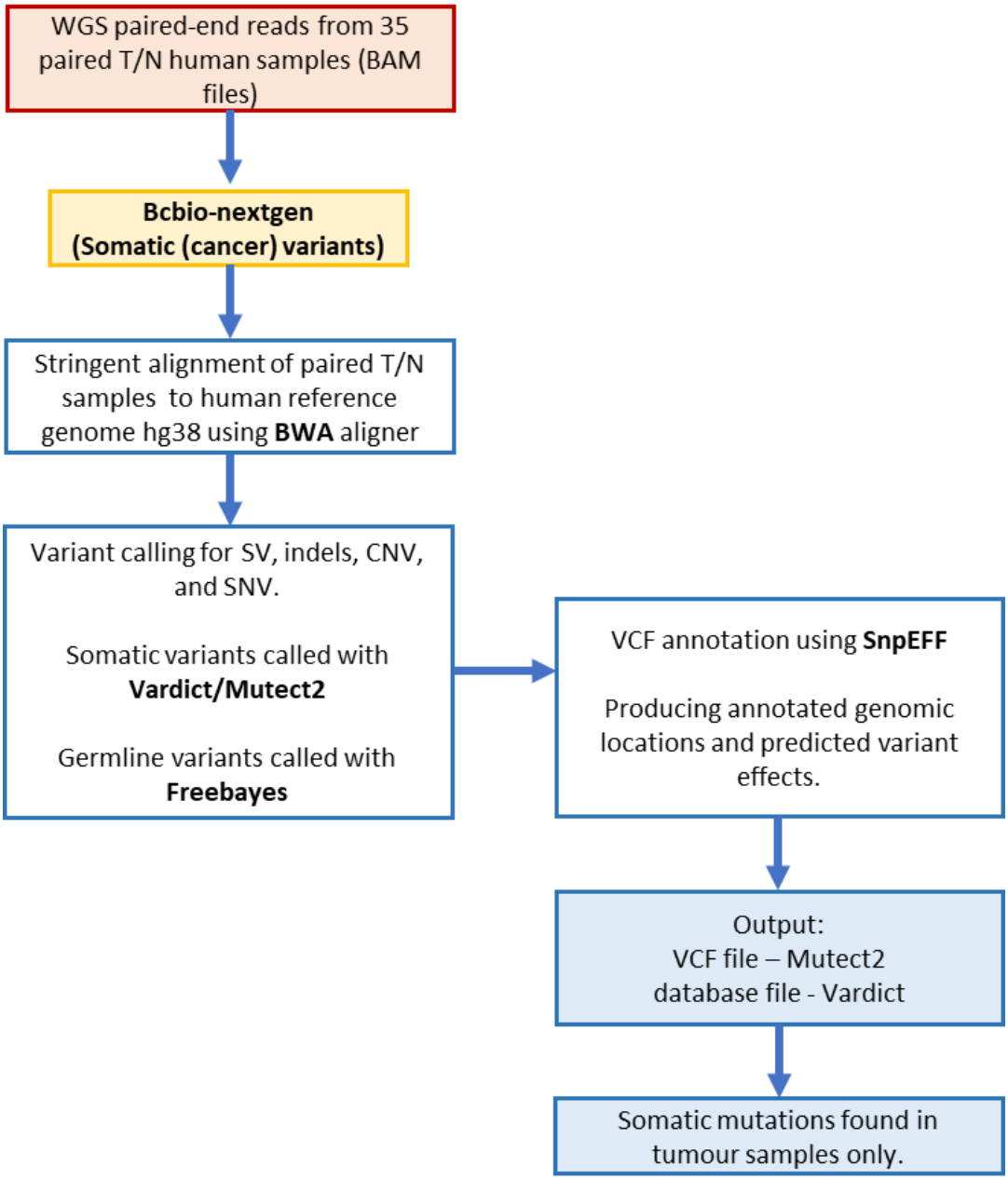


Figure 5.1: The bioinformatics workflow to detect somatic mutations using bcbio-nextgen. WGS = Whole Genome Sequencing, T/N = tumour and matched normal, SV = structural variants, CNV = copy number variants; SNV = single nucleotide variants, indels = insertions and deletions, VCF = variant call format.

5.2.2 Bcbio-nextgen Pipeline

Setting up a variant calling pipeline for small variants including SNV and indels, was the first step in the analysis process. It was decided to utilise an open source software package called bcbio-nextgen, a community developed "python toolkit providing best-practice pipelines for fully automated high throughput sequencing analysis" [352]. This software allows for the analysis of sequences through specialised pipelines with further visualisation and additional processing made possible. The variant calling analysis pipeline aligns reads to selected reference genomes, allowing for the identification of variants within the query sequences [353]. In this way, calls were compared against the common reference genome GRCh38 and multiple approaches for alignment, preparation and variant calling were incorporated into the pipeline to ensure an unbiased comparison of algorithms [352].

Bcbio-nextgen software (version 1.2.9) was downloaded and installed on the Ilifu server along with the necessary dependencies, aligners and the human reference genome GRCh38. Once the set up was confirmed, configuration files were constructed for each patient in the sample cohort following guidelines described in the software documentation [354].

All pipeline scripts used and query searches can be found in the online repository at https://github.com/VictoriaPatten/phd-scripts/tree/main/bcbio-nextgen.

5.2.3 Configuration Parameters

Bcbio-nextgen software is suitable for variant calling of small variants (SNP's) and indels, and supports the feature of tumour-normal paired calling. For tumour-normal samples, bcbio-nextgen is able to call both somatic (tumour-specific) and germline (pre-existing) variants (Chapman et al., 2021) using the parameters below:

BWA Aligner

The workflow involved in establishing this pipeline includes the setup of configuration files as YAML templates (Figure 5.1) which for each patient specify particular parameters and tools needed for the analysis. Raw patient BAM files were specified as the input files for each patient's configuration file, and reads were aligned to GRCh38 human reference genome using the Burrows-Wheeler Aligner (BWA 0.7.17) which maps low-divergent sequences against a large reference genome [355].

To option somatic and germline calls, individual variant callers were specified for each. Bcbio-nextgen carried out a single alignment for the normal sample first and then splits at the variant calling stage using the normal sample as a baseline for germline and somatic calling.

Mark Duplicates

In WGS it is a common best practice to mark or remove duplicate aligned reads as not doing so would lead to variant calling bias with incorrect results. Patient BAM files were prepared, aligned to GRCh37 at the Wellcome Sanger Institute, and duplicates were removed. However, in the bcbio-nextgen pipeline we realigned the BAM files to the latest, most up to date version of the reference genome (GRCh38) and thus the parameter *mark_duplicates* was set to 'TRUE' in the configuration files in order to mark duplicates again. Without this parameter, one would run the risk of preferentially amplified areas of the genome sequence being over-represented. Having these duplicates marked or removed increases the quality and reliability of the areas covered in sequencing [356].

Freebayes

Multiple different callers are supported for variant calling and in this instance germline SNV's and indels were called using Freebayes v1.3.6 [357], a genetic variant detector designed to locate small polymorphisms, specifically SNP's and indels. Freebayes is haplotype-based and calls variants based on the literal sequences of reads aligned to a particular target. The most likely combination of genotypes for the population at each position in the reference is determined through the short read alignments, and polymorphic positions are reported in variant calling file (VCF) format [357].

Vardict

Somatic calling uses the normal samples as a background to subtract existing (germline) calls, and in this instance, somatic SNPs and indels were called using Vardict [358], an ultra-sensitive variant caller that simultaneously calls SNV's and indels, performing local realignments for more accurate allele frequency estimation. Vardict further performs amplicon bias aware calling, rescue of long indels and improved scalability. Several downstream strategies are also employed to filter variants for the most likely cancer driving events [358].

SnpEff

VCF annotation was performed using the SnpEff tool [359]. The effects of variants in a genome sequence are rapidly categorised and annotated based on their genomic locations and prediction of coding effects. Using the VCF file generated by Vardict as input, SnpEff analyses and annotates the input variants and calculates the effects they produce on known genes. Annotated genomic locations include intronic and exonic regions, untranslated regions, intergenic regions, splice sites as well as upstream and downstream regions. The coding effects predicted include synonymous and non-synonymous amino acid replacements, start or stop codon gains or losses and frameshift variants [359].

Lumpy

Structural and copy number variants (CNV's) were called using Lumpy v0.3.1 [360], a probabilistic prediction framework for structural variant discovery. Multiple classes of evidence from multiple sources are accommodated in the same analysis run with multiple structural variant signals jointly integrated across a number of samples.

5.3 Pipeline Output Analysis and Results

When executing the pipeline and running the sample configuration files through the bcbio-nextgen, the alignment, mark duplicates, sorting and indexing of the BAM files is performed through a shell pipeline of BWA, samblaster and samtools [361,362]. The built-in multi-threading capabilities of BWA allow for the parallelized alignment of the BAM files. This is then followed by further parallelization of variant calling enabled by the simultaneous processing of partitioned genome regions that are bounded by spans of the genome where no callable reads in any of the samples that are being processed are contained [361].

```
- algorithm:
    aligner: bwa
    mark_duplicates: true
    remove_lcr: true
    tools_on: gemini
    variantcaller:
      somatic: vardict
      germline: freebayes
    svcaller: lumpy
    effects: snpeff
    effects_transcripts: canonical_cancer
    vcfanno: [gemini,somatic]
  analysis: variant2
  description: PD50649-normal
  files: ../input/HUMAN_1000Genomes_hs37d5_genomic_PD50649b.dupmarked.bam
  genome_build: hg38
  metadata:
    batch: bcbio
    phenotype: normal
- algorithm:
    aligner: bwa
    mark_duplicates: true
    remove_lcr: true
    tools_on: gemini
    variantcaller:
      somatic: vardict
      germline: freebayes
    svcaller: lumpy
    effects: snpeff
    effects_transcripts: canonical_cancer
    vcfanno: [gemini,somatic]
  analysis: variant2
  description: PD50649-tumor
  files: ../input/HUMAN_1000Genomes_hs37d5_genomic_PD50649a.dupmarked.bam
  genome_build: hg38
  metadata:
    batch: bcbio
    phenotype: tumor
fc_date: '2022-03-14'
fc_name: PD50649_original
upload:
  dir: ../final
resources:
  default:
    memory: 8G
    cores: 8
    jvm_opts: ["-Xms4G", "-Xmx8G"]
```

Figure 5.2: A prepared YAML configuration file for bcbio-nextgen somatic (cancer) variant calling. Both the tumour and normal BAM files for each patient were used as input. The aligner was set to BWA; mark_duplicates set to 'true'; Vardict and Freebayes are specified as somatic and germline variant callers respectively; lumpy was set as the structural variant caller (svcaller); and gemini and somatic vcf annotation is set to use SnpEff. https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbio-nextgen/bcbio_config.yaml

5.3.1 GEMINI Output

Annotation of the output was performed using the SnpEff tool [359]. GEMINI v.0.20.1 [363], was then used to create a database of the output to facilitate the query of the annotated vcf files. GEMINI is a genome mining tool for exploring human variations. This tool integrates genetic variation with a diverse and adaptable set of genome annotations into a unified database to facilitate interpretation and data exploration [363], thus when executing the bcbio-nextgen pipeline with the parameter *vcfanno: [gemini, somatic]* (Figure 5.2), a GEMINI database containing somatic variant tables is created for downstream query and analysis. The GEMINI tool provides a flexible database-driven Application Programming Interface (API) for the purposes of data visualisation and exploration. Accurately predicting the functional consequences and effects of the detected variants using GEMINI is however dependent on external tools, either SnpEff or VEP [364]. For this analysis, SnpEff was utilised and indicated in the YAML file as *effects: snpeff*. The documentation description states that "SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effect of genetic variants" [365].

One is able to pose intricate questions about one's data due to the expressive power of Structured Query Language (SQL), and the convenience of having all genetic variants stored in a database to facilitate variant interpretation [363]. Using the command line, the GEMINI output database files were explored and filtered to search for particular parameters of interest annotated within the variant/variant_impacts tables of the output database files. Variant consequence columns such as gene, impact and impact_severity (Table 5.2) were filtered out for specific variant searches. The impact_severity rankings (HIGH, MED(medium) or LOW) for each GEMINI impact term can be found in Appendix 2. The impact severity can be described as the functional consequence of a given variant [364].

GEMINI incorporates specific tools for the command line for the purpose of querying the output database files. This tools is called '*gemini query'* and below is an indication of how it was used to interact with the database files to filter for all HIGH impact severity mutations (Figure 5.3).

Table 5.2: Annotations of variant consequence columns found in GEMINI variant/variant_impacts tables within the output database files [364].

Variant Consequence	Description
anno_id	Annotation ID
gene	Gene name
transcript	Transcript name
exon	Exon number
is_exonic_is_lof	Loss of function in exon
is_coding	Is a coding region
condon_change	Change of codon
aa_change	Amino acid change
aa_length	Amino acid length
biotype	Biotype variant
impact	Impact of variant
impact_so	Sequence ontology for impact
impact_severity	Impact severity
polyphen_pred	Polymorphism phenotyping score
polyphen_score	Polymorphism phenotyping prediction
sift_pred	SIFT prediction
sift_score	SIFT score

```
gemini query --header --show-samples -q "select variant_id, chrom, start, end, gene, ref, alt, impact,
impact_severity from variants where impact_severity='HIGH'" PD39445-bcbio-vardict.db
```

Figure 5.3: *gemini query* command line to filter the database file of patient PD39445 for chromosome, start and end positions, gene, reference and alternate alleles, impact and impact severity of variants, where the impact severity is given as HIGH. https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbio-nextgen/HIGH_impact_gemini_search.sh

5.3.2 SnpSift Filtering

To ensure that the *gemini query* tool provided correct descriptions of variants from the database files, a second filtering tool was applied to all the DNA sequences to confirm the variants identified.

Following execution of the bcbio-nextgen pipelines, the SnpEff tool further generated variant identification output in VCF files for each patient, in addition to the GEMINI databases that were created. These VCF files could not be explored using *gemini query*, but instead required the use of SnpSift, a toolbox joined to the SnpEff package and allowing one to filter and manipulate VCF files in order to identify interesting and relevant variants [365].

Using SnpEff and SnpSift together, SNV's, indels and structural variants were identified with a simple assessment of the putative impact of each variant given as HIGH, MODERATE and LOW, in contrast to GEMINI'S HIGH, MED, LOW. Filtering of the VCF files was implemented using the *SnpSift filter* and *SnpSift extractFields* commands. Figure 5.4 shows

an example of filtering for all HIGH impact mutations on chromosome 1, extracting the fields of the gene, position, reference allele and alternate allele columns of the VCF file.

```
SnpSift filter "( CHROM = 'chr1' ) && ( ANN[*].IMPACT = 'HIGH')" bcbio-vardict-annotated.vcf |  SnpSift
extractFields - CHROM ANN[*].GENE POS REF ALT
```

Figure 5.4: Command line for *SnpSift filter* and *SnpSift extractFields* commands. The output of *SnpSift filter* is piped as input to *SnpSift extractFields.*

5.3.3 Search for Genes with Somatic Variants

5.3.3.1 Genes Identified in ERVCaller

Chapter 4, identified the locations of HERV's and possible translocations within the genomes of thirty patients. Output from this software generated a list of possible genes where HERV's are integrated (Table 5.3).

Using the GEMINI database files generated from bcbio-nextgen analysis, we endeavoured to identify whether any somatic mutations might exist within or near to the genes present on this list that may possibly link the HERV insertions to somatic mutations. The *gemini query* tool was used to search for the genes shown in Table 5.3. Figure 5.5 shows this *gemini query* command line search.

Table 5.3: Genes identified by ERVcaller with HERV-K insertions (Chapter 4) in the genomes of thirty patients.

TPM1	*WDHD1*	*MINCR*	*DPPA5P1*
TMEM117	*NVL*	*ZNF16*	*CAPN8*
LINC00559	*ZNF528*	*SCGB1D1*	*DUX4*
CHIAP1	*CCDC185*	*CHIAP2*	*GGT5*
ANO2	*DUX4L2*	*ZNF419*	*ZNF696*
TMED2-DT	*SUSD2*	*CNIH4*	*SCGB2A1*

```
gemini query --header --show-samples -q "select variant_id, chrom, start, end, gene, ref, alt, impact_so,
impact_severity from variants where gene in
('TPM1','TMEM117','LINC00559','CHIAP1','ANO2','TMED2DT','WDHD1','NVL','ZNF528','CCDC185','DUX4L2','SUSD2','MINCR
','ZNF16','SCGB1D1','CHIAP2','ZNF419','CNIH4','DPPA5P1','CAPN8','DUX4','GGT5','ZNF696','SCGB2A1') and call_rate
>=0.95" PD39445_bcbio_vardict.db
```

Figure 5.5: *gemini query* command line to filter the database file of patient PD39445 for chromosome, start and end positions, gene, reference and alternate alleles, impact and impact severity of variants, where the genes are specified from the given list of genes identified in ERVcaller. https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbio-nextgen/erv_gemini_search.sh

5.3.3.2 OSCC Associated Genes

Preliminary analysis performed at the Wellcome Sanger Institute of the first set of samples identified somatic variants in a list of genes previously reported to be associated with OSCC (Table 5.1) using Caveman and Pindel algorithms [302,303]. Our own analysis included use of the *gemini query* command (Figure 5.6) to filter for HIGH impact variants in these genes and use this as a further confirmation step (in addition to the two different filtering methods - GEMINI and SnpSift) that the bcbio-nextgen analysis pipeline that was set up was indeed accurate.

```
gemini query --header --show-samples -q "select variant_id, chrom, start, end, gene, ref, alt, impact_so,
impact_severity from variants where gene in
('TERT','AJUBA','BRCA1','BRCA2','CCND1','CDKN2A','CREBBP','CUL3','EGRF','EP300','ERBB4','FAT1','FAT2','FAT4','FB
XW7','KDM6A','KMT2C','KMT2D','LRP1B','NFE2L2','NOTCH1','NOTCH2','NOTCH3','PIK3CA','PREX2','PTCH1','RB1','TGFBR2'
,'TP53') and call_rate >= 0.95" PD39445_bcbio_vardict.db
```

Figure 5.6: *gemini query* command line to filter the database file of patient PD39445 for chromosome, start and end positions, gene, reference and alternate alleles, impact and impact severity of variants, where the genes specified are from the given list of genes identified with HIGH impact mutations by the analysis carried out by the Wellcome Sanger Institute. https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbio-nextgen/sanger_gemini_search.sh

5.3.3.3 All Genes with HIGH Impact Severity Variants

Once we had confirmed the previously identified genes using *gemini query*, a further search for all genes presenting HIGH impact variants was performed across all patient genomes. This was to determine the top genes with the highest number of HIGH impact mutations within the patient cohort. This was carried out on all thirty-five patients. Figure 5.3 above shows the *gemini query* command used to filter database files for all genes with HIGH impact severity mutations.

5.4 Gene Search

5.4.1 ERVcaller Gene Search

The search for somatic variants within genes containing HERV insertions as identified by ERVcaller software did not confirm the hypothesis that HERV insertion was associated with contiguous high impact mutations. The objective had been to link these HERV insertions sites to possible somatic mutations within or near to the genes of interest. Table 5.4 indicates that of the initial thirty patients sequenced, only four patients presented with MED (medium) impact severity mutations.

Table 5.4: GEMNI Search for variants within genes identified by ERVcaller as having HERV insertions

Patient Number	Chromosome	Gene	Ref	Alt	Type	Severity
PD39448	Chr22	*SUSD2*	C	T	missense variant	MED
PD39454	Chr4	*DUX4*	G	A	missense variant	MED
	Chr13	*LINC00559*	T	A	splice region variant	MED
	Chr14	*WDHD1*	A	G	splice region variant	MED
PD44699	Chr1	*CAPN8*	G	A	Missense variant	MED
	Chr1	*CAPN8*	A	AG	splice region variant	MED
PD44702	Chr15	*TPM1*	T	G	Missense variant	MED

*MED=Medium impact severity

Patient PD39448 had a MED severity missense mutation on chromosome 22 in gene SUSD2, a possible cytokine receptor or tumour suppressor involved in breast tumorigenesis [366]. Patient PD39454 had three MED severity mutations, one on chromosome 4 in gene *DUX4* (missense variant), one on chromosome 13, in gene *LINC00559* (splice region variant), and one on chromosome 14 in gene *WDHD1* (splice region variant). *DUX4* is a transcription factor in embryos and involved in muscular dystrophy [367]. *LINC00559* is a non-protein coding RNA expressed in the testis [368], and *WDHD1* is a DNA replication initiation factor [369]. Patient PD44699 has two MED impact severity mutations both on chromosome 1 in the *CAPN8* gene (missense variant and splice region variant respectively). *CAPN8* is involved in membrane trafficking in gastric pit cells and mucus cells in the stomach [370]. The fourth patient to present a MED impact severity missense variant was PD44702 on chromosome 15 in the *TPM1* gene which is involved in binding to actin filaments in muscle and non-muscle cells as well as associating with the troponin complex during striated muscle contraction [371].

With only four of thirty patients presenting somatic variants, and MED impact severity variants at that, it can be deduced that there is very little linkage between HERV insertions and somatic mutations. No HIGH impact severity variants were detected, further suggesting that there is no significant effect on the genes or nearby genes where the portions of HERV genomic DNA appear to be inserted. This does not support the hypothesis that HERV insertions or translocations could be linked to somatic mutations possibly as either a causal effect or a consequence.

Due to this finding, further investigations into the effects of HERV insertions were not explored and the focus was on investigating somatic mutations in the sample cohort.

5.4.2 Identification of Somatic Mutations

The list of genes shown in Table 5.1 was further used to confirm the bcbio-nextgen database output of our own analysis to prove that the pipeline was functional and accurate in its alignment, mapping and variant database generation. Table 5.5 indicates the number of patients with either or both MED/HIGH impact severity variants.

Table 5.5: bcbio-nextgen and GEMINI search of genes previously identified to be altered, yielding the indicated mutations in column 3. Column 2 lists the number of patients displaying the mutations identified. MED indicates medium impact severity mutations.

Genes Investigated	**Total number of patients with mutations**	**Number and Type of Variant**
AJUBA	4	2HIGH / 3MED
BRCA1	2	4MED
BRCA2	3	1HIGH / 2MED
CCND1	0	
CDKN2A	10	7HIGH / 3MED
CREBBP	0	
CUL3	0	
EGRF	0	
EP300	2	1HIGH / MED
ERBB4	2	2MED
FAT1	3	1HIGH / 9MED
FAT2	5	2HIGH / 4MED
FAT4	3	4MED
FBXW7	3	2HIGH / 1MED
KDM6A	1	1HIGH
KMT2C	3	1HIGH / 3MED
KMT2D	8	8HIGH / 2MED
LRP1B	2	2HIGH
NFE2L2	3	3MED
NOTCH1	10	3HIGH / 7MED
NOTCH2	0	
NOTCH3	3	3HIGH / 3MED
PIK3CA	3	3MED
PREX2	2	3MED
PTCH1	2	1HIGH / MED
RB1	1	1MED
TGFBR2	2	2MED
TP53	26	20HIGH / 7MED
TERT	1	1MED

From table 5.5, it can be seen that 24 genes that were previously identified in the Wellcome Sanger Institute's analysis had either or both MED/HIGH impact severity mutations when processed though our own bcbio-nextgen pipeline. This provides some confirmation that the results we obtained are reliable. As anticipated, *TP53* showed the highest number of variants with 20 patients presenting HIGH impact severity variants and 7 patients presenting MED impact severity variants. *CDKN2A* was another expected result with 10 patients

presenting 7 HIGH and 3 MED impact severity mutations. Both of these genes have been reported in the literature as being associated with a myriad of cancers, and more specifically, OSCC [6,30]. A recent systemic review and meta-analysis by Naseri *et al* (2021), showed that changes in the *TP53, CCND1, MDM2, NOTCH1/2/3, KMT2D, CDKN2A, PIK3CA* and *FAT1* genes were detected in more than 10% of OSCC patients in studies predominantly from China and Japan with a few studies coming from Brazil, Korea and Iran [30]. While the patient cohort in this study comprises only South Africans, it is encouraging to confirm that most of these commonly mutated genes in OSCC were identified as mutated in this patient cohort.

These findings indicate that the analysis pipeline of the WGS data that was performed in this study provided results that were in accordance with those reported in the literature.

5.4.3 HIGH Impact Severity Variant Gene Search

With the knowledge and confidence that our pipeline and results were accurate, a further search of the GEMINI database files was conducted using the *gemini query* command line (Figure 5.2) to filter and extract information on all the genes present in the sample cohort presenting with HIGH impact severity mutations.

Table 5.6 shows the results of this search indicating the number of patients with HIGH impact severity mutations in a given gene, thus providing a ranking of genes with the highest number of patients having mutations for each gene. Figure 5.7 shows the total number of variants detected across the patient cohort. For example, 18 out of 35 patients had HIGH impact mutations in the *TP53* gene, while 19 separate incidences were detected, suggesting that one patient had more than one mutation in the gene.

The results depicted in this table indicated a surprising finding. At the top of the list with 30 out of 35 patients presenting HIGH impact mutations, is the *MUC3A* gene. Furthermore, 258 incidences of mutations were detected across the patient cohort. These numbers far exceed those for genes such as *TP53, CDKN2A* and *KMT2D*. Table 5.7 gives a brief description of each gene and whether it has previously been reported to play a role in oesophageal cancer. 86% of this South African cohort display HIGH impact severity mutations in *MUC3A*, a gene not previously reported as being associated with OSCC. Figure 5.8 provides a breakdown of the number of *MUC3A* mutations detected per patient in the cohort. These findings prompted a number of questions into what could possibly be happening in this gene with such a significantly high number of mutations. What is the *MUC3A* gene responsible for and how is this affecting these South African OSCC patients?

Table 5.6: Top 20 genes with the most HIGH impact severity variants detected in the full thirty-five patient cohort.

Gene	Number of patients with mutations
MUC3A	30
TP53	18
HEATR9	14
VPS52	12
NCOR1	10
AHNAK	9
OR4D10	8
CDKN2A	8
FIGN	7
OR4D11	7
KMT2D	7
CST3	6
GRIN2A	6
DEF8	6
FGFR1	6
TBL1X	6
FAM135A	6
TDG	6
CST5	5
SLC4A3	5

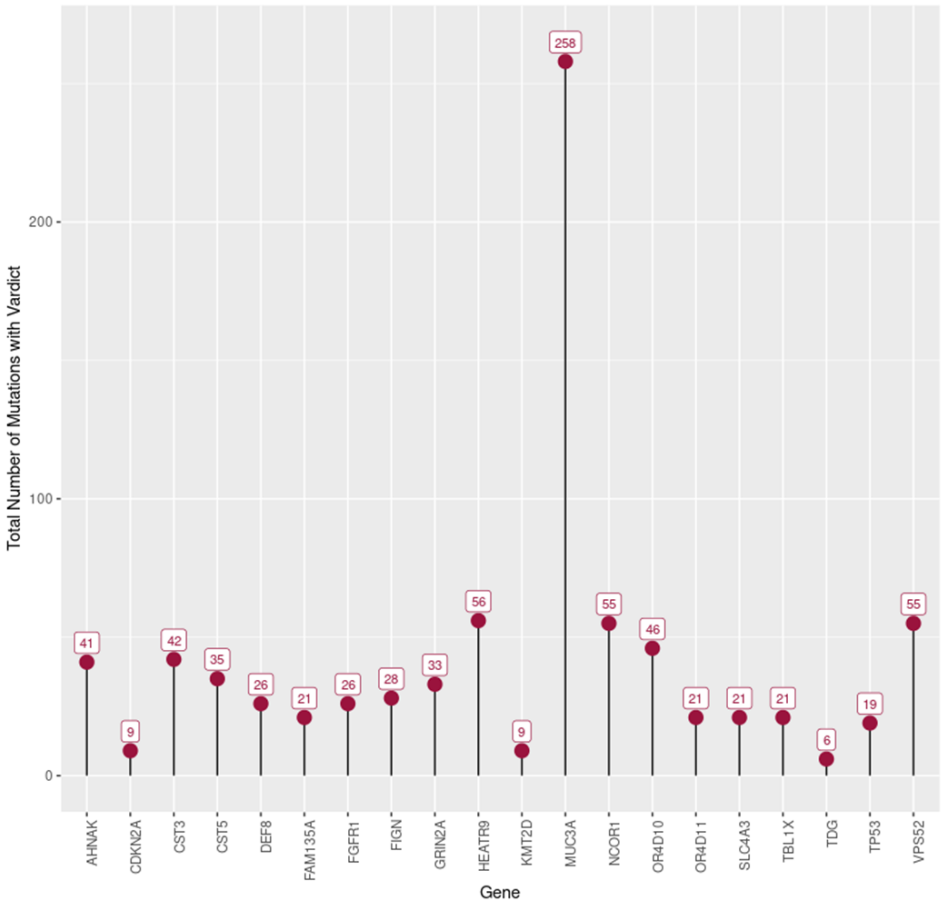


Figure 5.7: Lollipop plot indicating the distribution of somatic variants across the top 20 genes with the most HIGH impact severity variants detected using Vardict variant caller with bcbio-nextgen pipeline. The *MUC3A* gene shows a vastly greater total number of mutations detected across to whole cohort compared to all other genes.

Table 5.7: Description of the genes identified in Table 5.6 and whether they have previously been reported to be associated with oesophageal cancer.

Gene	Description	Reported role in Oesophageal Cancer
MUC3A	Epithelial glycoprotein involved in ligand binding and intracellular signalling [372]	-
TP53	Tumour suppressor protein involved in many cancers [373]	Yes [26,374,375]
HEATR9	Acts upstream of or within hematopoietic progenitor cell differentiation [376]	-
VPS52	Involved in vesicle trafficking from both early and late endosomes, back to the trans-Golgi network [377]	-
NCOR1	Mediates ligand-independent transcription repression of thyroid-hormone and retinoic-acid receptors by promoting chromatin condensation and preventing access of the transcription machinery [378]	-
AHNAK	May play a role in such diverse processes as blood-brain barrier formation, cell structure and migration, cardiac calcium channel regulation, and tumour metastasis [379]	Yes [380]
OR4D10	Interacts with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell [381]	-
CDKN2A	Tumour suppressor that induces cell cycle arrest in G1 and G2 phases [382]	Yes [26,383–385]
FIGN	Predicted to enable microtubule-severing ATPase activity, and to be involved in cytoplasmic microtubule organization [386]	-
OR4D11	Interacts with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell [387]	-
KMT2D	Methylates the Lys-4 position of histone H3 [388]	Yes [26]
CST3	Serves as a local regulator of cysteine proteinase activity [389]	Yes [390]
GRIN2A	Involved in long-term activity-dependent increase in the efficiency of synaptic transmission [391]	-
DEF8	Enables metal ion binding activity, and is involved in lysosome localization; positive regulation of bone resorption; and positive regulation of ruffle assembly [392]	-
FGFR1	Involved in intracellular tyrosine kinase activity and downstream signalling via PI3K and MAPK pathways [393]	Yes [394–396]
TBL1X	Plays a role in transcription activation mediated by nuclear receptors [397]	Yes [398]
FAM135A	Predicted to be involved in cellular lipid metabolic process [399]	-
TDG	Plays a key role in active DNA demethylation and removes thymine moieties from G/T mismatches [400]	Yes [70,401]
CST5	Plays a protective role against proteinases present in the oral cavity [402]	-
SLC4A3	Encodes a plasma membrane anion exchange protein [403]	-

MUC3A is a protein coding gene whose product is a highly *O*-glycosylated trans-membrane mucin expressed in various epithelial cells and located in the mucin cluster of chromosome 7q22 [256,266,404–406]. The extracellular domain of this transmembrane mucin would ordinarily function as a protection barrier, while the intracellular domain can be phosphorylated in turn activating various mucin-specific signal transduction pathways leading to the alteration of downstream cellular regulation processes such as inflammatory responses, epithelial cell adhesion, cellular differentiation and apoptosis [184]. In previous studies the aberrant expression of *MUC3A* has been associated with oncogenic profiles in breast, pancreatic, gastric, colorectal, prostate and renal cancers, eliciting poor prognosis in patients [257–259,405]. In normal epithelial cells the MUC3A protein plays an important role in the maintenance of the mucosal barrier function and prevention of invasion of pathogens at mucosal surfaces [184].

Further investigation into the *MUC3A* gene mutations were performed, with the intention to elucidate some understanding of a functional role in OSCC.

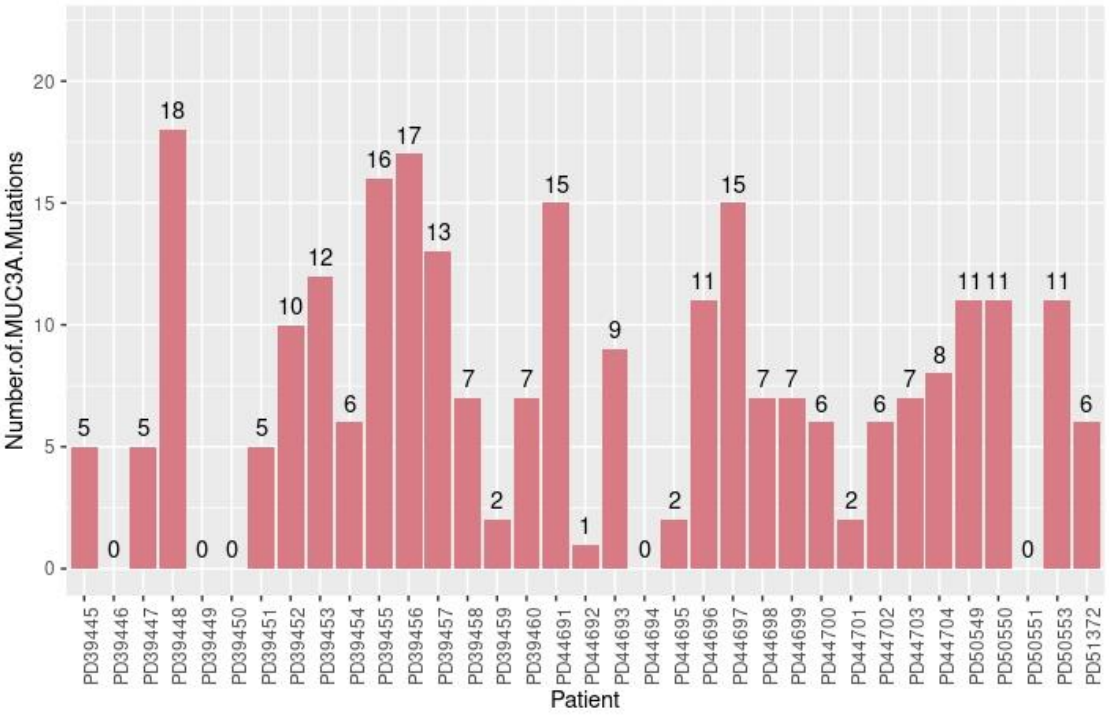


Figure 5.8: Bar graph of the number of *MUC3A* mutations detected per patient in the cohort (30 out of 35 patients). The number at the top of each bar indicates the number of mutations. https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbio-nextgen/MUC3A_gemini_search.sh

Analysis of all 258 individual mutations detected across the cohort showed that 95% of the mutations identified were frameshift variants caused by short indels. Only 14 out of the 258 variants were as stop codons. The full list of 258 variants detected are shown in Appendix 3. When sorting the variants by start and end position, it appeared that many patients shared a number of the same mutations and the positions identified were grouped together into five genomic regions, or clusters. This sorting is shown in Appendix 4. Cluster locations are indicated in Table 5.8. Using the NCBI resource Genome Data Viewer [407], the exact locations of these clusters all fall within the large second exon of the gene. The *MUC3A* gene consists of 12 exons with the largest being the second exon of 8805 base-pairs in length. Figure 5.9 shows the structure of the *MUC3A* gene with the mutation clusters indicated by vertical coloured lines.

Table 5.8: *MUC3A* mutations grouped into 5 cluster locations within exon 2 of chromosome 7.

	Start position	**End position**
Cluster 1	100953731	100953840
Cluster 2	100954996	100955249
Cluster 3	100957642	100958144
Cluster 4	100958183	100958658
Cluster 5	100958763	100959000

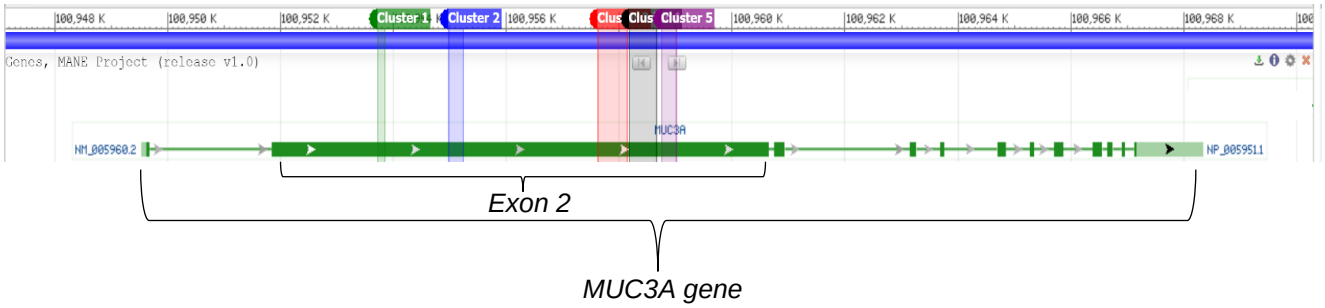


Figure 5.9: Visual representation taken from the NCBI Genome Data Viewer [407] depicting the *MUC3A* gene structure. The locations of the clusters of mutations detected through bioinformatics analysis are indicated in the second exon by coloured lines on the image. Cluster 1 = green, cluster 2 = blue, cluster 3 = red, cluster 4 = black and cluster 5 = purple.

It can clearly be seen from Figure 5.9 that clusters 3, 4 and 5 lie fairly close to each other, toward the second half of the second exon. Clusters 1 and 2 lie closer toward the beginning of exon 2 and are further apart from each other.

5.5 Laboratory Confirmation of Bioinformatics Data

At this point it was important to confirm the presence of these *MUC3A* mutations by PCR amplification and sequencing of patient DNA.

5.5.1 DNA Extraction for use in PCR

Genomic DNA was extracted from OSCC cell lines, and paired patient blood and tumour biopsies for PCR analysis.

Cell line DNA was used to optimise the PCR primers. Seven OSCC cell lines were grown in culture under standard conditions. WHCO 1, WHCO5 and WHCO 6 cells lines were originally established in South Africa from surgical biopsies of primary OSCC [295], while Kyse 30, Kyse 150, Kyse 180 and Kyse 450 cell lines were Japanese derived [296]. Extraction of DNA from cells was carried out using a standard phenol:chloroform:isoamylalcohol (25:24:1) protocol and eluted DNA was stored at -20°C until needed (see section 2.2.4.4).

DNA extraction from patient blood and biopsies were extracted in accordance with standard operating protocols as described in sections 2.2.1.2 and 2.2.1.3. Patient DNA was stored at -20°C until needed. DNA was thus available for PCR use from the 16 UCT patients that made up part of the patient sample cohort. DNA from the 14 WITS patients was not available in this laboratory due to shipping complications during the COVID-19 lockdown period.

5.5.2 PCR Primer Design

The clusters of mutations identified in exon 2 of the *MUC3A* gene served as guideline locations for the design of primers for PCR. Clusters 1, 2 and 5 where small enough in length to allow for a single set of primers each, while clusters 3 and 4 were too lengthy to allow for accurate amplification and post-PCR sequencing. For this reason, mid-way primers were designed for these two clusters.

The genomic *MUC3A* DNA sequence contains multiple tandem repeats, and exon 2 proved highly challenging for the design of robust and stable primers. The large number of multiple binding sites was a challenge as the repetitive nature of the DNA meant one or both of the primers in a pair would inevitably bind at more than one location. Eventually suitable primers were designed with the lowest possibility of multiple product lengths. Primer sequences for clusters 1-5 with their respective lengths, annealing temperatures, GC contents and the expected product lengths can be found in section 2.2.5.1, Table 2.7.

5.5.3 Primer Optimisation

The primers were prepared as 100µM stock solutions and stored at -20°C. 10µM working solutions were prepared from these stocks and used to prepare PCR reactions in order to avoid contamination of stocks. Primers for the individual five clusters were optimised separately using the DNA extracted from the OSCC cell lines. The integrity of this cellular DNA was confirmed by agarose gel electrophoresis as described in section 2.2.1.5.

A standard PCR protocol was used as a starting point for optimisation of each primer set, as described in section 2.2.5.2 [306], and thereafter, conditions were adjusted to obtain an optimal set of parameters for each primer pair. Reaction mixes were prepared in 25 µl volumes and thermocycling was carried out in an Applied Biosystems SimpliAmp thermocycler. Annealing temperatures were as per the manufacturer's specifications, and adjusted throughout the optimisation process. Primer sequences for Cluster 1 and Cluster 5 were easily optimised (Figure 5.10 **A** and **B**) and their specific cycling conditions are shown in Tables 5.9 and 5.10 respectively. Both of these primer sets amplified the cell line DNA at 60°C, cycling through forty repeats of the denaturation, annealing and elongation stages of the PCR.

Table 5.9: Optimised conditions for the primer set for Cluster 1 mutations.

Primer Pair	PCR Conditions		Product Size
Cluster 1 Forward:	95°C: 4 minutes		
5'- TAA GTA CAC TCA GCA CTC CTA -3'	95°C: 45 seconds		
	60°C: 45 seconds	40 cycles	889bp
Cluster 1 Reverse:	72°C: 1 minute		
5'- GAG ATC ATG GAT GTA GAA GTT ACC -3'	72°C: 7 minutes		
	4°C: 7 minutes		

Table 5.10: Optimised conditions of the primer set for Cluster 5 mutations.

Primer Pair	PCR Conditions		Product Size
Cluster 5 Forward:	95°C: 5minutes		
5'- ACC TCA CAT GAT ACT CCC -3'	95°C: 1 minute		
	60°C: 1 minute	40 cycles	752bp
Cluster 5 Reverse:	72°C: 1 minute		
5'- ATA TCA GTG GGT ATA GAG GGA AAG -3'	72°C: 7 minutes		
	4°C: 7 minutes		

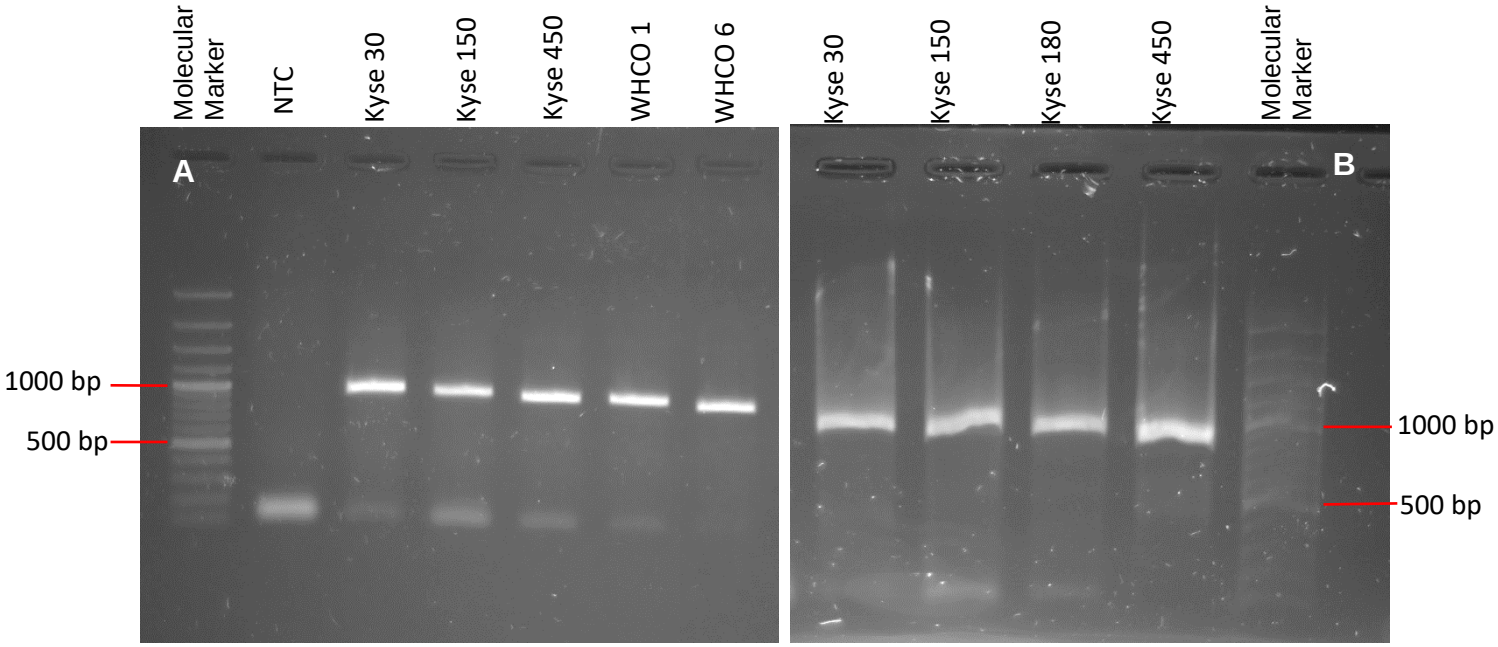


Figure 5.10: Optimisation of primers for **A**) Cluster 1 and **B**) Cluster 5 using several OSCC cell line DNAs. Both Primer sets were optimal at an annealing temperature of 60°C. PCR products were electrophoresed through a 1% agarose gel for 35 minutes at 100V and visualised under UV light for 20 seconds using Novel juice. Expected band sizes were observed for all samples for Cluster 1 (889 bp) and for Cluster 5 (752 bp).

The primer sets for Clusters 2, 3 and 4 (full product length primers as well as midway primers) were exceptionally difficult to optimise. Numerous adjustments to the protocols were made including magnesium titrations, primer concentration titrations, DNA concentration titrations and the addition of enhancers such as DMSO at varying concentrations, and a combination of Tris, KCl and Gelatine. Alternate PCR buffers and Taq polymerase were included in the reaction mix and multiple thermocycler machines were tested along with a myriad of different cycling profiles that were attempted. None of these yielded adequate template amplification.

In the absence of optimisation of clusters 2, 3 and 4, it was decided to proceed with PCR amplification of patient DNA using the two optimised primer sets for Clusters 1 and 5 while attempts at optimisation of the remaining primers continued in the background.

5.5.4 PCR of Patient DNA with Cluster 1 and 5 Primers

Using the optimised primer conditions for Clusters 1 and 5, PCR's were performed using genomic DNA previously extracted from patient tumour biopsies. The integrity of this DNA was verified by agarose gel electrophoresis as described in section 2.2.1.5. DNA was shown to be intact and suitable for PCR (Figure 5.11).

Table 5.11 below indicates the patients identified as having *MUC3A* mutations in clusters 1 and 5 (also shown in Appendix 3). Three patients were identified as having mutations in cluster 1 while four were identified in cluster 5. However, genomic DNA from only two UCT patients for cluster 1, and three patients for cluster 5 were available for analysis.

Table 5.11: Patients identified as having mutations falling into cluster 1 and cluster 5 *MUC3A* genomic locations. 'Institute' gives an indication of where the patients were recruited and biopsies collected.

	Patients	**Institute**
Cluster 1	PD39456	UCT/Groote Schuur
	PD39457	UCT/Groote Schuur
Cluster 5	PD39448	UCT/Groote Schuur
	PD39454	UCT/Groote Schuur
	PD39457	UCT/Groote Schuur

In all PCR reactions, a no-template control (NTC) was included. Figure 5.12 shows the PCR products from patients PD39456 and PD39457 respectively. The product size for this primer set is 889 bp according to the reference genome sequence, and sample amplification bands of the expected size were obtained.

PCR of cluster 5 primers with patient DNA had to be repeated a number of times to obtain useable PCR products for sequencing. A representative gel is shown in Figure 5.13 below.

The amplified PCR products obtained for cluster 5 primers appeared to be 'less clear' on the electrophoresis gels than those obtained for cluster 1 primers. The product size for this cluster was 752 bp according to the reference genome, and from Figure 5.13 the amplification bands visible on the gels appear to lie closer to the 1000 bp marker, suggesting that the PCR product may be slightly larger than anticipated. PCR products were subjected to bidirectional Sanger Sequencing as described in section 2.2.5.4.

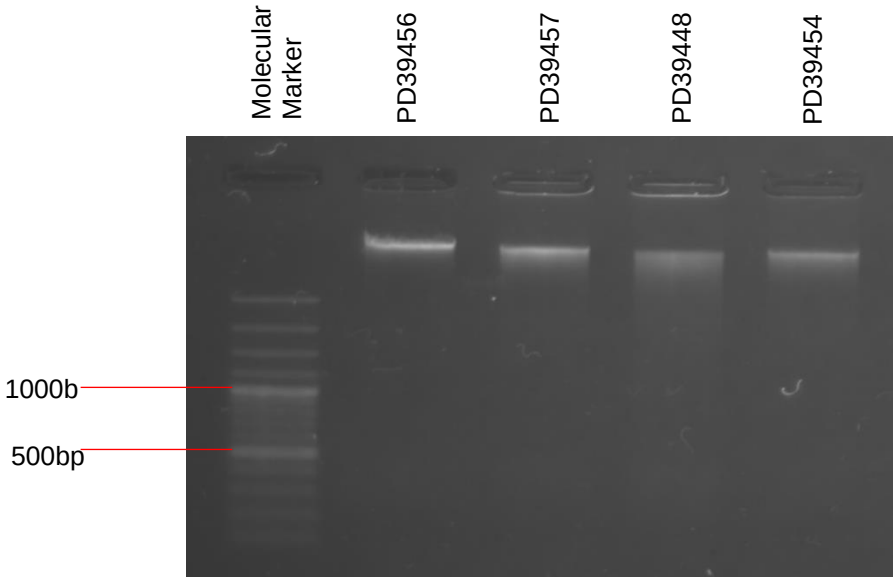


Figure 5.11: Visualisation of patient DNA electrophoresed through a 1% agarose gel for 35 minutes at 100V to verify the integrity of the DNA. The GeneLadder™ molecular marker is shown in lane 1 of the gel.

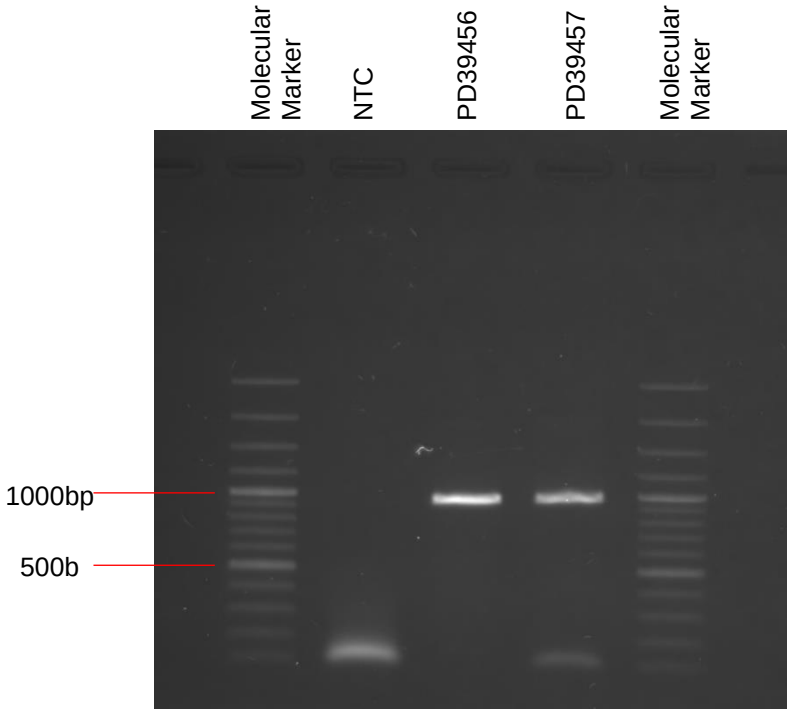


Figure 5.12: Visualisation of PCR products using primers for cluster 1 *MUC3A* mutations in patients PD39456 and PD39457 respectively. Post-PCR, products were electrophoresed through a 1% agarose gel for 35 minutes at 100V and visualised under UV light for 20 seconds using Novel juice. Expected amplification size are bands shown in the region of 889bp. NTC indicates no-template control.

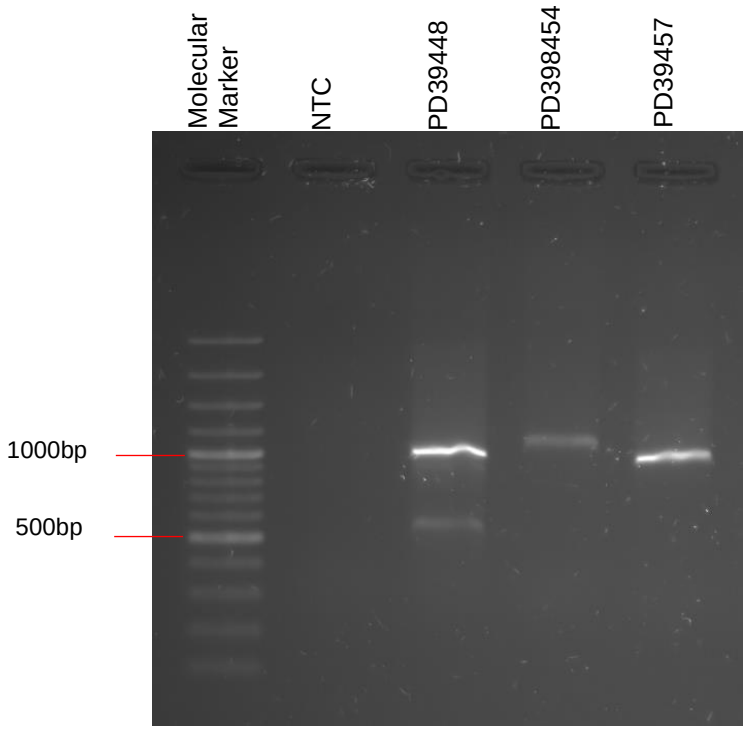


Figure 5.13: PCR using cluster 5 primers on patients PD39448, PD39454 and PD39457. All products were electrophoresed through 1% agarose gel for 35 minutes at 100 and visualised using Novel Juice, under UV light for 25 seconds. Amplification bands are visible slightly higher than the expected target size of 752bp. A non-specific amplification ban is visible for patient PD39448. NTC indicates no-template control.

5.5.5 Post-PCR DNA Sequencing Results

Cluster 1 Mutations in Patients:

The DNA sequence of the PCR products for patient PD39456 was very clear and easy to interpret (Figure 5.14 **B**). When analysing this chromatogram together with Table 5.12, it can be seen that the *MUC3A* variants identified through bioinformatics analysis in this patient, shown in the 'Alt' column in Table 5.12, could not be confirmed in the PCR product sequence. The variant positions have been indicated in the reference sequence and on the chromatogram to show where the alternate variants were expected, but the chromatogram sequences represented the reference sequence exactly (Figure 5.14 **A**), and none of the six variants were present in the patients.

Table 5.12: Cluster 1 variants identified in the *MUC3A* gene for patient PD39456 through bioinformatics analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
PD39456	chr7	100953731	*MUC3A*	CTC	T	Frameshift
	chr7	100953737	*MUC3A*	T	TGG	Frameshift
	chr7	100953752	*MUC3A*	T	TA	Frameshift
	chr7	100953758	*MUC3A*	TA	T	Frameshift
	chr7	100953774	*MUC3A*	AT	A	Frameshift
	chr7	100953777	*MUC3A*	C	CA	Frameshift

The clearly resolved peaks on the chromatogram show high confidence that the DNA sequence is accurate and that the PCR product matched the reference sequence.

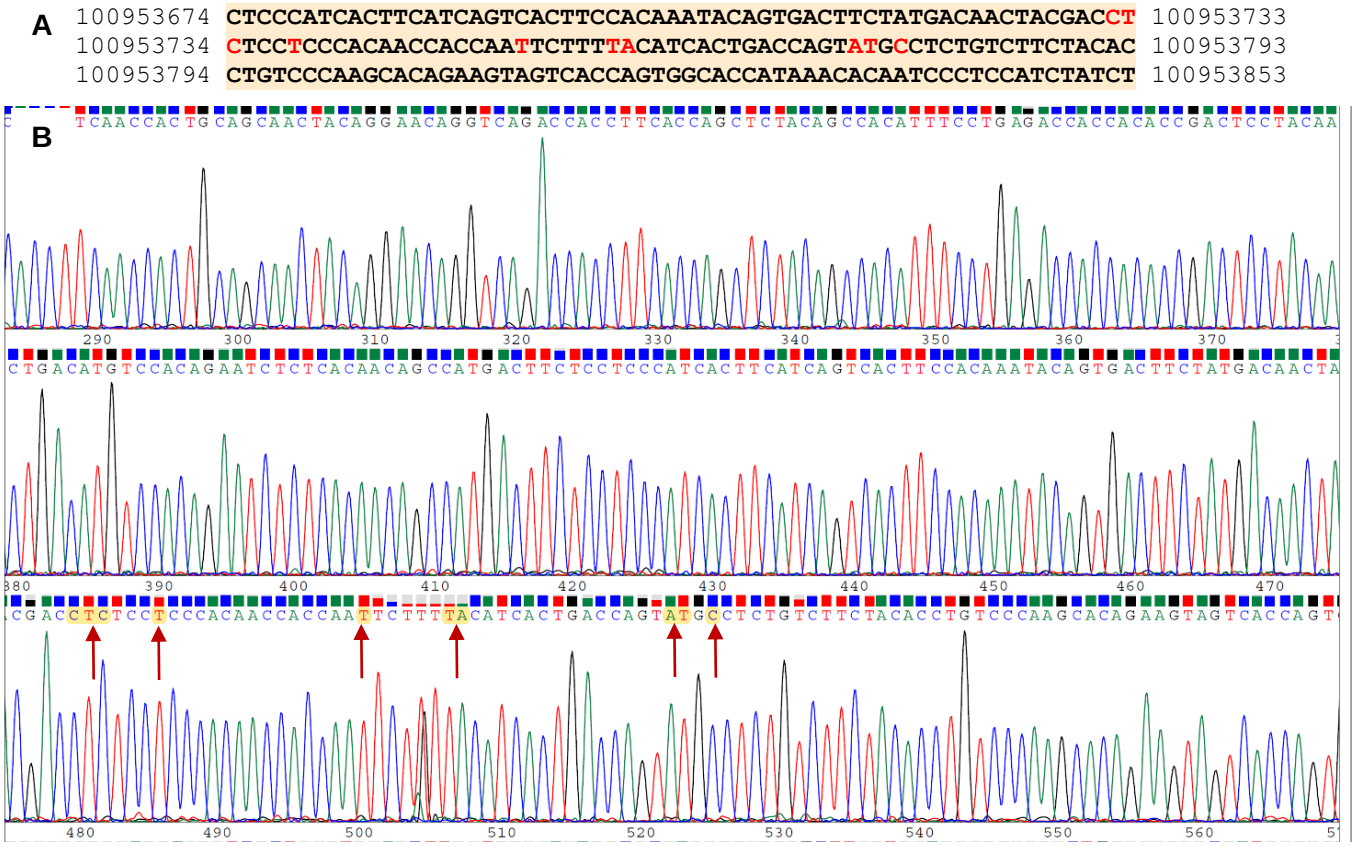


Figure 5.14: **A**) *MUC3A* reference sequence showing the locations of the cluster 1 mutations determined from bioinformatics data for patient PD39456 in red. **B**) Chromatogram of the PCR product sequence for patient PD39456 using the primer set for cluster 1 mutations. The annotated sequence is shown above each line of corresponding peaks and the nucleotides indicated by arrows represent the positions where the identified variants are located as identified in Table 5.12.

Patient PD39457 presented a further 6 *MUC3A* variants in the cluster 1 region, shown in Table 5.13 according to the bioinformatics analysis. The DNA sequence chromatogram obtained was again clear with well resolved peaks, and the positions for the expected variants are indicated on the chromatogram in yellow, but as can be seen when comparing the chromatogram sequence (Figure 5.15 **B**) to the *MUC3A* reference sequence (Figure 5.15 **A**), the variants once again are not present in the patients.

Table 5.13: Cluster 1 variants identified in the *MUC3A* gene for patient PD39457 through bioinformatics analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
PD39457	chr7	100953731	*MUC3A*	CTC	T	Frameshift
	chr7	100953737	*MUC3A*	T	TGG	Frameshift
	chr7	100953808	*MUC3A*	A	AG	Frameshift
	chr7	100953812	*MUC3A*	AG	A	Frameshift
	chr7	100953834	*MUC3A*	A	AT	Frameshift
	chr7	100953838	*MUC3A*	TC	T	Frameshift

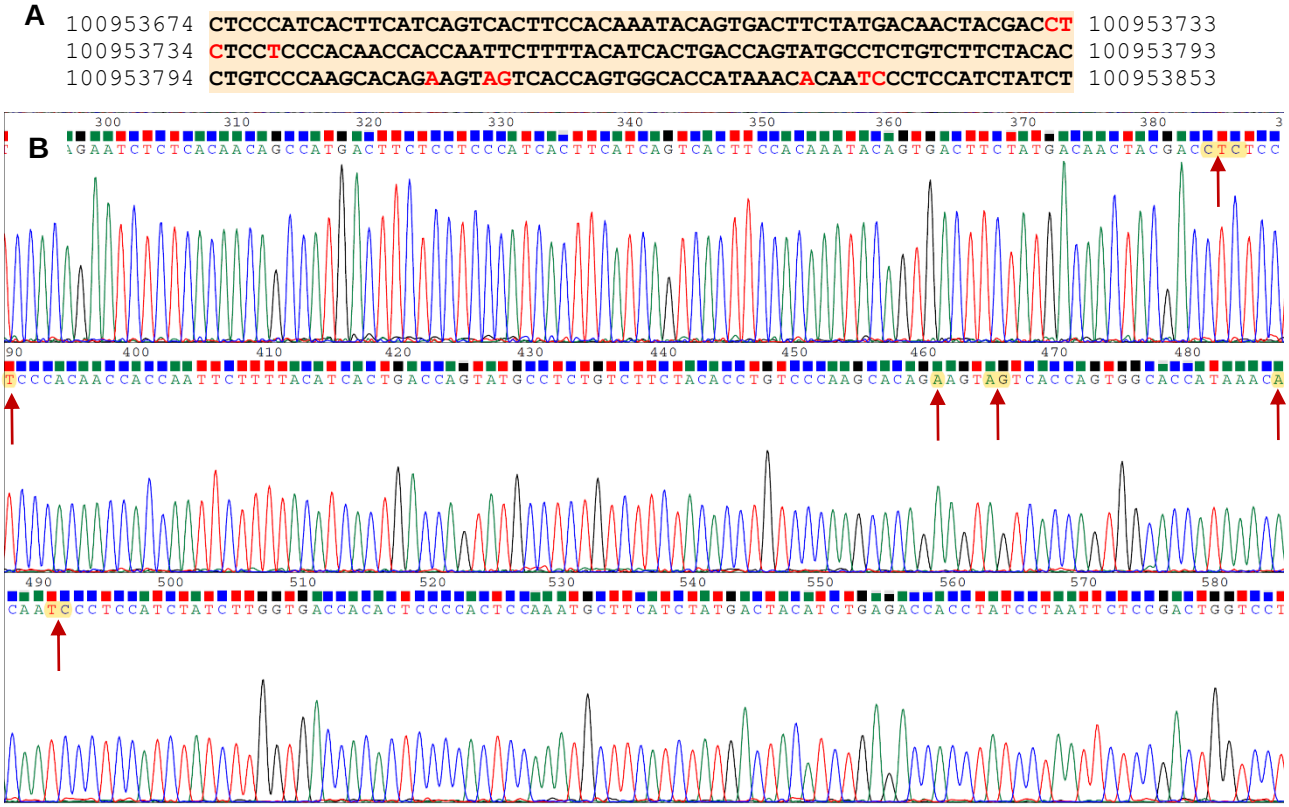


Figure 5.15: **A**) *MUC3A* reference sequence showing cluster 1 mutation locations for patient PD39457 in red. **B**) Chromatogram of forward direction Sanger Sequencing of the PCR product for patient PD39457 using the primer set for cluster 1 mutations. The annotated sequence is shown

above each line of corresponding peaks and the nucleotides indicated by arrows, indicate the positions of the variants identified by bioinformatics analysis of the WGS data.

Neither of the patients for the cluster 1 primer set showed the anticipated variants in their PCR products. This result was unexpected and perplexing, and we speculate that this could be due to low primer specificity and the high volume of tandem repeats in exon 2 of *MUC3A* influencing primer binding to a repeat region where no mutations are present.

Cluster 5 Mutations in Patients:

Sequencing of the PCR product from patient PD39448 provided a slightly less ideal chromatogram with a small degree of background noise (Figure 5.16). As with patient PD39456 and cluster 1, the *MUC3A* variants detected through bioinformatics analysis (Table 5.14) were not present in the resulting sequence chromatogram. The reference positions shown in Table 5.14 are indicated by the arrows on the chromatogram in Figure 5.16, and clearly match the reference genomic sequence. None of the alternate allele variants are present.

Table 5.14: Cluster 5 variants identified in the *MUC3A* gene for patient PD39448 through bioinformatics analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
PD39448	Chr7	100958806	*MUC3A*	C	CCAAGACCACCTC AACCAGTCCTCCCA GATTCACCT	Frameshift
	Chr7	100958808	*MUC3A*	T	TG	Frameshift
	Chr7	100958856	*MUC3A*	TC	T	Frameshift
	Chr7	100958858	*MUC3A*	T	TGG	Frameshift
	Chr7	100958859	*MUC3A*	TC	T	Frameshift

Because this sequence showed some degree of background noise, a portion of the sequence was analysed using the NCBI BLAST [284] to confirm the sequence alignment and identity. Figure 5.17 shows a screenshot of the BLAST results confirming that the sequence showed 100% query cover to *MUC3A* with 96% identities matched.

These sequencing and BLAST results suggest that the PCR product sequence obtained from patient PD39448 with cluster 5 primers matches the normal *MUC3A* genomic sequence exactly, and the variants identified through bioinformatics analysis are not present.

A

```
100958714 CCACCACGGAGACCACCTCACACAGTGCTCACAGCTTCACTTCTTCGATCACCACCACCG 100958773
100958774 AGACCACCTCACACAATACTCGCAGCTTCACTTCTTCGATCACCACCACCGAGACCAACT 100958833
100958834 CTCACAGTACTACCAGCTTCACTTCTTCGATCACCACCACCGAGACCACCTCACACAGTA 100958893
100958894 CTCCCAGCTTCAGTTCTTCAATCACCACCACTGAGACCCCCTTACACAGTACTCCTGGCC 100958953
100958954 TCACTTCGTGGGTCACCACCACCAAGACCACCTCACACATTACTCCTGGCCTCACTTCTT 100959013
```

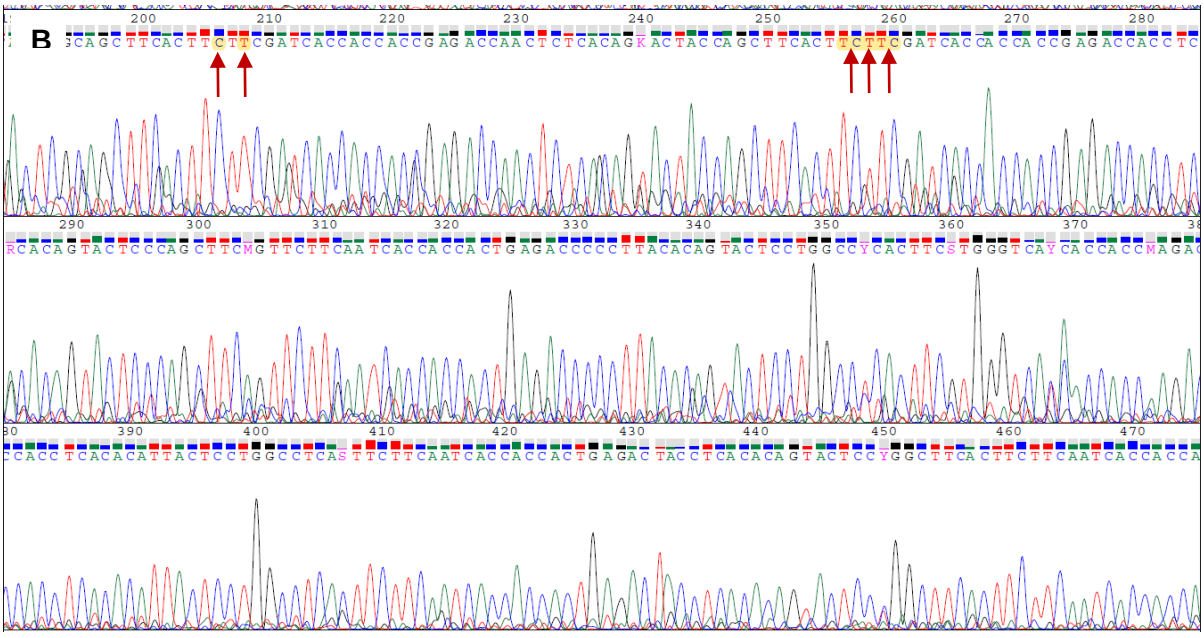


Figure 5.16: **A**) *MUC3A* reference sequence showing cluster 5 mutation locations for patient PD39448 in red. **B**) Chromatogram of forward direction Sanger Sequencing of the PCR product for patient PD39448 using the primer set for Cluster 5 mutations. The annotated sequence can be found above each line of corresponding peaks and the nucleotides indicated with arrows represent the positions where the identified variants should be located as identified in Table 5.14.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens mucin 3A, cell surface associated (MUC3A), mRNA	Homo sapiens	361	567	100%	2e-95	96.21%	11248	NM_005960.2

Homo sapiens mucin 3A, cell surface associated (MUC3A), mRNA

Sequence ID: NM_005960.2 **Length:** 11248 **Number of Matches:** 2

Range 1: 7181 to 7391 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
361 bits(195)	2e-95	203/211(96%)	0/211(0%)	Plus/Plus

```
Query  1     ACCACCGAGACCACCTCRCACAGTACTCCCAGCTTCMGTTCTTCAATCACCACCACTGAG  60
             |||||||||||||||| |||||||||||||||| |||||||||||||||||||||||||
Sbjct  7181  ACCACCGAGACCACCTCACACAGTACTCCCAGCTTCAGTTCTTCAATCACCACCACTGAG  7240

Query  61    ACCCCCTTACACAGTACTCCTGGCCYCACTTCSTGGGTCAYCACCACCMAGACCACCTCA  120
             |||||||||||||||||||||||||| |||||| ||||||| ||||||| ||||||||||
Sbjct  7241  ACCCCCTTACACAGTACTCCTGGCCTCACTTCGTGGGTCACCACCACCAAGACCACCTCA  7300

Query  121   CACATTACTCCTGGCCTCASTTCTTCAATCACCACCACTGAGACTACCTCACACAGTACT  180
             |||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||
Sbjct  7301  CACATTACTCCTGGCCTCACTTCTTCAATCACCACCACTGAGACTACCTCACACAGTACT  7360

Query  181   CCYGGCTTCACTTCTTCAATCACCACCACTG  211
             || ||||||||||||||||||||||||||||
Sbjct  7361  CCTGGCTTCACTTCTTCAATCACCACCACTG  7391
```

Figure 5.17: NCBI BLAST results of the sequence obtained from the PCR product of patient PD39448 with cluster 5 primers. 100% Query cover and 96% identities matched to the *MUC3A* genomic sequence. Max (maximum) score indicates the highest alignment score calculated from the sum of rewards for matched nucleotides and penalties for mismatches and gaps. Total Score provides the sum of alignment scores of all segments from the same subject sequence. Query Cover is the percentage of the query length that is included in the aligned segments. E-value denotes the

number of alignments expected by chance with the calculated score or better. And, Ident (identity) shows the highest percentage identity for a set of aligned segments to the same subject sequence.

Patient PD39454 had only two mutations according to the bioinformatics analysis, shown in Table 5.15. Figure 5.18 shows the *MUC3A* reference sequence (Figure 5.18 **A**) as well as the sequence chromatogram obtained by sequencing of the PCR products (Figure 5.18 **B**). This chromatogram sequence also matches the reference sequence exactly and does not indicate the presence of the variants detected in the bioinformatics study.

Table 5.15: Cluster 5 variants identified in the *MUC3A* gene for patient PD39454 through bioinformatics analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
	Chr7	100958995	*MUC3A*	CT	C	Frameshift
PD39454	Chr7	100958999	*MUC3A*	T	TC	Frameshift

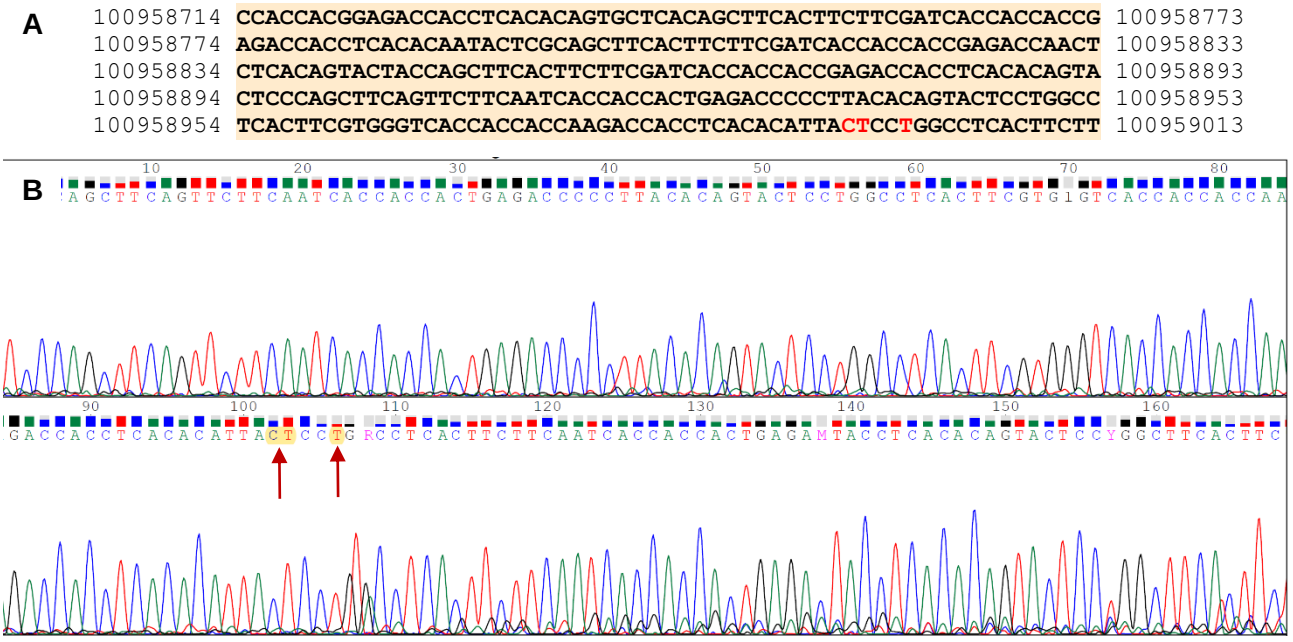


Figure 5.18: **A**) *MUC3A* reference sequence showing cluster 5 mutation locations for patient PD39454 in red. **B**) Chromatogram of forward direction Sanger Sequencing of the PCR product for patient PD39454 using the primer set for Cluster 5 mutations. The annotated sequence can be found above each line of corresponding peaks and the nucleotides indicated with arrows represent the positions where the identified variants should be located as identified in Table 5.15

Bioinformatics analysis of the WGS data of patient PD39457, had previously detected a single, large insertion (Table 5.16). The position for this expected variant is highlighted in

the *MUC3A* reference sequence (Figure 5.19 **A**) as well as on the chromatogram obtained from sequencing (Figure 5.19 **B**).

Table 5.16: Cluster 5 variants identified in the *MUC3A* gene for patient PD39457 through bioinformatics analysis. Positions indicated correspond to reference genome GRCh38, and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Chromosome	Position	Gene	Ref	Alt	Impact
PD39457	Chr7	100958763	*MUC3A*	A	ACCACCACCTGTTC ACTTCTTCTGTCGCC ACCATGGAGACCAC CTCACACAGTACTC CCAGCATCGCTACG TCAATCGCCACCAC TGAGATCATCTCAC ACAGCACTCCCAGC TATGCTTCTTCAA TTG	Frameshift

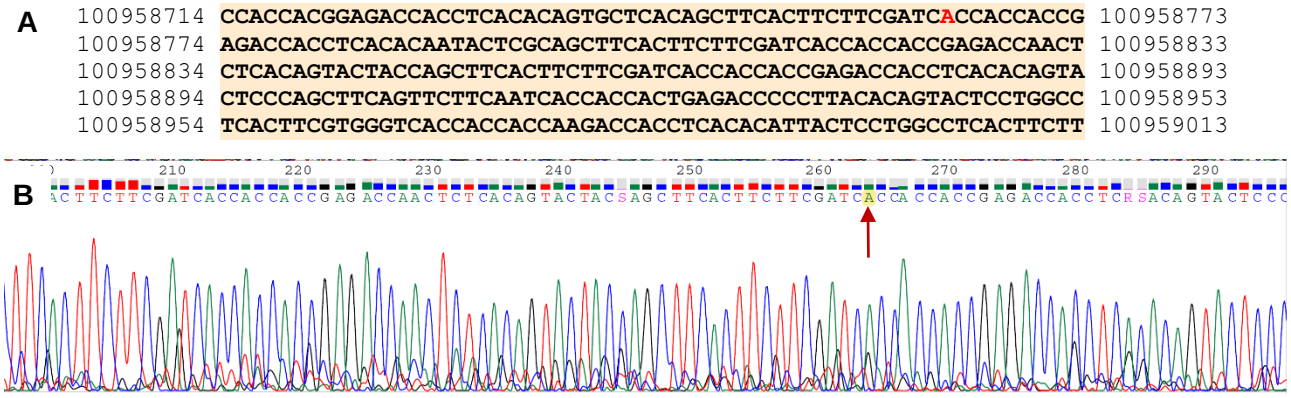


Figure 5.19: **A**) *MUC3A* reference sequence showing cluster 5 mutation locations for patient PD39457 in red. **B**) Chromatogram of forward direction Sanger Sequencing of the PCR product for patient PD39457 using the primer set for Cluster 5 mutations. The annotated sequence is shown above each line of corresponding peaks and the nucleotides indicated with an arrow represent the positions where the identified variant should be located as identified in Table 5.15

As with all four of the previous post-PCR sequencing, the sequence obtained for patient PD39457 with the cluster 5 primer set, did not confirm the large 129 base insertion that was expected based on the output obtained from the bioinformatics analysis performed on the samples. None of the bioinformatics variants identified were detected in these patients' post-PCR sequences, begging the question of, why and how?

Cluster 5 Screening:

As numerous OSCC biopsies had continuously been collected at UCT and Groote Schuur Hospital for the duration of the project, and DNA had been extracted and stored at -20°C, it was decided to screen some of these additional patients who were not part of the WGS cohort using the primers designed for the *MUC3A* cluster 5 mutations. Table 5.17 provides the list of additional patients with their respective ages and genders.

Table 5.17: Additional patients screened for cluster 1 and 5 mutations. Patient DNA was not part of the WGS cohort

UCT Patient Number	Gender	Age (years)
GT675	Male	50
GT676	Female	60
GT677	Male	70
GT678	Male	57
GT679	Female	84
GT680	Female	61
GT684	Male	76
GT685	Male	74

Three of these patients' chromatograms contained too much background noise and were unsuitable for any kind of analysis or interpretation. The remaining five patients' had the same outcome as patients PD3944, PD39454 and PD39457. None of the mutations identified through the bioinformatics analysis were present.

These perplexing findings were entirely unexpected and provided a great deal of consternation as to how the bioinformatics pipelines provided a set of mutations identified in the *MUC3A* that could not be confirmed in laboratory PCR experiments. In an attempt to resolve these contradictory findings, additional bioinformatics analyses were performed to determine the validity of the mutations observed in the *MUC3A* gene in this study. The results of these additional analyses are described below.

5.6 Assessment of *MUC3A* Mutation Validity

5.6.1 Integrative Genomics Viewer (IGV)

Visual inspection of sample read alignment and the complexity of the genomic region where the *MUC3A* gene is located were visualised using the high-performance IGV tool [408,409]. The genomic co-ordinates of the gene on chromosome 7 for each sample were selected and the aligned reads were manually reviewed and interpreted. Visual inspection of such reads can increase confidence in the variant calls identified, flag the possibility of false positives and provide visual insight into complex genomic regions [408]. Figure 5.20 is a representative image of a matched tumour and normal pair visualised on IGV. All tumour-normal pairs in the cohort showed similar patterns of noise in the *MUC3A* gene, particularly in the region of exon 2, which is exactly where putative mutations were observed using out methodology.

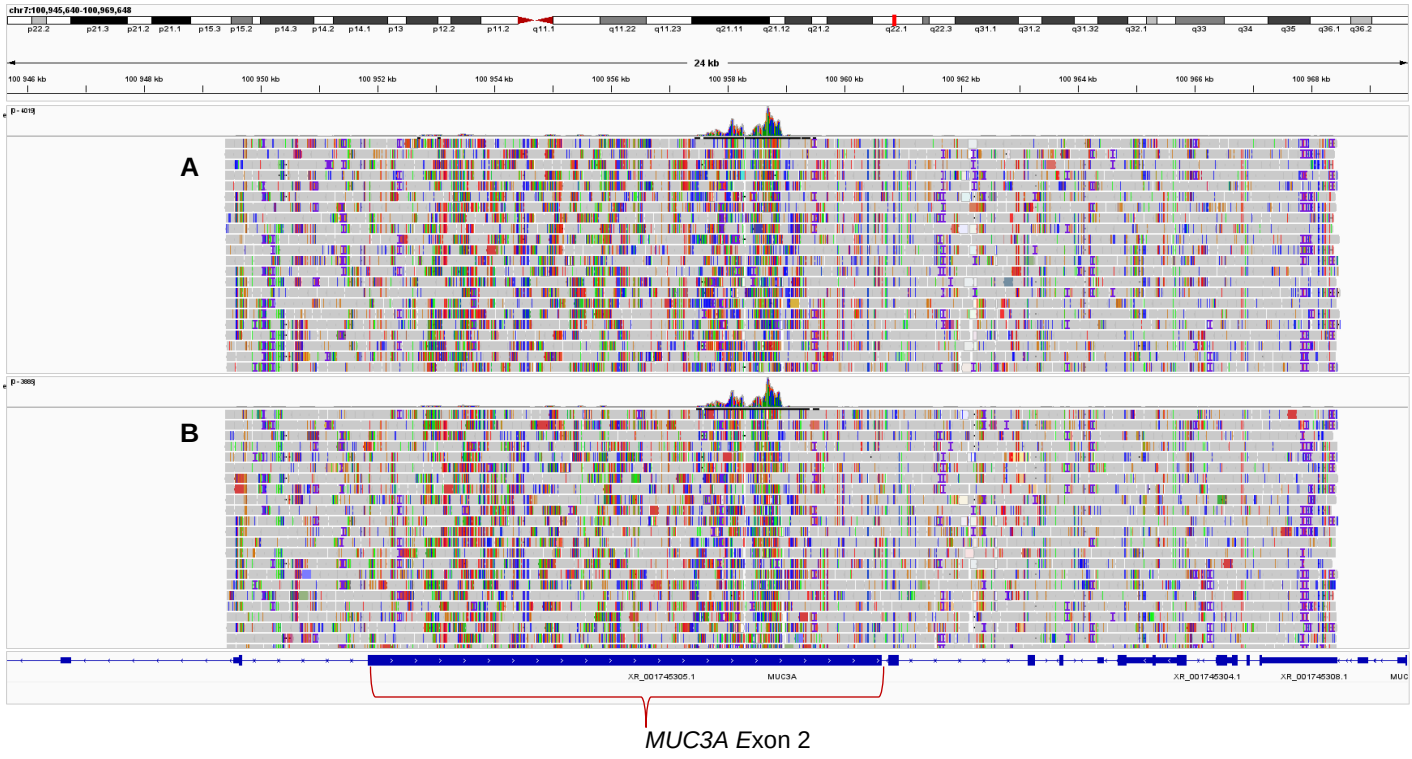


Figure 5.20: A representative IGV survey of the *MUC3A* gene on chromosome 7, showing excessive noise in both the A) tumour and B) matched normal sample, concentrated over exon 2. Samples were aligned against the GRCh38 human reference genome.

5.6.2 Panel of Normals

Following on from the wet laboratory findings where we were unable to confirm the *MUC3A* mutations through PCR, together with the noisy IGV assessments of exon 2 of *MUC3A*, it was imperative that we revisited the bioinformatics findings to further investigate the validity of the *MUC3A* mutations that were identified in the bcbio-nextgen pipeline. The high proportion of noise in both the tumour and normal samples seen in IGV images indicate that this particular gene, specifically exon 2, may be more problematic than originally expected and that it is likely that a large number of false positive somatic mutations in the variant calls may be present.

A 'Panel of Normals' (PON) was incorporated into the bcbio-nextgen pipeline to investigate whether the identified *MUC3A* mutations may have been false positives. Using a PON approach, a baseline level for variant calling is determined from a combined a set of normal samples typically derived from the same library preparation and sequencing workflow used for tumour samples to allow for non-sample specific system level biases to be subtracted [410]. In this way, variant calling results are improved as recurrent technical artifacts are removed. For short variant calling, it is recommended that the PON should be created and run using the variant caller Mutect2. Mutect2 is a variant detector for SNPs and indels and is part of the Genome Analysis Toolkit (GATK) [411,412]. An investigation by Bian et al., (2018) [413] examined the performance between several variant callers, including Vardict and Mutect2, run with bcbio-nextgen. Their findings indicated that Vardict produced the highest number of true positives, but also produced a high number of false positives. Mutect2 was found to be one of the best tools for detecting true positives and controlling for false positives.

PON files were created for each normal sample in the cohort (https://github.com/VictoriaPatten/phd-scripts/blob/main/Panel_of_Normals/Create_PON.sh) and combined into a single zipped VCF file (https://github.com/VictoriaPatten/phd-scripts/blob/main/Panel_of_Normals/combine_PONS.sh). The original bcbio-nexgen configuration files described above were edited to include Mutect2 as the somatic variant caller instead of the previously used Vardict, and the *background:* parameter was set to include the PON of all 35 normal samples. Bcbio-nextgen pipelines were re-run for all 35 tumour-normal pairs (https://github.com/VictoriaPatten/phd-scripts/blob/main/Panel_of_Normals/bcbio_with_pon.yaml) and the resulting VCF files were filtered for HIGH impact *MUC3A* mutations.

5.6.2.1 *MUC3A* Mutations When Using Mutect2 and PON

After re-running the bcbio-nextgen pipeline for all 35 tumour-normal pairs using Mutect2 and the PON approach, an entirely new set of HIGH impact *MUC3A* mutations was identified. More than 400 incidences of *MUC3A* mutations were now shown across all 35 samples in the cohort, with HIGH impact severity status. Out of these mutation events, a number were tagged with dbSNP (Single Nucleotide Polymorphisms Database) accession numbers (https://www.ncbi.nlm.nih.gov/snp/), indicating their inclusion in online databases as identified in previous studies. However, when checking these on the dbSNP database, mutations were listed as 'dubious' or synonymous and were non-pathogenic. Furthermore, all the mutations identified using Vardict variant caller were filtered out. This strongly suggests that the mutations in the *MUC3A* gene using the Vardict variant caller were false positives.

Figure 5.21 shows the total number of *MUC3A* mutations identified per patient using Mutect2 and the PON approach.

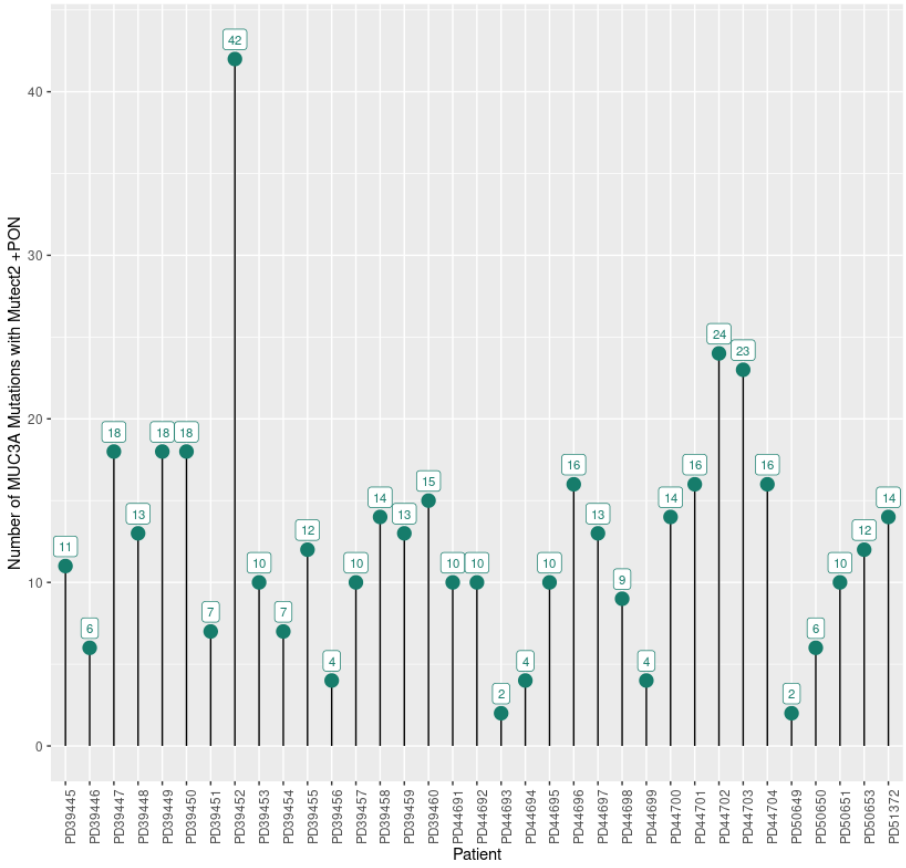


Figure 5.21: Lollipop plot of the number of *MUC3A* mutations per patient, detected using Mutect2 and the PON approach with bcbio-nextgen pipeline.

Table 5.18 provides a list of the *MUC3A* mutations that were common to a number of patients throughout the cohort. Several mutations with dbSNP accession numbers were detected in multiple patients and all mutations identified were of HIGH impact severity, being frameshift mutations. The list of all mutations detected are shown in Appendix 5.

Table 5.18: List of common mutations detected across the patient cohort, using Mutect2 and the PON approach with the bcbio-nextgen pipeline. Start refers to the variant genomic location, ref refers to the reference allele alt refers to the alternative allele, Impact indicates the type of variant identified, Impact severity denotes the severity of the variant, and Number of patients shows the number of patients within the cohort of 35 that shows the particular variant.

start	dbSNP Reference	ref	alt	Impact	Impact Severity	Number of Patients
100955010	rs1344850949;	C	T,CG	frameshift	HIGH	23
100956694	.	G	GA	frameshift	HIGH	23
100956690	.	TG	T	frameshift	HIGH	19
100955606	.	C	CAG	frameshift	HIGH	15
100955607	.	CCT	C	frameshift	HIGH	15
100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH	14
100954987	.	AC	A	frameshift	HIGH	13
100954991	.	C	CGGCG	frameshift	HIGH	13
100956808	.	ACC	A	frameshift	HIGH	13
100956813	.	G	GAA	frameshift	HIGH	13
100955595	.	TC	T	frameshift	HIGH	11
100955585	.	CTG	C	frameshift	HIGH	10
100955302	rs1584801296;	C	CA	frameshift	HIGH	9
100955539	.	ATG	A	frameshift	HIGH	9
100955560	rs1584801446;	AGT	A	frameshift	HIGH	9
100955296	.	GC	G	frameshift	HIGH	8
100955126	.	CTA	C	frameshift	HIGH	7
100955133	.	C	CTT	frameshift	HIGH	7
100955149	.	T	TTG	frameshift	HIGH	7
100955774	.	CA	C	frameshift	HIGH	7
100952647	rs773185111;	AGG	A	frameshift	HIGH	6
100952650	rs765973058;	AC	A	frameshift	HIGH	6
100953283	.	ATG	A	frameshift	HIGH	6
100953330	rs1237518715;	ACT	A	frameshift	HIGH	6
100955153	.	TCC	AGC,T	frameshift	HIGH	6
100955537	rs1305667129;	G	A,GTA	frameshift	HIGH	6
100955571	.	A	ATG,G	frameshift	HIGH	6
100953281	rs1229668308;	G	GTA,A	frameshift	HIGH	5
100953350	.	C	CAG	frameshift	HIGH	5
100953351	.	CCT	C	frameshift	HIGH	5
100953359	rs986103920;	C	CAA,T	frameshift	HIGH	5
100955565	.	G	GAA	frameshift	HIGH	5

100955779	rs1313584776;	C	CA,A	frameshift	HIGH	5
100963159	.	G	T	frameshift	HIGH	5
100963171	.	G	GTC	frameshift	HIGH	5
100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH	5
100953361	.	ACC	A	frameshift	HIGH	4
100953949	.	AC	A	frameshift	HIGH	4
100953954	.	A	AG	frameshift	HIGH	4
100955004	.	CAT	C	frameshift	HIGH	4
100955247	.	TTC	T	frameshift	HIGH	4
100963179	.	AGTGTCTGTGG	A	frameshift	HIGH	4
100954960	rs1584801081;	AAC	GAC,A	frameshift	HIGH	3
100954965	.	C	CTT	frameshift	HIGH	3
100955615	.	C	CAA	frameshift	HIGH	3
100955617	.	AC	A	frameshift	HIGH	3
100955920	rs1584801776;	A	AG	frameshift	HIGH	3
100963173	.	C	CCGTATCATTAA	frameshift	HIGH	3
100952718	.	CCT	C,ACT	frameshift	HIGH	3
100955916	rs1584801772;	AG	A	frameshift	HIGH	2
100952716	rs773095268;	C	CGT	frameshift	HIGH	2

From this PON approach with Mutect2 variant caller, a much larger number of *MUC3A* mutations were identified, all falling in the second exon of the *MUC3A* gene. However, the IGV image shown in Figure 5.20 indicates that in both tumour and normal samples, this genomic region is extremely noisy and it is very likely that most, if not all, of these putative mutations are not valid. All of the mutations previously identified with Vardict have been filtered out suggesting they were false positives. This could explain why we were unable to confirm the presence of these mutations experimentally. Given the experiences of this study we believe the present results in Table 5.18 and Figure 5.21 probably reflect false positive mutations as well. In the new set of mutations, every patient in the cohort was found to have *MUC3A* mutations which is highly improbable. It is likely that this revised data also suffers from false positives due to the difficult nature of the *MUC3A* genomic structure. We believe that extensive further analysis would be necessary to determine whether these mutations are in fact true and at this point we are unable to conclude their accuracy although we would suggest that they are probably all false positives.

5.8 Discussion

The investigations into the presence of somatic mutations in the sample cohort proved to be a tumultuous endeavour with conflicting results. The initial bioinformatics pipeline set-up and testing was a lengthy process requiring many re-configurations and troubleshooting of the software packages before the installation and set-up ran without errors. Once satisfied that the pipeline was running correctly, patient data files were run through the pipeline in their matched pairs, and this was carried out in parallel. Due to the large size of WGS data files, this analysis took approximately 4-5 days per patient. Both the GEMINI and SnpEff tools were utilised following pipeline execution to filter and search the output for somatic variants in specified genes, as well as for all variants in all genes with a HIGH impact severity.

When searching specifically for genes that had been identified in Chapter 4 as locations where, or near to where HERV insertions were found, the output showed that only four out of the initial thirty patient cohort showed MED (medium) impact severity mutations. No HIGH impact mutations were found thus eliminating the posibilty of linking the HERV insertions to functionally important somatic mutations. Without any HIGH impact mutations, nor a significant amount of MED impact mutations, it was clear that the insertion and possible translocations of HERV's in the genome couldn't be linked to any genomic variants that might lead to downstream anomalies.

To further test the validity of the bcbio-nextgen pipeline, filtering of the pipeline output was performed to search for genes previously identified by the Wellcome Sanger Institute's own analysis and reported as playing a role in OSCC. 24 out of these 29 mutated OSCC-associated genes appeared in our filtering analysis using both GEMINI and SnpSift, with both MED and HIGH impact mutations. Five of the genes however (*CCND1, CREBBP, CUL3, EGRF* and *NOTCH2*) were not identified as having HIGH impact severity variants in our pipeline and this could possibly be due to sensitivity differences of the different software used, the use of different variant callers or false positives reported in previous analyses.

We felt comfortable proceeding with further investigations and performed an extensive search of the database and VCF output files to identify all genes throughout the cohort where multiple instances of HIGH impact mutations were observed. The results from this search provided some expected results such as *TP53*, *CDKN2A* and *KMT2D*, but the most interesting finding was that a gene not previously described as associated with OSCC,

appeared to be the most highly mutated gene (with HIGH impact mutations) with 258 detected mutations across the cohort, 96% of which were frameshift variants.

Mutations in the *MUC3A* gene far exceeded numbers detected for other more commonly mutated genes. This transmembrane mucin gene has been found to be highly expressed in various epithelial cells of the intestines [256,406], whose protein product is reported to be involved in cellular protection through barrier function as well as in intracellular signal transduction pathways for regulation of inflammation, cell adhesion, cellular differentiation and apoptosis [184]. It is very interesting that while *MUC3A* has not yet been implicated in OSCC, it has been reported to play a role in other cancers including breast, prostate, gastric and renal cancers where elevated expression of the gene has been linked to poor disease prognosis [257–259,405]. These findings thus provided a clear path for the trajectory of the project and investigations into *MUC3A* became the focus of the study.

Following these observations it was necessary to confirm the presence of these mutations in the laboratory through PCR amplification and sequencing of patient genomic DNA. The design of suitable primers however, proved to be incredibly difficult due to the high number of tandem repeats causing a complex and complicated genomic structure with frequent multiple binding sites. Five sets of primers were designed to facilitate the PCR of the *MUC3A* mutations which grouped into five locations on exon 2 of the gene. These were colloquially referred to as ‘clusters’ for the purpose of this study. Optimisation of the primer sets proved just as difficult as the designing of the primers, and in the end only primers for clusters 1 and 5 were successfully optimised. Numerous PCR runs were performed and eventually suitable PCR products were subjected to post-PCR Sanger sequencing. Chromatograms of the resulting sequences however, showed that none of the expected *MUC3A* mutations identified in the bioinformatics analysis were present in the patient sequences, and rather, sequences matched to the human reference sequence exactly. Further screening of additional patients was done using cluster 5 primers, and again, no mutations were identified and sequences matched the reference. Given this observation, we became suspicious that the mutations identified through the bioinformatics pipeline might be false positives.

The inability to experimentally confirm the detected *MUC3A* mutations cast doubt over their validity and it was imperative to revisit and re-examine the data and bioinformatics pipeline. A colleague in the laboratory investigating different OSCC genes was able to confirm mutations identified through the bioinformatics pipeline in *PIK3CA* and *CDKN2A* genes

through PCR amplification and sequencing [414]. These results suggested the problem may lie with the *MUC3A* gene itself and the analysis thereof rather than with the pipeline explicitly. We postulated a number of different plausible explanations for this and a possible explanation might be that the genomic sequence of exon 2 of *MUC3A* is filled with tandem repeat structures with a high proportion of serine, threonine and proline residues [178,184,415], a difficult region to PCR, thus making primer specificity incredibly tricky due to multiple binding sites. It is possible that the primers that were designed were not specific enough and may have produced PCR products that were not specific to the region of interest. However, a more probable and likely explanation was that the mutations identified in the analysis were false positives given the nature of the *MUC3A* genomic sequence and the likelihood of WGS difficulties with this gene.

In order to validate the mutations and the bioinformatics pipeline it therefore was necessary to manually investigate the specific genomic region using the IGV tool. This confirmed the high degree of complexity of the gene at an individual read level, as all samples showed a high level of background noise and technical artefacts in both tumour and normal samples. This observation cast further doubt on the validity of the *MUC3A* mutations. The likelihood of false positives became the more apparent explanation and prompted the reanalysis of the WGS data using a PON approach and Mutect2 variant caller, with the hopes to eliminate these false positives. This was in fact achieved with the PON as following the re-running of the bcbio-nextgen pipeline, all mutations previously identified using Vardict variant caller were filtered out and we were able to conclude that they had indeed been false positives. However, a new set of *MUC3A* mutations was identified and further questions and speculations around their validity were given our previous experience and since further experimental confirmation attempts were not carried out due to time constraints.

A study conducted by Bian *et al* (2018) sought to compare the performance of a number of different variant callers, including Vardict and Mutect2 with bcbio-nextgen software. They questioned whether different callers might perform differently on different parts of the genome and whether guanine-cysteine content might affect analysis results. Their investigations showed that differences in true positives between callers was small, but the number of false positives varied far more. Furthermore, callers experienced diminishing accuracy when exposed to increasing levels of data complexity and that sequencing properties such as read depth, read quality, strand bias, and varying allele fractions can challenge a given caller's ability to accurately detect mutations. Their results showed that

Vardict produced the highest number of true positives, but in a trade-off, also produced a high number of false positives, while Mutect2 was among the best performing tools for detecting true positives and controlling for false positives [413].

The study by Bian *et al* (2018) brings into perspective the challenges faced in this doctoral study. It is evident that the *MUC3A* gene has a high degree of genomic complexity, especially in the region of the second exon, which has subsequently led to difficulties in variant calling of true positive variants, greatly affected by the variant callers used. The value of NGS data is wholly dependent on valid methods of interpretation and the accurate analysis and identification of mutations. Eliminating the presence of false positives is imperative and in this instance, extensive further investigations would be needed to conclude whether the new putative *MUC3A* mutations identified in the reanalysis with the PON approach were indeed true positives, however our prediction is that they too are false positives. Statistically it is improbable that all 35 patients in the cohort would have mutations in this gene. Furthermore, the high volume of mutations detected in MUC3A (as indicated in figure 5.7) raises the question of why no previous OSCC studies have identified and reported mutations in this gene. *MUC3A* has a highly complex and difficult to work with genomic structure and this is likely causing issues with its analysis.

In future, it would be advisable to include multiple variant callers in any bioinformatics analyses of *MUC3A* to integrate results and improve calling accuracy. We can conclude that the initial *MUC3A* mutations identified with Vardict were false positives. The putative *MUC3A* mutations identified using the Mutect2 and the PON approach require further experimental confirmation to determine whether they are true of false positives, however, the analysis of RNA-seq data was performed to try to identify whether these putative mutations were reflected in RNA expression patterns, and whether this affected gene expression within the patient cohort. The RNA-seq analysis will be discussed in Chapter 6.

Chapter 6: Analysis of Gene Expression

6.1 Introduction

6.1.1 Differential Gene Expression in OSCC

In order to understand the molecular basis of OSCC and to select potential candidate genes for the downstream development of anti-tumour drugs and treatment strategies, the analysis of the changes in gene expression between OSCC tumour samples and non-tumour samples is of critical importance. Molecular techniques have been developed to analyse changes in gene expression of thousands of genes simultaneously in an attempt to develop early diagnosis and prognosis strategies [416]. The discovery and identification of gene expression profiles may enable individualised targeted therapy for OSCC patients [417].

A study conducted by Lu *et al* (2014) found 1315 genes were differentially expressed in OSCC and of the 715 differentially expressed (DE) genes with known functions, 42.24% were related to the immune response, cell adhesion and cell metabolism [416], and EGFR has been reported to be expressed in 33.3% of OSCC cases [418]. It is speculated that some DE genes serve as beneficial novel therapeutic targets or diagnostic or prognostic markers [416].

Analysis of differential gene expression (DGE) between tumour and normal samples of OSCC patients may therefore provide valuable information for further investigations into the molecular basis underlying oesophageal cancer and tumorigenesis. In this way, more effective management, personalised therapy and potentially, early detection strategies may be developed from the elucidation and better understanding of gene expression profiles in OSCC.

6.1.2 RNA Sequence Analysis

With the rapidly advancing and ever improving technology present today, it is possible to analyse many aspects of RNA biology, from contributing to an essential understanding of genomic function, to investigating development and illuminating molecular dysregulation leading to cancer and other diseases [419]. RNA sequencing (RNA-seq) was developed more than a decade ago, and has been increasingly shown to provide insight into our understanding of genomic function [419,420].

High-throughput sequencing technology is widely acknowledged as the standard method for measuring RNA expression levels [421]. The recent large reduction in sequencing costs together with the advent of rapid sequencing technologies have enabled researchers to examine detailed profiling of gene expression levels [422]. Current RNA-seq technology facilitates a range of downstream analyses including comprehensive identification of gene isoforms, translocation events, nucleotide variations and post-transcriptional modifications [423,424], with one of the main objectives being the identification of differential gene expression (DGE) between two or more conditions, such as tumour and normal phenotype [425]. The correct identification of differentially expressed (DE) genes between specific conditions is critical in illuminating phenotypic variation [426]. Common methodologies for RNA-seq and DGE analysis include mapping of RNA sequences to a reference genome or transcriptome, followed by the estimation of gene expression levels and normalisation of mapped data using statistical and machine learning methods to identify DE genes. The obtained data is then evaluated for biological relevance [423,427,428]. It is expected that the number of sequenced fragments mapping to a transcript correlate directly with abundance levels and provide a measurable level of expression of each RNA unit. Sequencing depth is known to limit such expression signals. Recently, there has been much motivation to develop numerous statistical algorithms to implement a variety of approaches for normalisation and DGE analysis [425].

Most commonly, RNA-seq is used as a routine research tool to seek and identify DGE among sample groups, illuminating the underlying biology of the system of interest [423]. It involves multiple steps of a standard workflow, from RNA extraction from samples in a laboratory, followed by mRNA enrichment or ribosomal RNA depletion, cDNA synthesis and the preparation and sequencing of an adaptor-ligated sequencing library to a read depth of 10-30 million reads per sample on a high throughput platform (usually Illumina) [419,423,429].

Computational steps then follow via an automated analysis of raw sequence data to functionally annotated gene results. This requires the coordination of multiple steps and tools for the alignment and/or assembly of the sequence reads to a transcriptome or reference genome, quantification of reads overlapping reference transcripts, sample filtering and normalisation, and finally, statistical modelling for the reporting of significant changes in the expression levels of individual genes and/or transcripts between sample groups [419,429]. RNA-seq analysis involves the repetition of commands using various tools for the

quantification of gene expression, data quality control to identify DGE, as well as functional enrichment categories performed on a per-sample basis. [429].

The analysis of the raw RNA-seq data in this study was performed using bcbio-nextgen [352] in conjunction with a Bioconductor (BioC) [280] package called bcbioRNASeq [429] to aggregate the outputs of tools for RNA-seq quality control (QC), DGE and functional enrichment analysis. R Markdown templates available on installation of bcbioRNASeq were used for these analyses.

Figure 6.1 provides an overview of the bioinformatics workflow used.

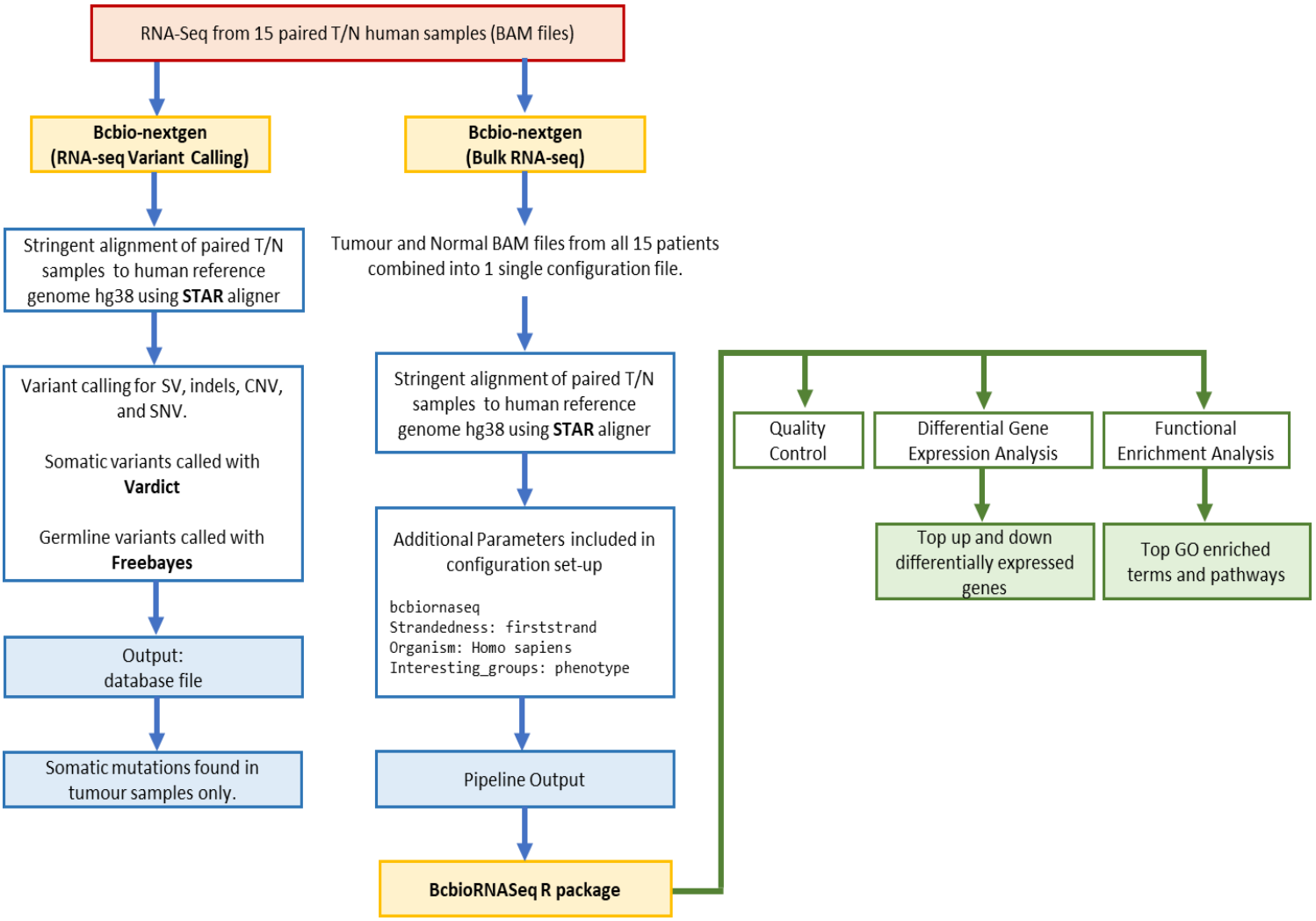


Figure 6.1: RNA-seq analysis pipeline overview using bcbio-nextgen and bcbioRNASeq R package to perform differential gene expression analysis and functional enrichment analysis. T/N = tumour and matched normal; GO = gene ontology.

6.1.3 Variant Calling with RNA-seq Data

6.1.3.1 RNA Sequence Data

RNA from fifteen patients was subjected to RNA sequencing as described in section 2.2.3. Eight of these patients had been recruited from UCT, while the remaining seven were from WITS University. Following sequencing, raw patient BAM files were transferred to the SANBI and Ilifu servers in South Africa for RNA-seq analysis.

6.1.3.2 Bcbio-nextgen pipeline for variant calling using RNA-seq data

As described for the WGS data, a variant calling pipeline was established using bcbio-nextgen software [352]. The RNA-seq pipeline configuration file in YAML specified the analysis type (RNA-seq) and the Spliced Transcripts Alignment to a Reference (STAR) aligner v2.7.10a [430] was used to re-align high throughput long and short RNA-seq data to the reference genome GRCh38. This aligner was developed to overcome common issues plaguing previous RNA-seq alignment methods such as high mapping error rates, alignment biases, low sensitivity, poor scalability, low mapping throughput and inability to detect non-linear chimeric RNAs [431].

The *strandedness* parameter was also specified in the configuration file. *Strandedness* can either be set as '*unstranded*' (default), '*firststrand*' or '*secondstrand*' depending on the library preparation methods used during the RNA sequencing. Setting the incorrect *strandedness* could account for up to 90% count loss in the analysis [354]. The RNA library preparation kit that was used incorporated dUTP during second strand cDNA synthesis thus according to the bcbio-nextgen documentation, *firststrand* was the correct specification.

The variant caller *vardict* [358] was used for somatic calling from tumour samples as described previously for WGS analysis. The configuration files used for this pipeline can be found at https://github.com/VictoriaPatten/phd-scripts/tree/main/bcbioRNAseq/RNAseq_VC_tumor.yaml.

Raw patient RNA-seq BAM files were specified as input for the pipeline which was then executed for each patient. The output from the pipeline was provided as a VCF file of detected variants.

6.1.3.3 RNA-seq Variant Calling Output

Output VCF files were separately obtained for each patient's tumour and normal samples. When analysing these files it was necessary to filter the data to validate the calls. High quality variants were labelled as PASS which indicates a depth ≥ 10 reads. From the variants that passed the quality control validation, we then filtered the data specifically searching for the *MUC3A* gene (Ensembl ID: ENSG00000169894) with HIGH or MED (medium) impact variants.

Because we had identified a large number of HIGH impact variants in the WGS data analysis, we hoped to find similar results with the RNA-seq analysis. However, in fifteen patient tumour samples, only 78 MED impact severity variants were detected in the MUC3A gene, and only two HIGH impact variants were discovered, one in patient PR50554 and one in patient PR50549 (Table 6.1). Furthermore, these two HIGH impact variants that were detected did not match any reference positions of any of the WGS variants that were identified and described in Chapter 5. A list of all variants detected is shown in Appendix 6.

Table 6.1: HIGH impact severity variants detected in MUC3A gene from RNA-seq analysis of patient tumour samples.

Patient	Position	Reference	Alternate	Mutation	Impact	Impact Severity
PR50554	100966942	CA	C	Deletion	frameshift_variant	HIGH
PR50549	100967142	AT	A	Deletion	frameshift_variant	HIGH

These results were perplexing as it was hoped to find a similar pattern of variation as observed in the analysis of the WGS data described in Chapter 5. Returning to the software documentation, we discovered reports suggesting that variant calling with RNA-seq data is frequently problematic. The bcbio-nextgen documentation reports high false negative rates when performing RNA-seq variant calling, indicating that 83% of variants are not called. When validating exome capture regions representing all protein coding genes, only a very small fraction of genes were expressed in normal samples and only these few genes had a read coverage suitable for variant calling. Indel calling specifically, was described as unreliable and imprecise [354]. Therefore a comparison of WGS results with RNA-seq variant calling is not suitable. However, as we believe the *MUC3A* mutations identified through WGS

are spurious, it was expected that variant calling with the RNA-seq data would not show similar mutations.

Most of the variants discovered with WGS pipeline were indeed small indels which may explain why this comparison provided conflicting results. It would be more suitable to investigate differential gene expression of RNA-seq data instead.

6.1.4 Bulk-RNA-seq Analysis to Determine Differential Gene Expression

6.1.4.1 Bcbio Pipeline

A new bcbio-nextgen pipeline was established, this time for bulk-RNA-seq analysis whereby all 15 patients' tumour and normal raw BAM files were used as input in a single YAML configuration file and executed through the pipeline as a singular job. STAR aligner was again specified but no variant callers were included. For a comprehensive quality control and differential gene expression analysis the *bcbiornaseq* parameter was included in the configuration set up. This supplies a dictionary of key-value pairs that are passed as options to the bcbioRNAseq R package [429,432]. The statistical computing software R, v 4.2.0 [433] was used for analysis and interpretation. This parameter supports *organism* as a key and takes the Latin name of the genome used, in this instance, *Homo sapiens*. The parameter further supports the *interesting_groups* key which was set as *phenotype* in this configuration to differentiate between 'tumour' and 'normal' samples.

After execution of the pipeline, quality control and gene abundance information was generated compatible with the bcbioRNASeq R package [432] for downstream analysis. Bcbio is able to assess the quality and complexity of the RNA-seq data using a combination of custom tools for the quantification of ribosomal RNA (rRNA) content as well as the genomic context of alignments in known transcripts and introns [429]. The important files and output generated from the bcbio run are saved by default in a '*final*' directory. Quality metrics of the data, provenance information and the data derived from the analysis that have been aggregated across all samples (i.e. count files) can all be found within a dated project directory created and saved within the parent *final* directory. In addition to this, individual directories corresponding to each patient sample are present containing the BAM files and count data for each sample. This *final* directory is later used as input for bcbioRNASeq [429].

6.1.4.2 bcbioRNASeq

bcbioRNASeq is a Bioconductor [280] package that aggregates the outputs of tools for RNA-seq QC, DGE and functional enrichment analysis [429]. The package relies on the output of the bcbio bulk-RNA-seq pipeline providing a unified package with objects, functions and pre-made templates for the continued next steps of RNA-seq analysis including the assessment of reads and alignment quality, the identification of outlier samples, clustering of samples, assessment of model fit, setting cut-offs and identifying DE genes. These R markdown templates are available upon installation of bcbioRNASeq [432] in R. They are ready-to-render and include the code for QC metrics, differential expression and functional enrichment analysis.

Before working with the markdown templates, it was necessary to create a structured S4 object containing all the necessary data and information from the bcbio run. By specifying the *uploadDir* argument which specifies the path to the *final* bcbio directory the object was created and used as input for the R templates. See https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbioRNAseq/bulk_rnaseq.yaml

6.1.4.2.1 Quality Control

RNA-seq analyses characteristically require QC assessments of reads, alignments, samples and models, and the data required for these assessments are generated by the bcbio run. The Qualimap tool [434] built into the bcbio pipeline generates several metrics to be used to assess the quality of the reads and consistency across the samples. These metrics are stored in the S4 bcbioRNASeq object (https://github.com/VictoriaPatten/phd-scripts/blob/main/bcbioRNAseq/bcb_obj.R) and when amending the QC template to specify the specific input object, these metrics are visualised graphically through the generation of specialised plots. These plots are useful in checking the data quality. Figure 6.2 **A-F** below show these graphical representations of the QC assessment of the RNA-seq data.

When analysing these statistical plots, total reads per sample and the mapping rate are given as metrics to identify any potential imbalances in sequencing depth or failure among the samples. Figure 6.2 **A** shows some variability, but in all cases the read coverage is higher than 100 million reads, which is considered as sufficient to provide good gene quality. Regarding mapping rates, it is expected that samples should have similar rates and should be greater than 75%. In this instance, all samples are well within the cut-off range having

around 90% reads mapping (Figure 6.2 **B**). Low genomic mapping rates would be indicative of sample contamination, poor sequencing quality or other artifacts. These graphs suggest good sequencing quality. To provide genomic context, it is expected that the majority of reads should map to exons and not introns. Ideally, at least 60% of total reads should map to exons. DNA or RNA contamination would result in high intronic mapping [429]. Figure 6.2 **C** and **D** clearly shows that nearly all samples are above 75% exon mapping, and far below the threshold for intron mapping, further indicating no DNA contamination.

Determining the number of genes detected relative to the number of mapped reads, is a further assessment of sample quality. Ideally, all samples should have similar number of detected genes/features and samples with a higher number of mapped reads will have more genes detected. Figure 6.2 **E** shows more features detected relative to normal samples, and all with fairly similar numbers. The gene saturation plot (Figure 6.2 **F**) clearly shows the relationship between the number of genes detected and the number of mapped reads.

As this trend appears to be linear, it is suggested that the sequencing was not yet saturated in detecting gene expression, indicating that more genes could possibly be detected if coverage were increased for the samples with lower numbers of reads.

It is further important to explore the fit of the model for the given dataset before performing DGE analysis. The normalized and transformed data can be used to assess the variance-expression level relationship in the data, to identify which method is best at stabilising the variance across the mean for downstream visualisation [429]. Plotting the mean standard deviation with the *plotMeanSD()* function wraps the output of different variance stabilising methods (Figure 6.3). The mean of the counts (log2) is plotted against the standard deviation using four separate normalisation tools, sf (size factor), vst (variance stabilising transformation, tmm (trimmed mean of M-values) and rle (relative log expression). This dataset shows that *tmm* and *rle* normalisation methods show smaller standard deviation and thus smaller variance for this dataset and are therefore more suitable for normalisation of the data moving forward.As this dataset can be classified into groups according to phenotype (tumour vs normal samples), it was further possible to investigate how similar the samples were to each other within the groups. To do this, bcbioRNASeq performed an Inter-Correlation Analysis (ICA) as well as a Principal Components Analysis (PCA) between the samples.

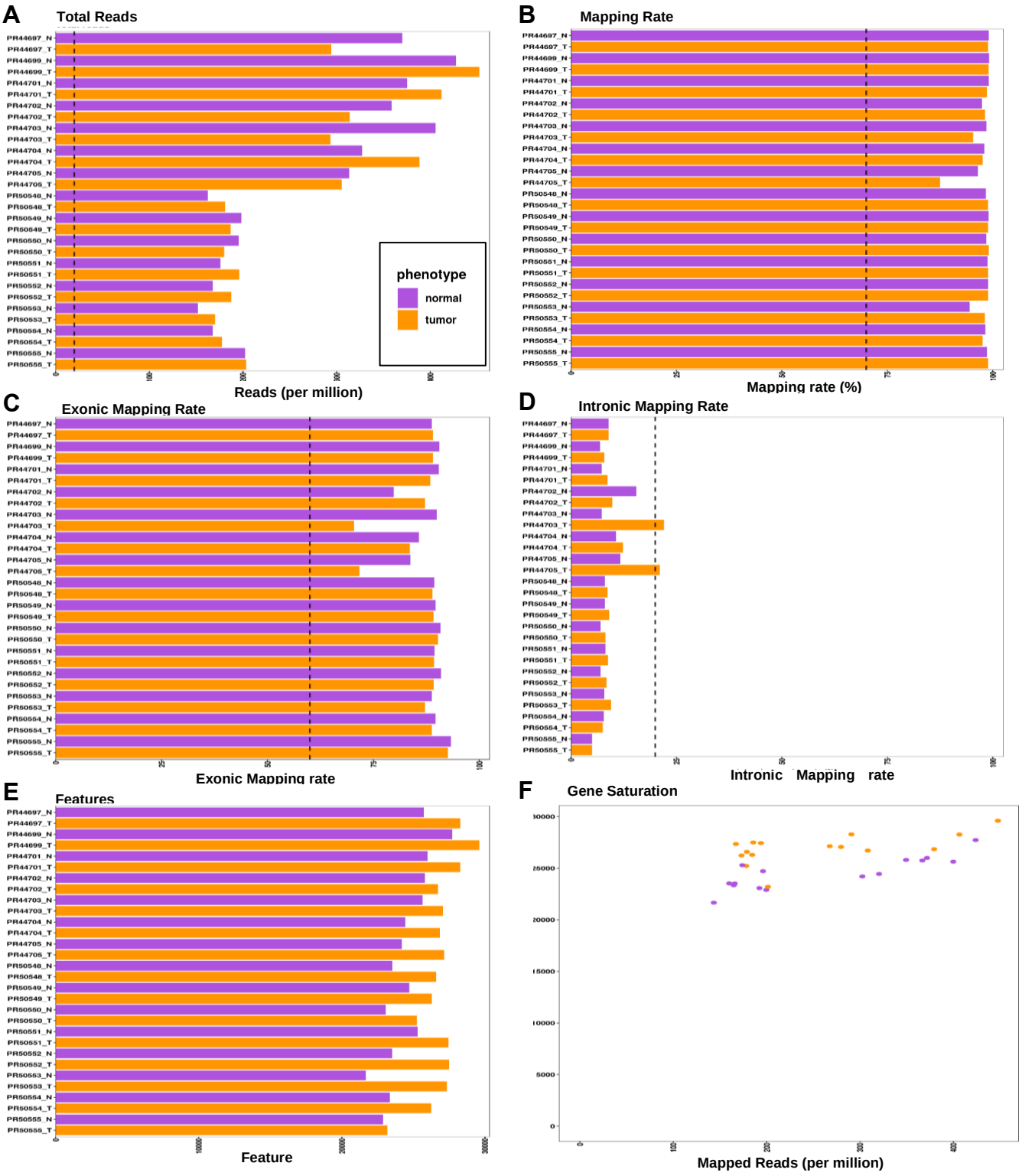


Figure 6.2: Read alignment, genomic context, and gene detection statistics. **A**) The total reads plot indicates the total number of reads sequenced per sample. **B**) Shows the percentage of reads mapping to the reference genome. The exonic and intronic mapping rate plots (**C** and **D**) indicate the percentage of reads mapping to exons or introns, respectively. The genes/features detected plot (**E**) indicates the total number of genes for each sample with at least one mapped read. The gene detection saturation plot (**F**) shows the relationship between the number of reads mapped and the number of genes detected. Vertical dashed black lines indicate suggested cut-off values. Normal samples are presented in purple and tumour samples in orange.

The ICA carried out a pair-wise correlation between expression profiles of all samples, followed by clustering based on these correlation patterns. Similar samples are described as highly correlated and are generally found to cluster together, thus one would expect samples from the same group to cluster together. Results of this correlation clustering are presented as a heatmap (Figure 6.4) where one is also able to identify outliers present in the dataset (samples with low correlation). PCA is a further technique used to summarise variation within the dataset where expression levels are transformed in principal component space, reducing each sample to a single point [435]. This allows one to separate samples by expression variation and to further identify outlier samples (Figure 6.5).

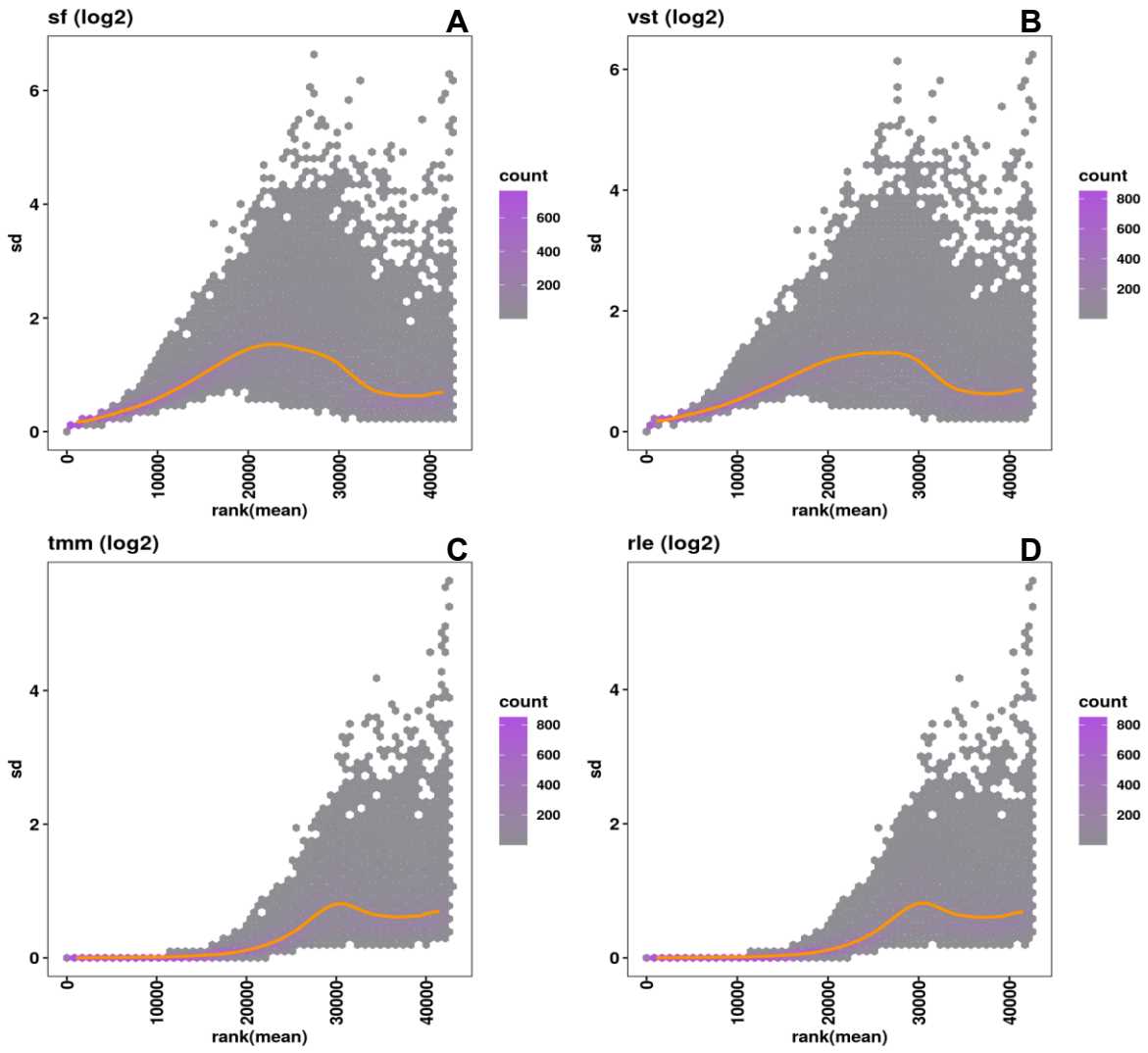


Figure 6.3: Variance stabilizing transformation plots showing the standard deviation of normalised counts using **A**) sf (size factor), **B**) vst (variance stabilising transformation), **C**) tmm (trimmed mean of M-values), and **D**) rle (relative log expression). The orange line denotes the running median estimator. These panels show that rle and tmm normalisation greatly reduce the standard deviation and variance across the mean. Count refers to normalised read count, expressed in log2.

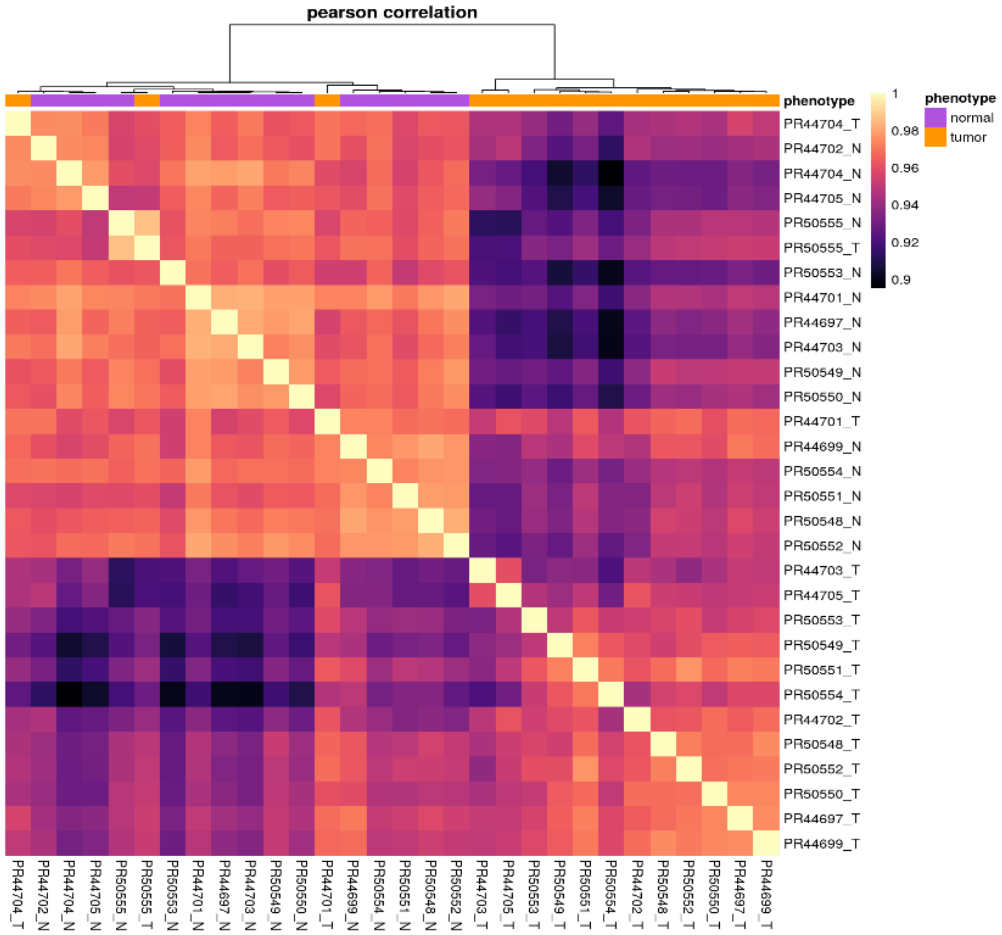


Figure 6.4: Correlation heatmap of a pairwise sample Pearson correlation where correlations are clustered by both column and row. The sample class (phenotype) is highlighted across the top of the heatmap. Three tumour sample outliers are detected within the purple (normal) cluster.

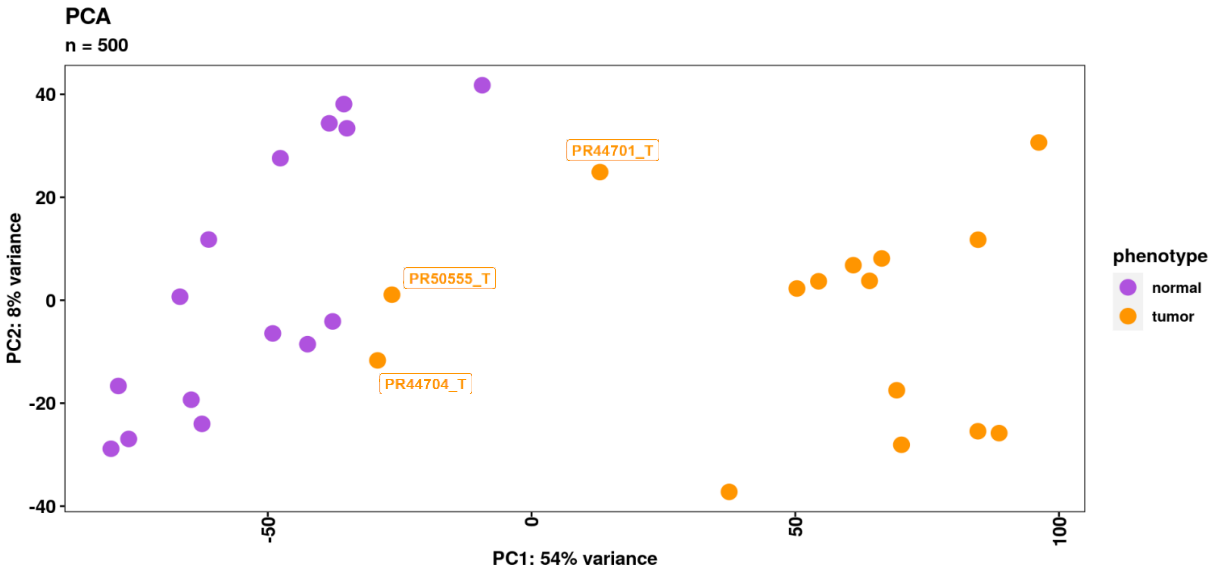


Figure 6.5: PCA plot for tumour and normal paired samples. Clustering of samples within the tumour and normal groups is shown, however three tumour outliers were detected, labelled on the plot. n=500 is a default value of the top 500 transcripts (genes) used that differentiate between tumour and normal samples.

The ICA and PCA plots give an indication of the clustering and therefore, the correlation of the samples within the tumour and normal groups. In both plots, it can clearly been seen that three outliers appear to exist within the tumour sample group, with these samples clustering close to and within the normal grouping. Three orange (tumour) blocks can be seen grouped within the purple (normal) bar across the top of the heatmap, while the same three samples are seen clustering with the normal samples on the PCA plot. Tumour samples PD50550, PD44701 and PD44704 appear to be more closely resemble normal samples as the biopsies from which the RNA was extracted were reported to show low percentage of tumour purity (section 2.1.1.1, Table 2.1).

Sample outliers were removed from the dataset and the bcbio analysis pipeline and QC metrics were repeated with the amended group so as to avoid skewed results. The raw data BAM files for these patients were removed from the dataset which was then re-run through the pipeline. A new *final* output directory was produced and this was used to create a new S4 object used as input as before in a new QC analysis template. Figures 6.6 and 6.7 show the updated ICA an PCA plots with outliers removed from the dataset

Removing outlier samples from the dataset eliminates samples with possible RNA material issues, the presence of high levels of normal tissue or instances where issues in RNA library preparation have occurred. In doing so, the best, most reliable set of sample data is put forward for further analysis of differential gene expression.

6.1.4.2.2 Differential Gene Expression

Once QC of the sample data had been performed and outliers removed from the dataset, the next step was to identify genes that were differentially expressed between the sample groups. As with QC, a bcbioRNASeq R markdown template was used for this analysis incorporating the DESeqAnalysis package [436] library which uses Salmon-derived abundance estimates imported with *tximport* from the bcbio object. The object with removed outlier tumour samples was used as input going forward. Tools used in the DGE analysis follow the approach of using regression-based models to estimate expression differences for a given gene, followed by statistical testing based on the null hypothesis that differences are close to zero [437]. In order to determine whether the read count differences between phenotypes for a given gene are greater than would be expected by chance, DGE tools estimate the differences based on the replicates of each condition [437].

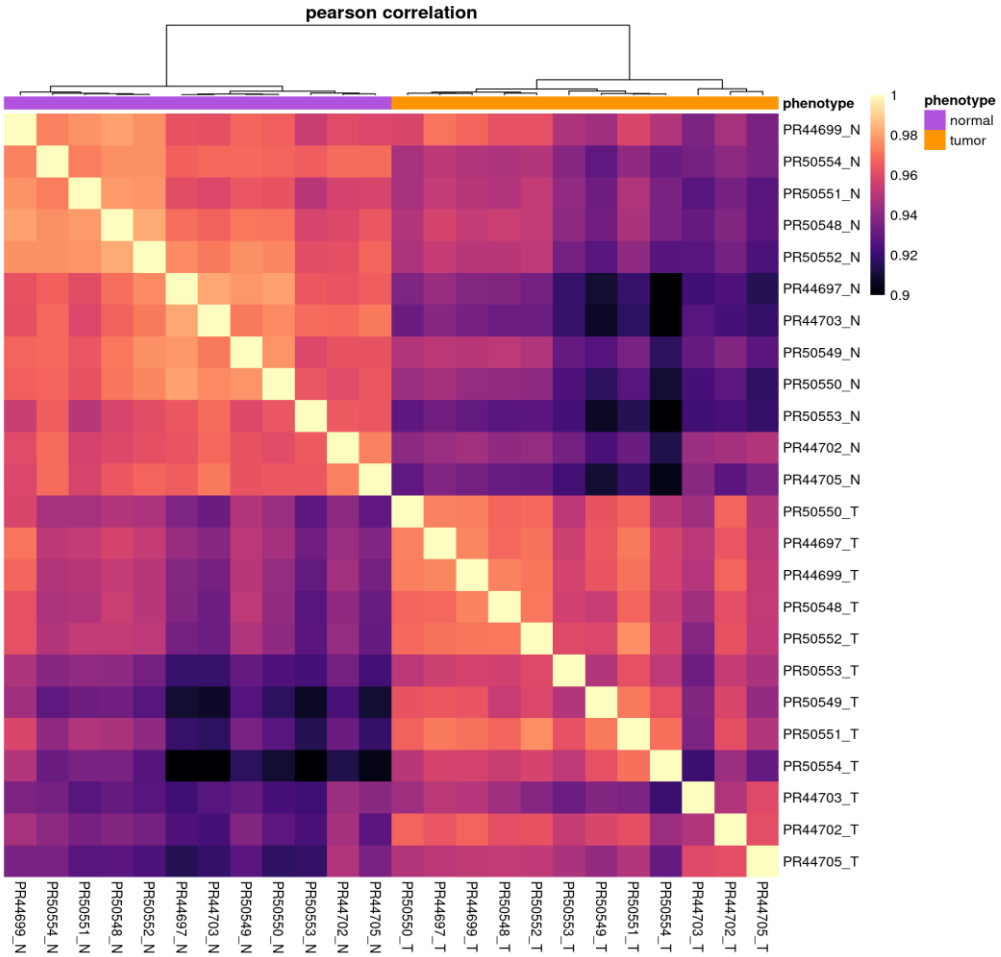


Figure 6.6: ICA plot of amended dataset with outlier tumour samples removed.

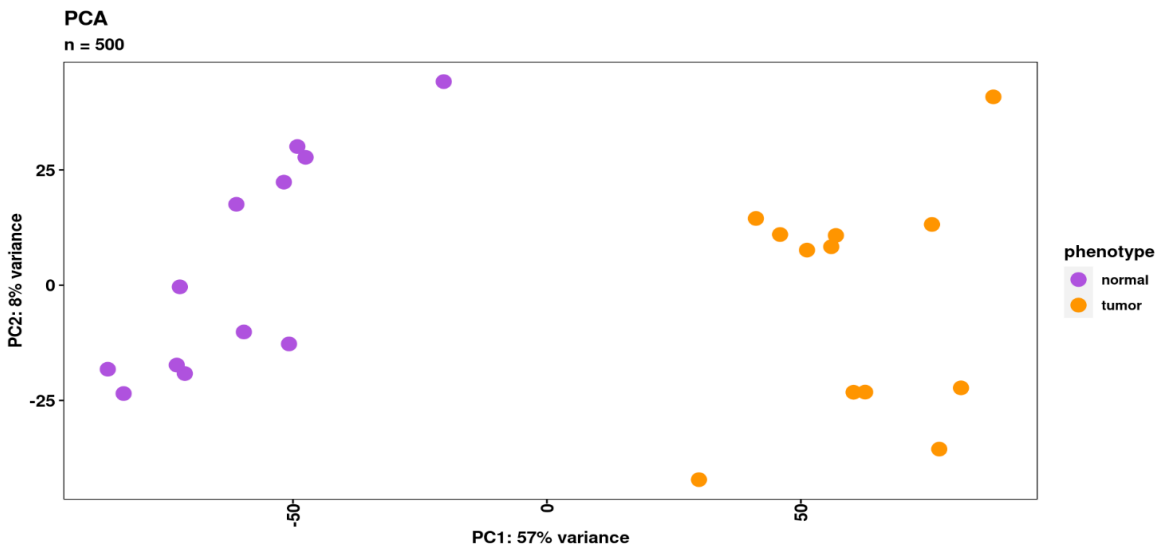


Figure 6.7: PCA plot of amended dataset with outlier tumour samples removed. n=500 is a default value of the top 500 transcripts (genes) used that differentiate between tumour and normal samples.

The DESeq2 tool utilised in this analysis makes use of this regression based model and was applied to every gene, relying in a negative binomial model to fit the observed read counts to arrive at the estimate for the expression differences [437].

Dündar *et al* (2015) [437] described that if *variance = mean,* as with Poisson distribution, precise estimates of mean read counts through RNA-seq analysis give an accurate indication of the kind of variance to be expected. This therefore allows for the identification of genes showing greater differences between two conditions (i.e. tumour vs normal phenotypes) than can be expected by chance.

Results obtained from the analysis include *P* values associated with each gene detected, as well as adjusted *P* values that are corrected for multiple testing including false discovery rates. The Benjamini Hochberg (BH) statistical method [438] was set as the default statistical test correction to control for false discovery rate (FDR). The alpha-threshold was set at 0.01 (the significance cut-off used for optimising the independent filtering) and the log2 fold change (LFC) threshold set at 1 (a non-negative value which specifies the LFC threshold. The default value is 0, corresponding to a test that the LFC's are equal to zero). LFC=1 was set to reject the null hypothesis (that no expression change occurs) and to include all small and large DGE changes in the analysis.

The *plotMA()* function was incorporated to visualise the mean of the normalised counts versus the log2 fold changes for all genes tested (Figure 6.8). Here each dot on the plot represents one gene. The LFC y-axis indicates the significant DGE of each gene, plotted against the mean expression of all samples. 2421 DE genes were identified, of which, 1564 were upregulated and 857 were downregulated, relative to the norm. The *plotVolcano()* function was also used to produce a volcano plot comparing significance (using BH-adjusted *P* values) for each gene against the LFC scale (Figure 6.9). Significantly upregulated genes are plotted on the positive axis, while downregulated genes are plotted on the negative axis.

The *plotDEGHeatmap()* function was further used to produce a gene expression heatmap (Figure 6.10) for the visualisation of the expression of differentially expressed (DE) genes across the samples. This heatmap shows DE genes on a per-sample bases, using an additional LFC cut off. By default, the heatmap plot is scaled by row, and uses the *ward.D2* method for clustering [439]. Figure 6.11 shows a further PCA plot from the *plotDEGPCA()* function demonstrating the correlation and clustering of the tumour and normal samples where the LFC threshold was set at 1 and the alpha threshold set at <0.01.

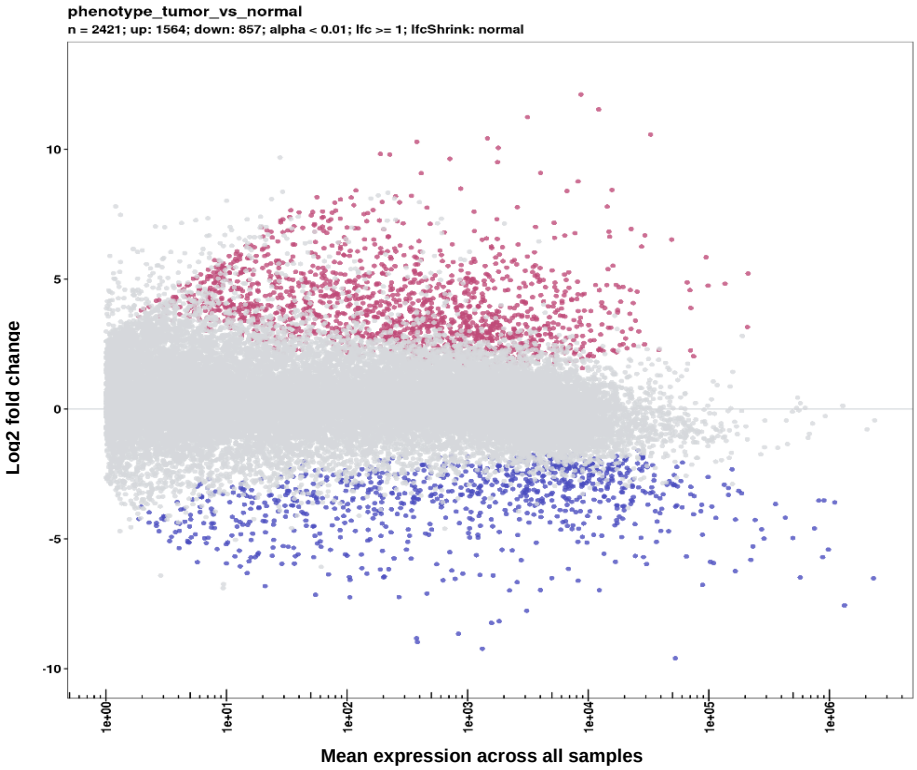


Figure 6.8: Mean Average expression levels across all samples plotted against log2 fold change observed in the contrast of interest on the y-axis. 2421 DE genes were detected of which 1564 genes were upregulated (red dots) and 857 genes were downregulated (blue dots). Each point on the plot represents one gene. LFC threshold was set at 1 while alpha value was plotted at <0.01.

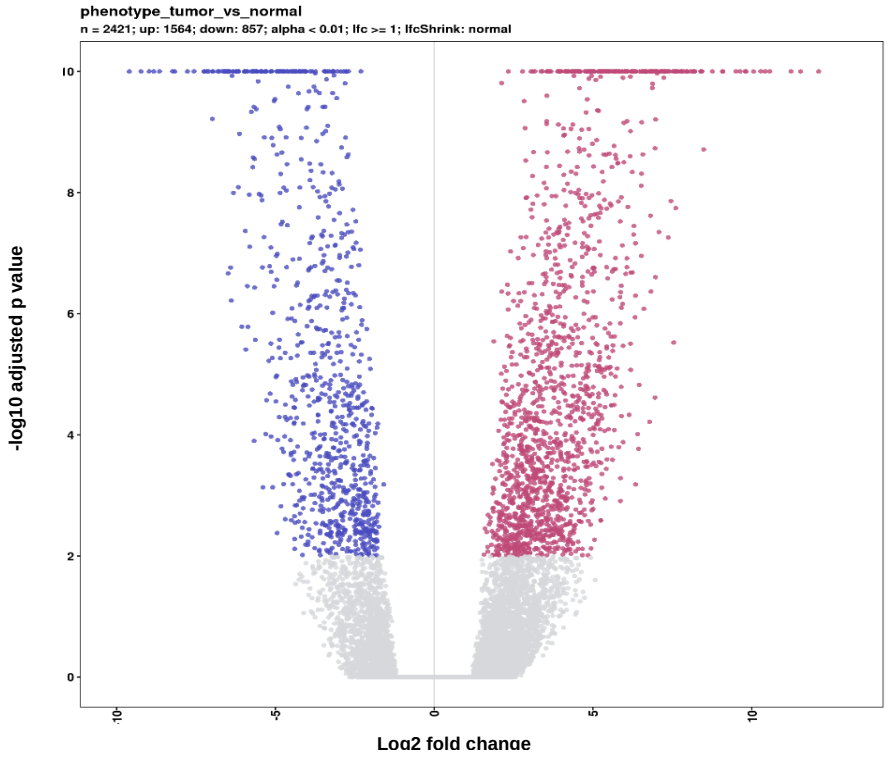


Figure 6.9: Volcano plot comparing the amount of gene expression to the significance of that change, plotted as -log10 transformation of the multiple test adjusted *P* value. Genes identified as significantly upregulated are plotted in red, while downregulated genes are plotted in blue.

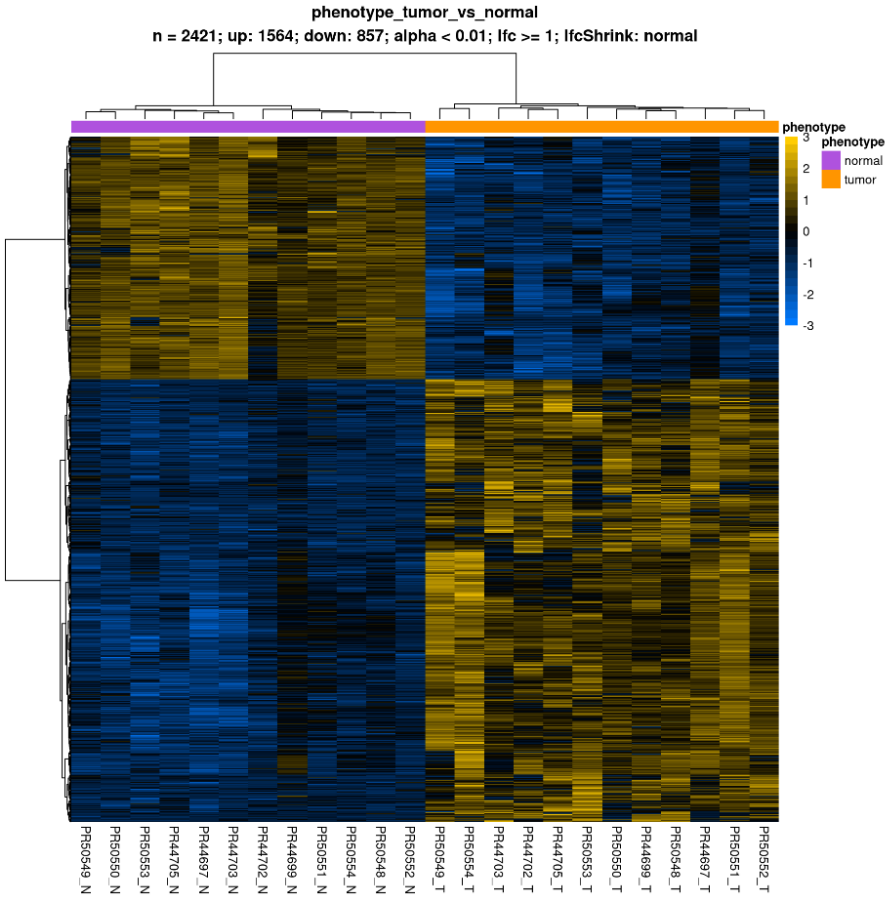


Figure 6.10: Differentially expressed genes heatmap showing DE genes on a per-sample basis. Results are clustered by both row and column, with sample classes (tumour vs normal) highlighted across the top of the heatmap.

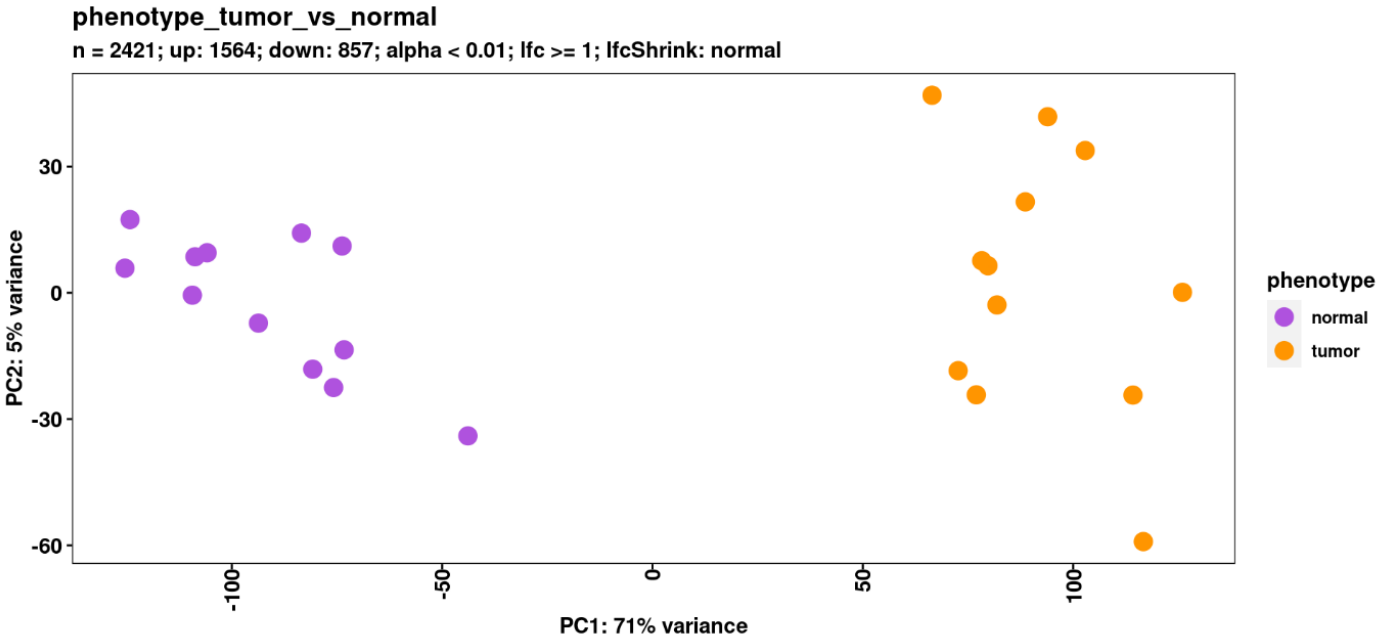


Figure 6.11: PCA plot of tumour and normal correlation clustering for the 2421 DE genes detected.

At the end of the analysis, results tables were generated using the *results()* function in the template whereby tables were extracted with log2 fold change, *P* values and adjusted *P* values for all 2421 DE gene. 1564 upregulated genes and 857 downregulated genes were recorded, with the significance of the differential expression indicated by the adjusted *P* values, the log2 fold change and the base mean (mean of the normalised counts per gene for all samples) per gene. Table 6.2 shows the top 10 up- and down- regulated genes detected in the DGE analysis, where the most significantly up- and down-regulated genes were MMP11 and *AC005532.1* respectively.

Table 6.2: Top 10 up- and downregulated genes in the tumour samples when compared to the normal samples.

	Ensembl ID	Gene Name	Log2 Fold Change	P Adjusted Value
Upregulated Genes				
1	ENSG00000099953	*MMP11*	7.17	$1.22e^{-53}$
2	ENSG00000123500	*COL10A1*	11.24	$2.69e^{-47}$
3	ENSG00000196611	*MMP1*	10.57	$7.15e^{-44}$
4	ENSG00000149968	*MMP3*	12.11	$2.85e^{-42}$
5	ENSG00000163064	*EN1*	8.21	$3.04e^{-35}$
6	ENSG00000106366	*SERPINE1*	7.79	$3.68e^{-34}$
7	ENSG00000120708	*TGFBI*	6.26	$1.04e^{-32}$
8	ENSG00000137745	*MMP13*	11.54	$2.04e^{-29}$
9	ENSG00000133110	*POSTN*	6.63	$1.97e^{-28}$
10	ENSG00000170454	*KRT75*	9.50	$4.53e^{-28}$
Downregulated genes				
1	ENSG00000230825	*AC005532.1*	-9.23	$3.24e^{-78}$
2	ENSG00000096006	*CRISP3*	-9.60	$6.14e^{-49}$
3	ENSG00000267709	*AC024592.2*	-8.83	$1.34e^{-43}$
4	ENSG00000244040	*IL12A-AS1*	-7.24	$2.70e^{-30}$
5	ENSG00000196260	*SFTA2*	-8.24	$6.55e^{-30}$
6	ENSG00000226051	*ZNF503-AS1*	-5.98	$8.59e^{-28}$
7	ENSG00000077063	*CTTNBP2*	-5.45	$5.01e^{-27}$
8	ENSG00000150672	*DLG2*	-4.84	$5.01e^{-27}$
9	ENSG00000228789	*HCG22*	-7.77	$2.23e^{-25}$
10	ENSG00000258616	*LINC02303*	-8.97	$2.23e^{-25}$

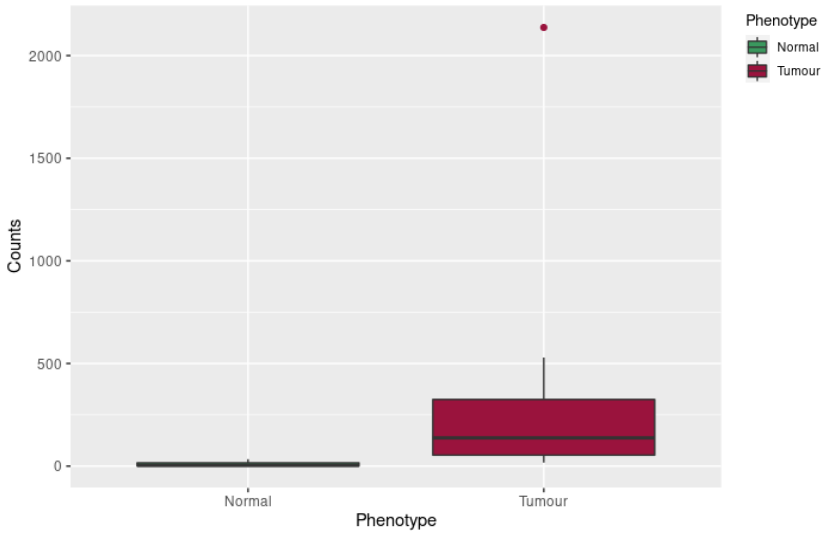
Although *MUC3A* in tumour samples did not feature in the top 10 upregulated genes, it had an adjusted *P* value of $7.05e^{-06}$ and a LFC of 4.6 (Table 6.3). This indicates that the expression of *MUC3A* was indeed 24.25 times higher (i.e. if LFC=4.6, then fold change = $2^{4.6}$) in the tumour samples than the corresponding normal samples, and this change in expression can be viewed as highly significant.

Table 6.3: *MUC3A* upregulation in tumour samples.

	Ensembl ID	Gene Name	Log2 Fold Change	P Adjusted Value
570	ENSG00000169894	*MUC3A*	4,6	$7.05e^{-6}$

The differential expression of *MUC3A* was ranked at 570 out of 1564 upregulated genes. This falls close to the top third of all upregulated genes detected. While it may not fall in the top 100 genes, its increased expression is still a worthwhile observation as it aligns with a number of other studies reporting similar findings in other cancers, for example, in previous studies the increased expression of *MUC3A* has been found to exert oncogenic profiles in breast, pancreatic, gastric, colorectal, prostate and renal cancers, eliciting poor prognosis in patients [257–259,405].

Furthermore, when analysing the *MUC3A* gene counts for tumour and normal samples, this significantly increased expression is clearly visible (Figure 6.12). The *tximport()* function incorporated into the bcbio pipeline provides transcript-level estimates of counts and abundance.



Patient	Normal	Tumour
PR44697	13	295
PR44699	33	112
PR44702	0	35
PR44703	26	358
PR44705	0	56
PR50548	3	245
PR50549	2	151
PR50550	7	508
PR50551	5	16
PR50552	20	2045
PR50553	0	17
PR50554	1	109

Figure 6.12: Box plot of *MUC3A* counts per sample, tumour vs normal taken from the *tximport()* function in bcbio, providing transcript level count estimates. A tumour outlier is observed at 2045. Corresponding count values are shown in the table on the right hand side. Values indicate the increase in counts observed in tumour samples relative to their paired normal samples.

When examining the top ten up- and down- regulated genes, it is interesting to note that aberrant expression of a number of these genes have also previously been associated with OSCC in literature. Li *et al* (2020) and Song *et al* (2021) reported that the expression of *COL10A1* was found to be significantly increased in OSCC tumour tissue compared to

matched normal tissue [440,441], while Liu *et al* (2016) [442] reported that *MMP1* expression is increased in OSCC and associated with poor outcomes through promotion of tumour growth and metastasis. This pattern of high expression and upregulation has also been reported for *MMP3* [443], *SERPINE1* [444], *TBFB1* [445], *MMP13* [446], and *POSTN* [447], all found to promote OSCC invasion and metastasis.

Of the list of the top downregulated genes, only *HCG22* has previously been reported to be associated with OSCC. Under normal conditions, *HCG22* inhibits tumour proliferation, invasion and migration. Li *et al* (2020) [448] described that in OSCC, it is remarkably downregulated, no longer exerting it's anti-tumour role.

These observations are all in keeping with current literature views and provide a further indication of the legitimacy of the bioinformatics analysis performed.

6.1.4.2.3 Functional Enrichment Analysis

In order to gain a greater insight into the list of DE genes, a functional enrichment analysis was performed using the R Markdown template available upon installation of the bcbioRNASeq [432] package in R. Interpreting gene lists and genome-wide regions of interest has become increasingly more prevalent in genomic research through the widely implemented functional enrichment analysis techniques [449]. The analysis of pathway enrichment is an essential step toward identifying logical themes that are most characteristic of high-throughput sequencing data [450].

The purpose of this analysis was to identify significantly enriched biological processes, molecular functions and cellular components classes among the list of DE genes. The Markdown template generates gene ID's for the list of DE genes previously obtained, as well as a background list of genes, using LFC values (generated in DGE analysis) for significant results. The *enrichGO()* function uses the significant genes list, the background gene list and the ontology to test as input to perform statistical enrichment analysis of gene ontology (GO) terms using hypergeometric testing. Category net-plots, dot plots and enrichment GO plots were produced in order to visualise the enriched processes identified.

When performing analyses for Biological Processes, Molecular Functions and Cellular components, we observed that our gene of interest, *MUC3A,* was only associated with the Molecular Functions ontology. Therefore, only plots from the Molecular Function analysis

are shown here. All plots from Biological Processes and Cellular Components are shown in Appendix 7. The dot plot of Molecular Function (Figure 6.13) shows the number of DE genes associated with the top 25 GO terms (size) and the *P*-adjusted values for these terms. The order of GO terms is based on the gene ratio (number of significant genes associated with the GO term / total number of significant genes). LFC threshold was increased to 2 in order to reduce the large number of DE genes and consider those that have a large (statistically significant) difference in expression between tumour and normal conditions. Thus we consider the GO functions of the important genes which generated clear plots.

The top GO enrichment terms were used to plot category net-plots, and the plot for Molecular Function (Figure 6.14) showed that *MUC3A* was significantly associated with the GO term 'extracellular matrix structural constituent'. The category net-plot shows the relationship between the genes associated with the top five most significant GO terms (Figure 6.13) and the LFC's (threshold set at 2) of the associated significant genes. The size of the GO term reflects the *P*-values of the terms, with the more significant terms being larger. The significant fold change of *MUC3A* can thus be viewed as an important factor in this functional enrichment analysis.

The GO term 'extracellular matrix structural constituent' is the sixth highest GO enrichment for the Molecular Function ontology, and the category net-plot shows links to 'signalling receptor regulator activity', 'receptor ligand activity' as well as 'signalling receptor activator activity'. It is not surprising to observe the other genes linked to the same gene ontology produce common extracellular matrix proteins. Eleven collagen genes are linked to this ontology (*COL1A1, COL1A2, COL3A1, COL4A1, COL5A1, COL5A2, COL6A3, COL10A1, COL11A1, COL12A1* and *COL22A1*). Studies investigating the collagen family of genes found that *COL1A1, COL10A1* and *COL11A1* showed upregulated expression in OSCC tissue compared to normal controls (J. Li et al., 2019), and increased expression of *COL1A2* and *COL11A1* promotes proliferation and metastasis in OSCC [452,453]. Furthermore, increased collagen content has been associated with MAPK and PI3K/Akt signalling pathways leading to chemotherapy resistance in OSCC [454].

Matrix metalloproteinases (MMPs) (associated with the 'metallopeptidase activity' gene ontology in Figure 6.14) along with the numerous collagen isoforms, laminins, glycoproteins and proteoglycans are known contributors to extracellular matrix stiffness in OSCC, and are strongly associated with important cellular events during tumour progression during invasion

and metastasis, activating oncogenic signalling pathways involved in cytoskeleton alterations during adhesion and migration [455]. Plausibly, *MUC3A* may play a similar role as an extracellular matrix structural constituent. This link with *MUC3A* and the cell signalling ontologies ties in with the discussion in section 1.3.2.1 suggesting that transmembrane mucins, including *MUC3A*, are involved in cell signalling through the phosphorylation of the intracellular cytoplasmic tails activating signal transduction pathways.

Gene Set Enrichment Analysis

A gene set enrichment analysis (GSEA) was also included in this functional enrichment analysis. GSEA is a computational method used to determine whether a pre-defined set of genes show any statistically significant concordant differences between two biological states, i.e. tumour vs. normal phenotypes (Mootha et al., 2003; A. Subramanian et al., 2005). The previous GO analyses were based on the list of DE genes in order to identify large differences in expression, but not necessarily small differences. The GSEA analysis addresses this limitation by aggregating per gene statistics across genes within a gene set, to detect all genes in a predefined set that change in a small but co-ordinated way. It is likely that many relevant phenotypic differences are established by small but consistent changes in a set of genes [432]. The basic assumption of GSEA is that although large changes in individual genes has significant implications on biological pathways, some weaker but co-ordinated changes in sets of genes that are seen to be functionally related (i.e. in related pathways) might also cause significant effects [437].

In this analysis, GSEA was performed using the clusterProfiler tool (T. Wu et al., 2021) together with the Kyoto Encyclopaedia of Genes and Genomes (KEGG) gene sets and the LFC measured previously as input to identify enriched biological pathways with genes that exhibit coordinated fold changes larger than can be expected by chance. The clusterProfiler library provides a set of functions to uncover biological functions and pathways through over-representation analysis, and makes use of GO and KEGG to support thousands of organisms.

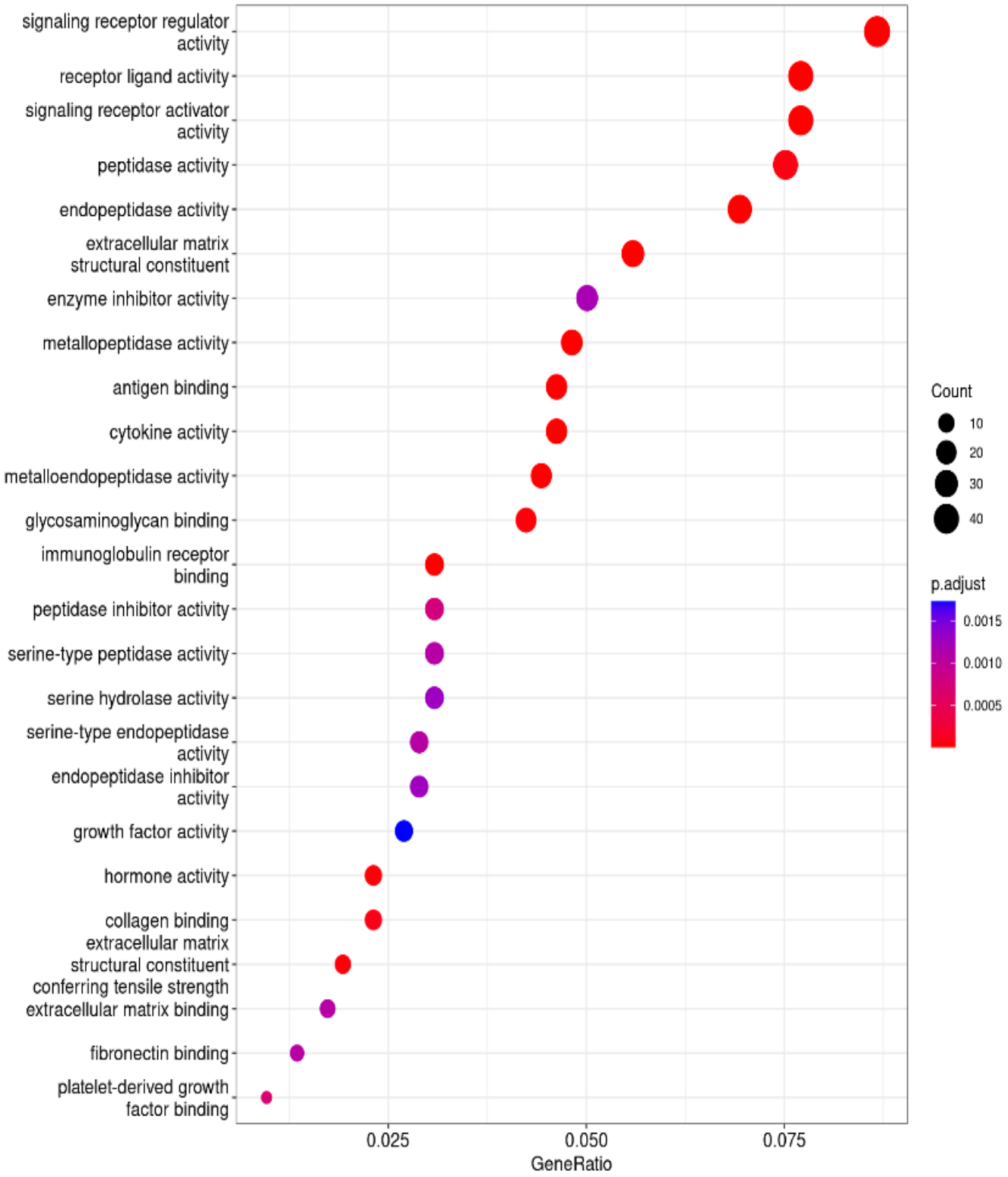


Figure 6.13: Dot plot showing the top 25 enriched GO terms for Molecular Function, where the LFC threshold was set at 2. The largest gene ratios are plotted in order of gene ratio while the size of the dots represents the number of genes in the significant DE genes list. Dot colour represents the *P*-adjusted values (BH).

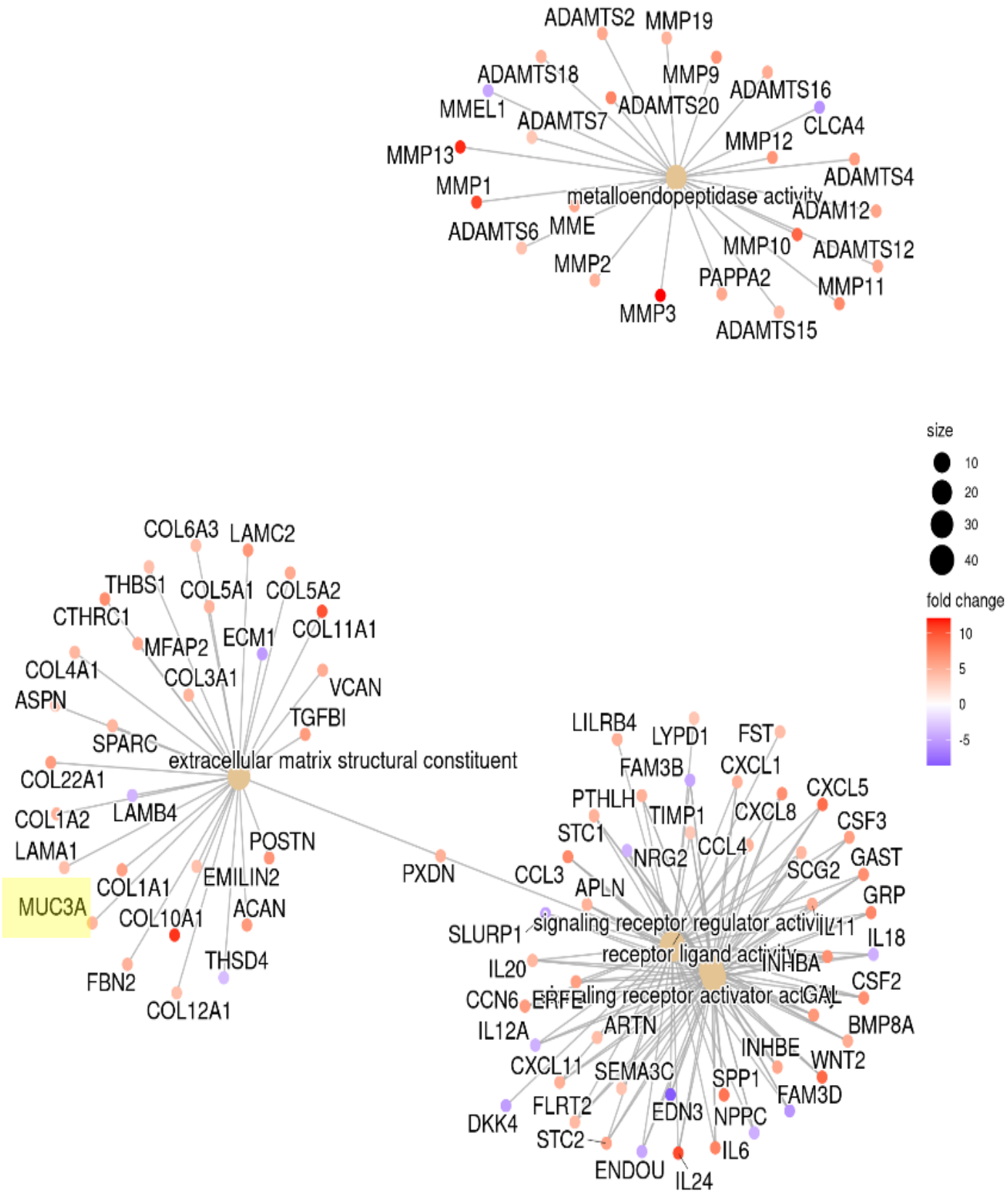


Figure 6.14: Category net-plot of the top most significant GO terms (*P*-adjusted values) plotted and connected with lines to associated DE genes, and where the LFC threshold was set at 2. Colours represent the fold changes of the significant genes associated with the GO terms while the size of the terms reflects the *P* values, with the more significant terms being larger. *MUC3A* associated with extracellular matrix structural constituent GO term.

GSEA tools make use of the ranked genes (ranked by LFC) without the use of a cut-off/threshold. In this way, tools are able to use more information to identify enriched pathways. By using LFC rankings, pathways involving genes exhibiting coordinated fold changes larger than expected by chance are identified. Significantly dysregulated pathways show a FDR (q-value) of <0.05. Enriched pathway images can be viewed at https://github.com/VictoriaPatten/phd-scripts/tree/main/kegg_pathways_images. Table 6.4 below shows the top ten KEGG enriched pathways detected in the analysis.

The top GSEA KEGG enriched pathways show a heavy leaning to cytokine/chemokine influence with IL-17 (interleukin-17) signalling pathways (hsa04657) as the most enriched in these OSCC patients. IL-17 is a well-known pro-inflammatory cytokine that has been proven to play an important role in numerous inflammatory and autoimmune diseases [458]. There is evidence that shows IL-17 induces OSCC tumour cells to produce increased amounts of chemokines (L. Lu et al., 2013) and in OAC, reactive oxygen species (ROS)-NF_KB signalling pathways are activated by IL-17, subsequently upregulating the expression of a number of matrix metallopeptidases (MMP's) to promote invasiveness [460–462]. In Table 6.4, a number of MMP genes are shown to be associated with IL-17 signalling pathway. Numerous CXC, CSF and CCL genes are listed which are all chemokines or cytokines commonly involved in inflammation and the immune response, as reflected in the enriched pathways listed.

Interestingly, two of the top enriched pathways implicate viral influences within the patients; these being 'Viral protein interaction with cytokine and cytokine receptor' (hsa04061), and 'Human cytomegalovirus infection' (hsa05163). Viral cytokine and cytokine receptors, together with structurally unique, soluble, cytokine-binding or cytokine-receptor-binding proteins present molecular strategies for subversion and alteration of host cytokine networks by large DNA viruses. This may lead to the activation or inhibition of cytokine signalling in the host and may affect aspects of host immunity [463,464].

The human cytomegalovirus (HCMV) is an enveloped, dsDNA virus known to cause significant morbidity and mortality in immunocompromised individuals. Reports have shown that HCMV can lead to the activation of the oncogenic PI3K/AKT pathway, catalysing oncogenesis. Furthermore, HCMV gene products and antiapoptotic proteins interfere with major histocompatibility complex (MHC) class 1 antigen presentation and avoid immune clearance of infected tumour cells [465,466].

Table 6.4 Top 10 GSEA KEGG enriched pathways.

KEGG ID	Enriched Pathway	P-adjusted value	q-value (FDR)	Genes
hsa04657	IL-17 signalling pathway	0,0005	0,0003	MMP3/MMP13/MMP1/CXCL5/IL6/CXCL8/ CSF2/PTGS2/MMP9/CSF3
hsa05323	Rheumatoid arthritis	0,0118	0,0091	MMP3/MMP1/CXCL5/IL6/CCL3/CXCL8/CSF2/ IL11/CXCL1/CD80/CXCL6/CTSL/TLR2/ ATP6V0D2
hsa04061	Viral protein interaction with cytokine and cytokine receptor	0,0235	0,0181	IL24/CXCL5/IL6/CCL3/CXCL8/CXCL11/ CCR3/CCL4/CXCL1/CCR1/IL20/CXCL6/ CXCL10/IL2RA/ACKR4/PPBP
hsa04620	Toll-like receptor signalling pathway	0,0235	0,0181	SPP1/IL6/CCL3/CXCL8/CXCL11/CCL4/CD80/ CXCL10/TLR2/TLR8/CD14/LBP/TLR4/ CTSK
hsa04926	Relaxin signalling pathway	0,0312	0,0239	MMP13/MMP1/GNGT1/MMP9/COL1A1/ COL1A2/MMP2/COL3A1/COL4A1
hsa04062	Chemokine signalling pathway	0,0321	0,0246	CXCL5/CCL3/GNGT1/CXCL8/CXCL11/CCR3/ CCL4/CXCL1/CCR1/CXCL6/CXCL10
hsa04974	Protein digestion and absorption	0,0321	0,0246	COL10A1/COL11A1/COL22A1/COL1A1/ MME/COL5A2/COL1A2/COL5A1/COL3A1/ COL4A1/SLC7A7/COL6A3/COL12A1/ SLC16A10/COL4A2/COL24A1/COL8A1
hsa05163	Human cytomegalovirus infection	0,0321	0,0246	IL6/CCL3/GNGT1/CXCL8/PTGS2/CCR3/ CCL4/CCR1
hsa05205	Proteoglycans in cancer	0,0321	0,0246	WNT2/MMP9/COL1A1/PLAUR/COL1A2 /HOXD10/MMP2/PLAU/CTSL/TLR2/SHH/ ITGA5/THBS1/LUM/GPC3/FN1/TLR4
hsa04060	Cytokine-cytokine receptor interaction	0,0321	0,0246	IL24/CXCL5/IL6/CCL3/CXCL8/CSF2/INHBA/ CSF3/INHBE/IL11/BMP8A/CXCL11/CCR3/ IL31RA/CCL4/CXCL1/CCR1/IL20/CXCL6/ CXCL10/IL13RA2/IL2RA/ACKR4/IL7R/PPBP/ XCL1/TNFRSF12A/TNFRSF17/BMP8B/ TGFB2/CCL11/IL12RB2/TNFRSF9/TNFRSF1B /LIF/CCR8/TNFRSF4/IL3RA/TNFSF13B/ TNFRSF8

The functional analysis results suggest genes and pathways that may be involved with the condition of interest, however, it is important to note that the results are not conclusive and all identified processes and pathways would require further experimental validation.

6.2 Immunohistochemical Analysis

Given the high somatic mutation rates observed in the WGS investigations together with the RNA-seq analysis that indicated a significantly increased gene expression in the patient cohort, further investigations were carried out to assess the protein levels by immunohistochemistry (IHC) staining. Given the WGS and DGE analysis results already obtained, we investigated MUC3A protein levels in order to gain a clearer insight.

The specific advantage of this technique over others is the ability to correlate specific antigens with certain cellular or tissue compartments thus aiding in functional analyses of pathological conditions [467]. It is frequently performed on formalin-fixed paraffin-embedded (FFPE) tissue which requires an antigen retrieval step to retrieve antigens that are masked by fixation thereby facilitating more accessible antibody binding [468]. This significantly increases IHC sensitivity allowing for a more expansive application of the technique [469,470].

Primary antibodies can be either monoclonal or polyclonal. Generally, monoclonal antibodies have a greater specificity, while polyclonal antibodies are more sensitive [471]. Quality control of the IHC staining is of critical importance and both positive and negative controls are routinely used [472]. Negative controls are run to eliminate any possibility of nonspecific antibody binding and consist of sample tissue stained with only secondary antibody, and no primary [472]. Visualisation of staining results is commonly performed under a light microscope for identification of the presence or absence, or the localisation of the molecule of interest.

6.2.1 Results: Immunohistochemistry for MUC3A

In this study, seventeen patient biopsies obtained at Groote Schuur Hospital in Cape Town had previously undergone histological testing and fixation in FFPE blocks for long term storage in the Department of Pathology at UCT. IHC staining was performed on all of the FFPE sample blocks with the exception of four samples that had been damaged whilst in storage. IHC staining was carried out in the Department of Pathology at UCT as described in section 2.2.6, using duodenum FFPE tissue as a positive control [473,474].

Standard IHC staining techniques were carried out as discussed in section 2.2.6, with the inclusion of antigen retrieval steps from the FFPE sections. IHC sections were examined under light microscopy. FFPE blocks for samples PD39451, PD39456, PD39457 and

PD39459 were unfortunately damaged during storage and were thus not available to be sectioned and stained. Of the remaining thirteen samples, three cases showed focal membrane and cytoplasmic staining within neoplastic squamous cells. Ten cases were negative for MUC3A IHC stain in the presence of positive external control which showed cytoplasmic and membrane positivity. Some of these ten cases showed very weak nuclear positivity within the squamous epithelium but this was interpreted as a negative staining. Positive IHC for MUC3A is recorded as cytoplasmic and/or membrane staining. Table 6.5 shows the positive and negative staining of the tumour biopsies.

When evaluating the thirteen stained samples, Table 6.5 shows that ten patients with putative *MUC3A* mutations in their WGS data stained negatively, and three stained positively, as shown in Figures 6.16 and 6.17 while Figure 6.15 shows duodenum staining as a positive control.

Table 6.5: MUC3A Staining results of FFPE sections of OSCC patients from Groote Schuur Hospital. Patients whose FFPE blocks had been damaged are highlighted in blue.

Patient Number	***MUC3A* mutations found in WGS data using Vardict**	***MUC3A* mutations found in WGS data using Mutect2**	**Staining Result**	**% Tumour Cells**
PD39445	Yes	Yes	Negative	n.d.
PD39446	No	Yes	Positive	n.d
PD39447	Yes	Yes	Negative	28
PD39448	Yes	Yes	Negative	44
PD39449	No	Yes	Positive	57
PD39450	No	Yes	Negative	64
PD39451	Yes	Yes	*No FFPE block*	
PD39452	Yes	Yes	Negative	64
PD39453	Yes	Yes	Negative	22
PD39454	Yes	Yes	Negative	n.d
PD39455	Yes	Yes	Negative	91
PD39456	Yes	Yes	*No FFPE block*	
PD39457	Yes	Yes	*No FFPE block*	
PD39458	Yes	Yes	Negative	30
PD39459	Yes	Yes	*No FFPE block*	
PD39460	Yes	Yes	Positive	66
PD51372	Yes	Yes	Negative	27

*n.d. = not determined

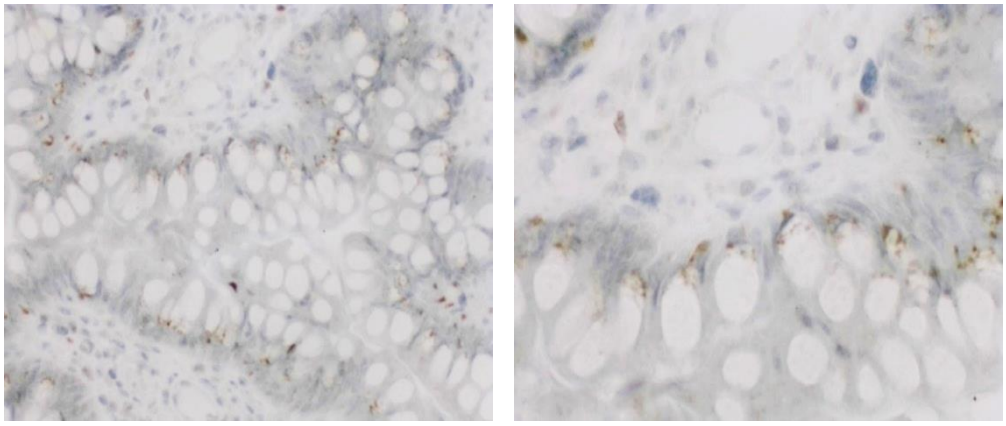


Figure 6.15: Positive cytoplasmic staining of duodenum tissue as a positive experimental control. 3-4µm slices of duodenum FFPE tissue were stained using a MUC3A primary antibody. Brown stains indicate positive staining for MUC3A.

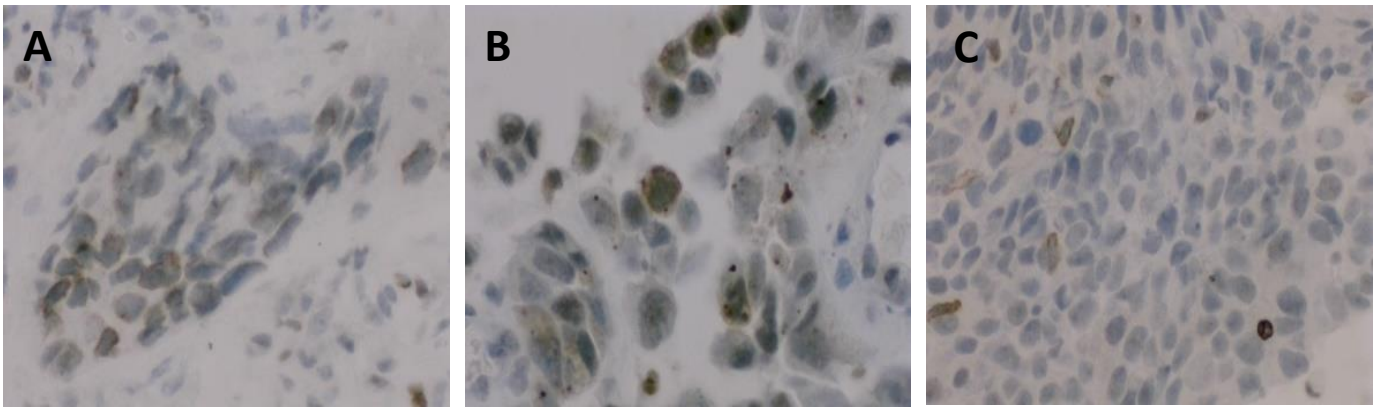


Figure 6.16: IHC on haematoxylin stained FFPE tissue sections of samples. **A** - **C** show positive staining of samples PD39446, PD39449 and P39460 respectively.

The positively stained OSCC samples showed concentrated focal membrane staining in some areas, as well as diffuse and cytoplasmic staining, while the control duodenum tissue staining was visible on one side of cells, possibly secretory vesicles. These positive staining differences are to be expected due to the nature of the different tissues and cells.

The IHC results show that in ten out of thirteen samples no expression of the MUC3A protein was detected via IHC staining, while the remaining three samples did show the presence of MUC3A protein. Given that we currently view the putative *MUC3A* mutations identified in

WGS as spurious, it is not surprising that the pattern of MUC3A staining doesn't seem to associate or correlate with the putative mutation profile as previously determined.

Ideally it would be useful to have a larger sample group to give a better reflection of the protein expression profile of this gene in this cohort. This small sample set showed that 10 out of 13 tumour samples showed no MUC3A protein upon IHC staining. This observation seemingly contradicts the significantly upregulated gene expression observed with DGE analysis.

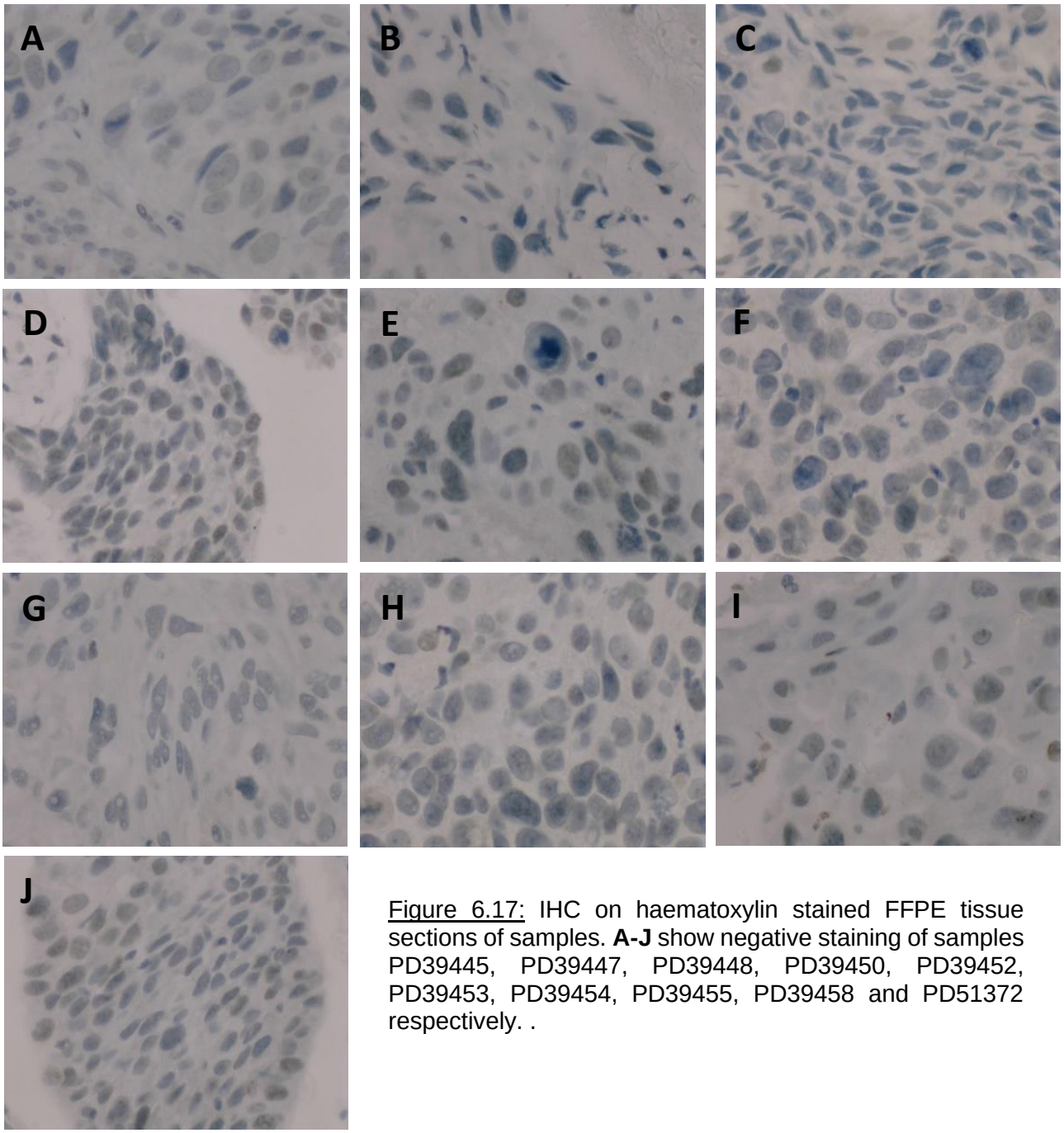


Figure 6.17: IHC on haematoxylin stained FFPE tissue sections of samples. **A-J** show negative staining of samples PD39445, PD39447, PD39448, PD39450, PD39452, PD39453, PD39454, PD39455, PD39458 and PD51372 respectively. .

6.2.2 Correlation of Putative *MUC3A* Mutations with Expression

Literature describes that aberrant and upregulated expression of *MUC3A* has been implicated in a variety of different cancers including breast, pancreatic, gastric, colorectal, renal and prostate cancers [257–261], although the exact mechanism of *MUC3A* upregulation remains unclear. MUC3A expression may play an important role in oncogenesis, particularly in metastasis, invasion and the progression of cancer [267]. Table 6.6 provides an overview of the different techniques used, indicating samples utilised in each analysis.

DNA sequence data from thirty five South African OSCC patients indicated putative *MUC3A* mutations in the entire cohort, however, based on our previous experience, we currently view these putative mutations as false positives given the high frequency of alterations, the observation that these specific alterations were not detected in the previous round of analyses, and that other authors had not made similar findings (especially given the high frequency). RNA from fifteen of the cohort samples was subjected to RNA-seq. When performing variant calling on this data, some moderate impact severity and two high impact severity (frameshift) *MUC3A* mutations were identified. It was subsequently determined however, that variant calling for small variants and indels using RNA-seq data is unreliable and inaccurate with the bcbio-nextgen software. Although, the inability to identify HIGH impact *MUC3A* mutations in the RNA-seq data is consistent with the view that the WGS analysis mutations were in fact false positives. Surprisingly though, the analysis of DGE indicated a highly significant increase in expression of *MUC3A* in tumour samples compared to their paired normal samples indicating some degree of dysregulation of the gene. IHC on FFPE sections gave the opposite results, with negative MUC3A protein staining in 10 out of 13 samples, indicating that although a significant increase in gene expression was observed at a transcript level, this increased expression is not reflected at the protein level. The role mechanism for this aberrant protein expression remains unclear.

Table 6.6: Overview of *MUC3A* analysis through WGS, RNA-seq analysis, DGE and IHC. Patients whose DNA was subjected to WGS are listed the first column, and the number of putative *MUC3A* mutations detected per patient using Vardict and Mutect2 respectively are listed in the second and third columns. The patients whose RNA was subjected to RNA-seq are listed in the fourth column, followed by the number and type of *MUC3A* variants detected in RNA variant calling in the fifth column. RNA-seq data used for DGE is indicated in the sixth column while FFPE blocks used for IHC are indicated in the seventh column with their respective staining results.

WGS Patients	# HIGH impact WGS MUC3A variants detected with Vardict	# HIGH impact WGS MUC3A variants detected with Mutect2	RNA-seq Patients	RNA-seq MUC3A variants	Used in DGE	IHC Stain
PD39445	5	11	NA	NA	NA	Negative
PD39446	0	6	NA	NA	NA	Positive
PD39447	5	18	NA	NA	NA	Negative
PD39448	18	13	NA	NA	NA	Negative
PD39449	0	18	NA	NA	NA	Positive
PD39450	0	18	NA	NA	NA	Negative
PD39451	5	7	NA	NA	NA	n.a.
PD39452	10	42	NA	NA	NA	Negative
PD39453	12	10	NA	NA	NA	Negative
PD39454	6	7	NA	NA	NA	Negative
PD39455	16	12	NA	NA	NA	Negative
PD39456	17	4	NA	NA	NA	n.a.
PD39457	13	10	NA	NA	NA	n.a.
PD39458	7	14	NA	NA	NA	Negative
PD39459	2	13	NA	NA	NA	n.a.
PD39460	7	15	NA	NA	NA	Positive
PD44691	15	10	NA	NA	NA	n.a.
PD44692	1	10	NA	NA	NA	n.a.
PD44693	9	2	NA	NA	NA	n.a.
PD44694	0	4	NA	NA	NA	n.a.
PD44695	2	10	NA	NA	NA	n.a.
PD44696	11	16	NA	NA	NA	n.a.
PD44697	15	13	PR44697	4 Moderate	Yes	n.a.
PD44698	7	9	NA	NA	NA	n.a.
PD44699	7	4	PR44699	7 Moderate	Yes	n.a.
PD44700	6	14	NA	NA	NA	n.a.
PD44701	2	16	PR44701	2 Moderate	Yes	n.a.
PD44702	6	24	PR44702	2 Moderate	Yes	n.a.
PD44703	7	23	PR44703	21 Moderate	Yes	n.a.
PD44704	8	16	PR44704	3 Moderate	Yes	n.a.
			PR44705	3 Moderate	Yes	n.a.
PD50649	11	2	PR50548	7 Moderate	Yes	n.a.
PD50650	11	6	PR50549	1 High, 1 Moderate	Yes	n.a.
PD50651	0	10	PR50550	1 Moderate	Yes	n.a.
			PR50551	12 Moderate	Yes	n.a.
PD50653	11	12	PR50552	5 Moderate	Yes	n.a.
PD51372	6	14	PR50553	NA	Yes	Negative
			PR50554	1 High, 6 Moderate	Yes	n.a.
			PR50555	NA	Yes	n.a.

*NA = not applicable

*n.a. = no paraffin section available

6.3 Discussion

Variant calling of RNA-seq data appears to be notoriously difficult with numerous issues plaguing alignment methods. These include high mapping error rates, low sensitivity, alignment biases and low mapping throughput [431]. The official documentation for bcbio-nextgen software describes variant calling with RNA-seq data as frequently problematic with a high false negative rate commonly reported. Specifically, variant calling for indels may be inaccurate and unreliable [354]. It was therefore no surprise that we were not able to identify the *MUC3A* mutations in the RNA-seq data that had been identified in the WGS analysis. Variant calling on RNA-seq data only identified two patients each with a single HIGH impact variant in the *MUC3A* gene. Some MED impact mutations were identified but did not match those found in WGS. However, when attempting to validate the WGS *MUC3A* mutations using the PON approach and Mutect2 variant caller (described in Chapter 5), all initial *MUC3A* mutations were filtered out and an entirely new set of mutations identified. This indicates that the initial mutations were in fact false positives. It is also highly likely that most, if not all of the new *MUC3A* mutations identified are also spurious. Thus the results we obtained when performing variant calling on RNA-seq data could either be due to the problematic and unreliable nature of the pipeline, or that the WGS *MUC3A* mutations were in fact false positives. For future work, it would be pertinent to investigate different software packages and pipelines more suitable and tailored for RNA-seq variant calling. Unfortunately this was not possible due to time constraints in this project. Thus the focus of the RNA-seq analysis was quickly changed to investigate the more common analysis methods of DGE and functional enrichment.

A subsequent bcbio-nextgen pipeline was established to perform RNA-seq alignment steps and generate QC and gene abundance information from the RNA-seq data. Using R statistical analysis software [433], this information from all patients was converted into a structured S4 object containing the necessary data for downstream analysis. R markdown templates specific to the bcbioRNASeq R package were utilised to perform QC, DGE and functional enrichment analysis. The QC output gave an indication of the quality of the RNA-seq data and implemented tests to identify whether the data was suitable for continuing with DGE analysis. Figure 6.1 confirms that the sequenced data shows high coverage and mapping rates well above the cut-off thresholds. This confirms the quality of the sequence data, indicating good sequencing with no presence of background artifacts or contamination. Exon vs intron mapping also confirmed the integrity of the sequencing with all samples

mapping well above the exon threshold and well below the intron threshold. A further assessment of sample quality comes from determining the number of genes detected relative to the number of mapped reads, where samples should ideally have similar numbers of features detected. Tumour samples are shown to all have an increased number of features detected than compared to their respective normal samples, and are all in a similar range to each other.

When assessing the PCA and ICA plots it was evident that in the initial dataset, three outliers were present in the group showing low correlation to their respective grouping and leading to a level of potential skewedness of the data. After these three patient outliers were removed (both tumour and paired normal), the PCA and ICA plots showed clear clustering of tumour and normal sample data suggesting a suitable and reliable dataset to use for DGE analysis. These QC measures all confirmed that the RNA-seq data was of high quality, with accurate mapping and clustering.

DGE was then assessed using the same S4 object utilised for QC with outliers removed from the dataset. Correctly identifying DGE between specific conditions is critical to the understanding of phenotype variation between the two [426]. 2421 DE genes were identified in the analysis, with 1564 genes shown to be upregulated and 857 genes downregulated when the LFC threshold was set at 1. Confidence in the list of DE genes produced was assessed through examining the tumour vs normal clustering. The heatmap shown in Figure 6.9 together with the PCA plot Figure 6.10 demonstrate clear clustering of tumour samples, well separated from the clustering of normal samples. This represents a major trend of the data confirming that differences in gene expression were accurately detected between the two phenotypes and the trend was represented across all samples in each group. Furthermore, the robustness of the LFC values reported can be used as an indication of the differential expression effects between the phenotypes.

When examining the list of DE genes, a number of genes were identified that have previously been associated with OSCC. This finding serves as a control of sorts further verifying the robustness of the analysis in keeping with literature observations. When looking specifically at *MUC3A,* expression was significantly increased with a 4.6 log2 fold increase from the normal (24.25 fold change), and a *P*-adjusted value of $7.05e^{-6}$. This upregulation falls within the top third of all upregulated genes detected. Furthermore, when assessing the transcript level count estimates for each patient generated in the initial bcbio-nextgen pipeline analysis

using the *tximport()* function, tumour samples displayed far greater transcript counts when compared to their paired normal samples as shown in figure 6.12. This further confirms the observations of increased *MUC3A* gene expression in tumour samples. These results suggest that even though the putative WGS *MUC3A* mutations identified are likely spurious and we cannot link them to the DGE findings, this gene is indeed upregulated in the patient cohort. Ideally, it would be beneficial to confirm these results with PCR and qRT-PCR, however, as described previously, the *MUC3A* gene is notoriously difficult to PCR due to the high degree of tandem repeats in the genomic sequence and presents a difficult limitation to the study. It is important to note the recurrent elevation of *MUC3A* expression levels in tumour samples compared to paired normal samples. This internal control provides substantial weight to the finding of *MUC3A* upregulation and suggests that there may be some involvement of this gene in this patient cohort. Many genes are upregulated in cancer whose exact roles are not always well understood, and *MUC3A* is one of them.

Functional enrichment analysis was carried out utilising the list of DE genes. The purpose of this analysis was to bring to light the functional implications of DGE in these OSCC patients. The GO analysis of biological processes, molecular functions and cellular components identified the top most enriched GO terms among the samples and the number of genes associated with these terms (Figures 6.12-6.14). Interestingly, *MUC3A* was associated with the 'extracellular matrix structural component' GO term of the Molecular Functions ontology class, further cementing the idea that *MUC3A* might have an interesting role in these OSCC patients. This GO terms was also linked to a number of collagen (COL) genes which have been shown to promote proliferation and metastasis in cancers including OSCC. In support of the literature [451–453], the expression of the COL genes were found to be increased in this patient cohort. A number of MMP genes also showed significantly increased expression along with laminins, glycoproteins and proteoglycan genes known to contribute to extracellular matrix stiffness and tumour progression through the activation of oncogenic signalling pathways [455]. Thus the extracellular matrix in these OSCC appears to be highly affected, and a mutated, upregulated *MUC3A* gene could possibly play a similar role in signalling dysregulation.

A GSEA analysis was carried out using the KEGG library gene set to identify significantly enriched pathways influenced by smaller co-ordinated changes in related genes. The IL-17 signalling pathway was revealed to be the most enriched pathway. The general trend points to the enrichment of pathways involved in inflammation, and cellular regulatory mechanisms

(IL-17 pathway, Toll-like receptor pathway, chemokine and cytokine pathways and proteoglycans in cancer) and suggest a pro-inflammatory environment exists in this OSCC patient cohort. Chronic inflammation is known to play a pivotal role in pathogenesis and development of malignant disease [462].

Furthermore, the enrichment of two viral associated pathways (viral protein interactions with cytokine receptor and human cytomegalovirus infection) is of interest. These viral pathways suggest that the patients in the cohort have experienced viral infections influencing cytokine and cytokine-receptor interactions, affecting the immune response and possibly the activation of oncogenic signalling pathways, further contributing to the overall tumour microenvironment.

IHC staining for *MUC3A* was performed on thirteen OSCC biopsies set in FFPE blocks. The resulting stain images showed clear positive staining in three samples and negative staining in ten samples. The role and mechanism of the reduced MUC3A protein levels in the sample cohort is currently unclear and requires further investigation.

DGE analysis of RNA-seq data found that the transcription of *MUC3A* gene was significantly increased in tumour samples. One can speculate that although RNA levels are upregulated, translation of the RNA either results in a protein product which is rapidly degraded, or a truncated protein which corresponds to the extracellular domain of MUC3A which is released from the cell. This truncated protein could bind to EGF receptor (as discussed in section 1.3.2). This in turn could lead to the activation of signalling cascades, enhancing cell regulatory mechanisms [178,184] and providing a positive feedback loop for increased and sustained transcription. However, these suggestions remain speculative.

Although these techniques together suggest an interesting possibility regarding *MUC3A* and its involvement in OSCC, there still remains some degree of uncertainty. With such a high proportion of the cohort still presenting with putative *MUC3A* mutations when using the PON approach with Mutect2 variant caller, it is perplexing that these observations were not detected in previous studies in Asia and other African studies given the high incidence of OSCC in these regions, and gives rise to the question of validity of these results. The initial *MUC3A* mutations detected when using Vardict variant caller have been completely eliminated as false positives due to the inability to confirm their presence with PCR and their complete absence when using the PON approach analysis. The new set of mutations identified however are also to be viewed with a high degree of caution as these may also

be spurious given the inconsistencies with individual variant callers and the complexity of the *MUC3A* genomic sequence. Furthermore, a search of the TCGA and COSMIC databases showed no reported *MUC3A* mutations in OSCC, and only three *MUC3A* mutations reported in OAC. We suggest that in future, multiple variant callers should be incorporated into the analysis pipeline to integrate the results and improve performance. Continued and further analyses into a larger South African cohort and other populations would be beneficial. It is also pertinent for future investigations to note the complexity and difficult nature of the *MUC3A* gene, and that all mutation analyses results should be interpreted with caution.

Chapter 7: Conclusions

7.1 Novel Viral Insertions

Viruses are constantly bombarding the human genome leading to the development of several diseases as a consequence of acute infection [53]. With many families of viruses capable of integrating their DNA into the human genome, it is speculated that between 8% and 40% of the human genome may therefore house these hidden potential carcinogens [53,56,57]. Statistical analyses have indicated causal associations between several viruses and cancer [309], and it is now widely accepted that viruses are directly responsible for a number of human cancers with estimates suggesting that up to 20% of all human cancers have a viral aetiology [72,73]. The aim of this study was to investigate whether any novel viral insertions might be present in South African OSCC patients which could be associated with the development of this disease.

While there initially appeared to be interesting observations of insertions of *Autographa Californica* Nucleopolyhedrovirus in the preliminary WGS data, these results could not be confirmed by PCR analysis. Subsequent reanalysis of the WGS data using different analytical tools eliminated the presence of *Autographa Californica* Nucleopolyhedrovirus, *Emiliana Huxleyi Virus 86*, the *Gryllus Bimaculatis Nudivirus* or the *Trichodisplasia Spinulosa* Associated Polyomavirus. Although this evidence shows these viruses are not associated with OSCC, another family of exogenous viral genes, Human Endogenous Retroviruses (HERV's) were more common.

One can speculate that with a larger sample cohort and a robust bioinformatics pipeline, novel viral integrations may possibly be identified. As discussed in section 1.2, many viruses have the ability to integrate their DNA into host genomes through viral integration [55] thus it is quite plausible that novel viral integrations exist and are yet to be discovered.

After completing the functional enrichment analysis of KEGG pathways (Chapter 6) on fifteen RNA-seq data samples from a larger OSCC patient cohort, it was shown that two viral associated pathways (viral protein interactions with cytokine receptor and human cytomegalovirus infection) were enriched in these patients. This finding suggests that there might be a role for viral integrations in OSCC.

We accept that the small sample size used in this analysis made it difficult to identify viral insertions which could be involved in the development of OSCC. However, at the time this study was carried out, the inconclusive results and time constraints led to a re-focussing of the project objectives and investigations were redirected to HERV's and their possible link to somatic mutations. These investigations were initially carried out on the three pilot sets of WGS data and later on a larger sample cohort.

7.1.1 Limitations and Future Directions

This exploratory study experienced limitations due to unreliable bioinformatics pipelines used for the initial analysis. Much time was spent on PCR analysis when it was later discovered that the viral sequences not present in the genomes analysis, or at least not at the sites indicated by the bioinformatics analysis. It was interesting to note that PCR-amplified DNA sequences mapped to bacterial cloning vectors which raised the question of how these were picked up in the bioinformatics analysis and should raise caution for future investigations.

A further limitation experienced was the reliance on external bioinformatics analysis with no knowledge of the pipelines or processes involved. It was only after new collaborations were formed that a more hands-on approach and deeper understanding of the requirements and objectives of the analyses was acquired. In future, it may be interesting to revisit investigations into novel viral insertions with the development of suitable analysis pipelines. This however fell outside of the time and scope of this particular thesis.

7.2 Human Endogenous Retroviruses

It is well known that endogenous retroviruses became integrated into the human genome millions of years ago and are now stably integrated in the genome and part of our inherited genetic material [78–81]. A number of HERV's are implicated in a variety of different cancers such as breast cancer [96], lung cancer [97], prostate cancer [98], hepatocellular carcinoma [99], melanomas [100], germ cell tumour [101], leukaemia [102], and lymphoma [103]. Investigations into HERV-K integration in South African OSCC patients was performed to determine where these Transposable Elements are integrated and whether differences exist between tumour and matched normal samples. Tumour tissue had increased numbers of insertions

compared to normal tissue, suggesting duplication and/or translocation of integrated DNA, which aligns with the literature reports that increased expression of HERV elements is associated with numerous cancers [104–106]. We were able to identify a number of upstream and downstream genes in close proximity to the HERV insertion sites and although a number of the genes identified were previously reported to be involved in OSCC, when performing variant calling on the WGS data, only four patients showed medium impact severity mutations and we were unable to conclusively form any links between the HERV insertions and the presence of somatic mutations in close proximity to these genes.

Thus while HERV-K insertions in these patients cannot be linked to somatic mutations that influence OSCC development or progression, it was still interesting to observe the differences in number of insertions in tumour tissue compared to their matched normal tissue, confirming the retroviral movement within the human genome, and it is possible that they may still be the cause of host genome instability [108]. The disruption of host gene regulation through hijacking and manipulation by retroviral elements can influence the expression of host genes, leading to long-lasting effects on the genome [57], and it is speculated that HERV's might transform benign cells and induce cancer though several other mechanisms, not within the scope of this thesis.

7.2.1 Limitations and Future Directions

Due to COVID-19 lockdown and regulations, extensive delays in sequencing and data analyses were experienced. In future it would be interesting to delve further into the role that translocated/duplicated HERV insertions may play in OSCC, and whether they influence the activation of immune-suppressor pathways, homologous recombination, or promote oncogenes aiding propagation and progression of tumorigenesis.

7.3 MUC3A in Oesophageal Squamous Cell Carcinoma

Initial bioinformatics analysis of WGS data yielded unexpected results where 258 somatic variants were detected in the *MUC3A* gene in thirty out of thirty-five South African OSCC patients. This gene has not previously been associated with OSCC, however it has been implicated in a number of other cancers where upregulated expression has been associated with poor prognosis and a decrease in overall survival [257–259,405]. This gene was the most

mutated gene in the patient cohort with 96% of the variants reported to be frameshift in nature. The fact that we were unable to validate these mutations in the laboratory raised questions around the robustness and accuracy of the bioinformatics pipeline used for analysis, as well as the original whole genome sequencing. The *MUC3A* gene is very large and the genomic sequence contains variable tandem repeats that may influence the mapping of the sequenced reads leading to incorrect variant calling. This highly repetitive nature of the gene impacted on PCR primer design and specificity. Gum *et al* (1997) [251] stated that the long stretches of tandem repeats make the cloning of *MUC3A* 'extraordinarily difficult', and PCR seemed to face the same problem. We must also consider the possibility of the mutations only occurring on one allele which was preferentially amplified during PCR reactions, and/or the presence of false positive results generated by the variant caller used in the analysis pipeline. As primer optimisation was successful for only two clusters of mutations, we were unable to validate all 258 variants that were detected using PCR as a validation tool, thus we cannot be certain of the degree of possible false positivity. Different variant callers have been reported to perform differently for various datasets and levels of genomic complexity. Different sequencing properties at individual variant sites, including read depth, read quality, strand bias and varying allele fractions frequently challenge the ability of variant callers to detect mutations accurately and increasingly complex data makes the detection of true positives and minimising false positives more difficult [413]. Two separate bcbio-pipelines were built and samples were run through each, producing the same output. Furthermore, two separate filtering tools were used (GEMINI and SnpSift), with each indicating the same high level of *MUC3A* mutation. Thus investigations continued with the analysis of RNA-seq data. While we were comfortable with the bcbio-nextgen pipeline as a whole, the question of the validity of the mutations due to the complexity of the gene was still pertinent. Using the IGV tool we manually examined the target genomic location (chromosome 7, *MUC3A* gene, exon2) and found a high degree of noise in both tumour and normal samples. This prompted the re-analysis of the data using a Panel of Normals approach with the Mutect2 variant caller. The study by Bian *et al* (2018) [413] revealed that Mutect2 was one of the top performing variant callers in their comparative study. The new bcbio-nextgen analysis using the PON approach filtered out all previous *MUC3A* mutations identified using Vardict, and instead showed an entirely new set of *MUC3A* mutations. A number of the new mutations were tagged with dbSNP accession numbers and are reported on the dbSNP database. Although there is the possibility that these new mutations are real,

the vast number of mutations identified and exact repetitions between patients, together with the high level of complexity of the genomic region do not inspire confidence in the quality of the data. There is a strong likelihood that these new mutations are also false positives and extensive laboratory validation would be necessary to conclude their validity with certainty.

RNA sequencing was performed on fifteen of the patients in the cohort. Initially, we attempted to perform variant calling using bcbio-nextgen software, however some evidence indicates that this is unreliable and imprecise for indel calling [354] and one should rather perform DGE analysis for better consideration of the RNA data. This was reflected in the results we obtained for variant calling where only two high and 78 medium impact severity variants were detected, none of which corresponded to the WGS variants.

DGE analysis was performed on the RNA-seq data using bcbio-nextgen in conjunction with bcbioRNASeq R software package [429,432]. Assessment of the quality of the data was performed using the QC Markdown template and this provided an indication that the data was of high quality. Interestingly, three tumour samples (PD50555, PD44701 and PD44704) were observed to group more closely to the normal samples. Upon inspection of the initial percentage tumour purity (section 2.1.1.1, Table 2.1) we noted that these three patients all presented with low tumour content. These outliers were subsequently removed from the dataset. DGE of the amended dataset showed 2421 upregulated genes and 857 downregulated genes. *MUC3A* was significantly upregulated with a log2 fold change of 4.6 (24.25 times higher than expression in normal samples) with a *P*-adjusted value of $7.05e^{-6}$. This observation echoed literature reports of upregulated expression of *MUC3A* in multiple cancers [256–258].

The functional enrichment analysis that followed DGE highlighted the pro-inflammatory status of the patients, as predominantly pro-inflammatory chemokine and cytokine influencing pathways and genes were enriched in the samples. *MUC3A* was significantly associated with the ‘extra cellular matrix structural constituent’ Gene Ontology of the Molecular Function class which was linked to a number of collagen genes that are well documented to play a role in tumour proliferation and metastasis, and influencing various signalling pathways. MMP genes were also upregulated and associated with the Molecular Functions ontology class. Thus conceivably, *MUC3A* may play a similar role in influencing the extra cellular matrix and signalling pathways in these patients.

While expression analysis results aligned with what has been reported in the literature regarding *MUC3A* gene expression in cancer, it was pertinent to investigate the protein levels to further characterise these findings. IHC was performed on thirteen FFPE sections, ten of which stained negatively for MUC3A protein. This altered protein expression seemingly contradicts the highly significant upregulation of the gene seen at a transcript level. Taken together, we observed an over expression of the gene and transcript production, but no overexpression based on protein levels. We thus speculate that the MUC3A gene may in fact be dysregulated, but the exact role of this altered MUC3A protein expression is unclear and requires further investigation. Ideally, a repeat of the immunohistochemistry staining with a larger cohort would be beneficial.

We can speculate that the increased transcription may be due to production of a truncated protein that structurally, may expose the EGF-like domains for binding to various receptors, and/or the SEA domain for autoproteolysis that can then affect cell migration and invasion through the phosphorylation of HER/ErB2, affecting the PI3K/Akt signalling pathway [260,265]. Shedding of the extracellular domain of MUC3A protein can lead to signalling pathway activation linked to proliferation and apoptosis [184], thus it is possible that a truncated protein may act in a similar way to a 'post-shedding' protein, with EGF-like domains interacting with various receptors, leading to phosphorylation of the cytosolic tail and initiating cell growth and proliferation pathways [181,184,199]. It has been suggested that mutations in *MUC3A* may be associated with its expression [450], and a study investigating gastric cancer showed that recurrent mutations in *MUC3A* were associated with increased risk of tumorigenesis. It was also speculated that mutations in the *MUC3A* gene are highly likely to change the extracellular structure of the protein, or the structure of the glycosylation site, thereby affecting it's barrier function efficacy [475].

MUC1 and *MUC4* are the most widely studied mucin genes and since they are transmembrane mucins like *MUC3A*, we could speculate that *MUC3A* might function in a similar fashion with regards to intracellular signalling. Through interactions with the EGF receptor, MUC1 has been shown to activate JAK/STAT3 signalling as well as NF_KB and Wnt signalling while MUC4 interacts with HER2 to influence the PI3K/Akt pathway. An exposed EGF-like domain on a truncated MUC3A protein therefore could activate these cell regulatory signalling pathways and contribute to OSCC tumorigenesis and tumour progression in a continuous loop together with the upregulated transcription of the *MUC3A*

gene in a microenvironment that is already in a state of chronic inflammation as shown in the pathway enrichment analysis.

The results from the DGE analysis and functional enrichment analysis do show that in this sample cohort, *MUC3A* is significantly differentially upregulated. Furthermore, when assessing IHC staining, 77% of the samples stained negatively for the MUC3A protein, albeit in a small sample size. Taken together, these results suggest that the *MUC3A* gene is dysregulated in this patient cohort.

7.3.1 Limitations and Future Directions

This study evolved as it progressed and a number of limitations were experienced.

The value of WGS data is heavily dependent upon accurate analysis and identification of mutations, thus it would be pertinent to include multiple variant callers in the bioinformatics pipeline when analysing WGS data in future to minimise false positives detected and improve the overall performance.

It would also be more ideal to have performed RNA sequencing on all thirty five patients rather than only fifteen patients, thus increasing the sample size for DGE and functional enrichment analyses. Similarly, it would have been beneficial to have access to the DNA from all thirty-five patients for PCR analysis, and all thirty-five FFPE blocks for IHC staining. Unfortunately this was not possible as the DNA from patients recruited from WITS/Charlotte Maxeke Hospital were not shipped due to COVID lockdown regulations. FFPE fixation of tumour biopsies was also only performed on biopsies recruited from patients at Groote Schuur Hospital in Cape Town. Performing these analyses on only a subset of the total cohort was definitely a limitation to the study.

A further limitation was that a number of patients displayed low tumour purity. This was detected in the QC of RNA-seq data and patients were removed from the dataset, however, all of the patients were included in WGS analyses regardless of tumour purity. Tumour content from three patients was reported as ‘not determined’ and a further 8 patients showed a tumour content of below 30%. It is possible that this may have affected the WGS analysis and contributed to false positive variant discovery.

In future, it would be beneficial to expand the investigations to include a much larger patient cohort, as well as to include a more diverse population for comparisons and indication of

prevalence. It would also be pertinent to test different variant callers with the bcbio-nextgen software and perform all analyses on all patients.

7.4 Final Remarks

While the viral insertion investigations proved inconclusive, it may be useful to revisit the hypothesis with a more robust bioinformatics pipeline, in a larger patient cohort.

While no links with HERV-K integrations and somatic mutations could be made in these OSCC patients, it was interesting to observe the difference in number of insertions between tumour and matched normal samples, suggesting further investigations into HERV translocation and duplication in the genome could be fruitful.

Although we initially detected a high degree of somatic mutations in the *MUC3A* gene, these mutations proved to be false positives when reanalysing the data together with a PON approach. An entirely new and different set of *MUC3A* mutations were identified with the PON approach and Mutect2 variant caller, however, we view these with caution and conclude they were probably false positives as well given the complexity of the genomic structure and the variation in variant caller accuracy. Further laboratory confirmation would be needed to confirm these mutations. However, the significant upregulation of *MUC3A* observed in DGE analysis and the contradictory IHC results suggest that *MUC3A* is dysregulated in this patient cohort in some way. We feel that the literature surrounding *MUC3A* is missing key information regarding the complexity of this gene and the difficulties that lie in its analysis. It is pertinent for future investigators to be aware of these difficulties and the challenging nature of the genomic sequence when performing future analyses. A number of limitations were experienced in this investigation and future studies should take care to address these issues in their endeavours to elucidate the true role of *MUC3A* in OSCC. We suggest that future investigations incorporate multiple different variant callers into their analysis pipelines and integrate the results to improve performance. In depth laboratory confirmations methods should be employed to validate any mutations detected although this may be complicated by the complexity of the genomic sequence of this particular gene.

References

1. Bray, F., Laversanne, M., Weiderpass, E. & Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029–3030 (2021).

2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209–249 (2021).

3. Huang, F. L. & Yu, S. J. Esophageal cancer: Risk factors, genetic association, and treatment. *Asian J Surg* **41**, 210–215 (2018).

4. Thrift, A. P. Global burden and epidemiology of Barrett oesophagus and oesophageal cancer. *Nature Reviews Gastroenterology & Hepatology 2021 18:6* **18**, 432–443 (2021).

5. Lagergren, J., Smyth, E., Cunningham, D. & Lagergren, P. Oesophageal cancer. *The Lancet* **390**, 2383–2396 (2017).

6. Smyth, E. C. *et al.* Oesophageal cancer. *Nat Rev Dis Primers* **3**, (2017).

7. Ilson, D. H. & van Hillegersberg, R. Management of Patients With Adenocarcinoma or Squamous Cancer of the Esophagus. *Gastroenterology* **154**, 437–451 (2018).

8. Abnet, C. C., Arnold, M. & Wei, W. Q. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* **154**, 360–373 (2018).

9. Kim, J. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169 (2017).

10. Uhlenhopp, D. J., Then, E. O., Sunkara, T. & Gaduputi, V. Epidemiology of esophageal cancer: update in global trends, etiology and risk factors. *Clin J Gastroenterol* **13**, 1010–1021 (2020).

11. Chen, W., Zheng, R., Zeng, H., Zhang, S. & He, J. Annual report on status of cancer in China, 2011. *Chinese Journal of Cancer Research* **27**, 2 (2015).

12. Pickens, A. & Orringer, M. B. Geographical distribution and racial disparity in esophageal cancer. *Ann Thorac Surg* **76**, S1367–S1369 (2003).

13. Thrumurthy, S. G., Chaudry, M. A., Thrumurthy, S. S. D. & Mughal, M. Oesophageal cancer: Risks, prevention, and diagnosis. *The BMJ* **366**, (2019).

14. Xie, S. H. & Lagergren, J. Risk factors for oesophageal cancer. *Best Pract Res Clin Gastroenterol* **36–37**, 3–8 (2018).

15. Arnold, M., Soerjomataram, I., Ferlay, J. & Forman, D. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut* **64**, 381–387 (2015).

16. Asombang, A. W. *et al.* Systematic review and meta-analysis of esophageal cancer in Africa: Epidemiology, risk factors, management and outcomes. *World J Gastroenterol* **25**, 4512 (2019).

17. Ferlay, J. *et al.* Global cancer observatory: cancer today. *Lyon, France: international agency for research on cancer* **3**, 2019 (2018).

18. Arnold, M. *et al.* Obesity and the incidence of upper gastrointestinal cancers: An ecological approach to examine differences across age and sex. *Cancer Epidemiology Biomarkers and Prevention* **25**, 90–97 (2016).

19. Qin, J. *et al.* Factors associated with overall survival and relief of dysphagia in advanced esophageal cancer patients after 125I seed-loaded stent placement: a multicenter retrospective analysis. *Diseases of the Esophagus* **32**, 1–8 (2019).

20. Gavin, A. T. *et al.* Oesophageal cancer survival in Europe: A EUROCARE-4 study. *Cancer Epidemiol* **36**, 505–512 (2012).

21. Li, X. *et al.* Systematic review with meta-analysis: the association between human papillomavirus infection and oesophageal cancer. *Aliment Pharmacol Ther* **39**, 270–281 (2014).

22. Lin, D. C. *et al.* Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet* **46**, 467–473 (2014).

23. Zhang, L., Wu, J., Ling, M. T., Zhao, L. & Zhao, K. N. The role of the PI3K/Akt/mTOR signalling pathway in human cancers induced by infection with human papillomaviruses. *Mol Cancer* **14**, 1–13 (2015).

24. Pai, S. I. & Westra, W. H. Molecular Pathology of Head and Neck Cancer: Implications for Diagnosis, Prognosis, and Treatment. *Annu Rev Pathol* **4**, 49 (2009).

25. Testa, U., Castelli, G. & Pelosi, E. Esophageal Cancer: Genomic and Molecular Characterization, Stem Cell Compartment and Clonal Evolution. *Medicines* **4**, 67 (2017).

26. Du, P. *et al.* Comprehensive genomic analysis of Oesophageal Squamous Cell Carcinoma reveals clinical relevance. *Scientific Reports* **7**, 1–9 (2017).

27. Liu, X. *et al.* Genetic Alterations in Esophageal Tissues From Squamous Dysplasia to Carcinoma. *Gastroenterology* **153**, 166–177 (2017).

28. Su, P. *et al.* Identification of the Key Genes and Pathways in Esophageal Carcinoma. *Gastroenterol Res Pract* **2016**, (2016).

29. Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature 2014 509:7498* **509**, 91–95 (2014).

30. Naseri, A. *et al.* Systematic Review and Meta-analysis of the Most Common Genetic Mutations in Esophageal Squamous Cell Carcinoma. *J Gastrointest Cancer* **1**, 1–10 (2021).

31. Odera, J. O. *et al.* Esophageal cancer in Kenya. *Am J Dig Dis (Madison)* **4**, 23 (2017).

32. Higginson, J. & Oettlé, A. G. Cancer Incidence in the Bantu and “Cape Colored” Races of South Africa: Report of a Cancer Survey in the Transvaal (1953–55). *JNCI: Journal of the National Cancer Institute* **24**, 589–671 (1960).

33. Frances Rose, E. Esophageal Cancer in the Transkei: 1955–69. *JNCI: Journal of the National Cancer Institute* **51**, 7–16 (1973).

34. Blot, W. J. Esophageal cancer trends and risk factors. *Semin Oncol* **21**, 403–410 (1994).

35. Lagergren, J., Bergström, R., Lindgren, A. & Nyrén, O. Symptomatic Gastroesophageal Reflux as a Risk Factor for Esophageal Adenocarcinoma. *https://doi.org/10.1056/NEJM199903183401101* **340**, 825–831 (1999).

36. Sammon, A. M. Carcinogens and endemic squamous cancer of the oesophagus in Transkei, South Africa. Environmental initiation is the dominant factor; tobacco or other carcinogens of low potency or concentration are sufficient for carcinogenesis in the predisposed mucosa. *Med Hypotheses* **69**, 125–131 (2007).

37. Sewram, V., Sitas, F., Oconnell, D. & Myers, J. Diet and Esophageal Cancer Risk in the Eastern Cape Province of South Africa. *http://dx.doi.org/10.1080/01635581.2014.916321* **66**, 791–799 (2014).

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

39. Sumeruk, R., Segal, I., te Winkel, W. & van der Merwe, C. F. Oesophageal cancer in three regions of South Africa. *S Afr Med J* **81**, 91–93 (1992).

40. Dlamini, Z. & Bhoola, K. Esophageal cancer in African blacks of Kwazulu Natal, South Africa: an epidemiological brief. *Ethn Dis* **15**, 786–789 (2005).

41. Pacella-Norman, R. *et al.* Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans. *British Journal of Cancer 2002 86:11* **86**, 1751–1756 (2002).

42. Sitas, F. *et al.* The relationship between anti-HPV-16 IgG seropositivity and cancer of the cervix, anogenital organs, oral cavity and pharynx, oesophagus and prostate in a black South African population. *Infect Agent Cancer* **2**, 1–9 (2007).

43. 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 Epidemiol* **41**, 113–121 (2016).

44. Koonin, E. V, Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biol Direct* **1**, 29 (2006).

45. Forterre, P. & Prangishvili, D. The origin of viruses. *Res Microbiol* **160**, 466–472 (2009).

46. Berliner, A. J., Mochizuki, T. & Stedman, K. M. Astrovirology: viruses at large in the universe. *Astrobiology* **18**, 207–223 (2018).

47. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun* **8**, 1–12 (2017).

48. Emerman, M. & Malik, H. S. Paleovirology—Modern Consequences of Ancient Viruses. *PLoS Biol* **8**, e1000301 (2010).

49. Koonin, E. v., Dolja, V. v. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).

50. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).

51. Greene, S. E., Reid, A. & Francisco, S. *Viruses throughout life and time: Friends, Foes, Change Agents*. www.asmscience.org (2013).

52. Elena, S. F. Evolutionary transitions during RNA virus experimental evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150441 (2016).

53. Cantalupo, P. G., Katz, J. P. & Pipas, J. M. Viral sequences in human cancer. *Virology* **513**, 208–216 (2018).

54. Nelson, P. N. *et al.* Human endogenous retroviruses: transposable elements with potential? *Clin Exp Immunol* **138**, 1–9 (2004).

55. Katzourakis, A. & Gifford, R. J. Endogenous Viral Elements in Animal Genomes. *PLoS Genet* **6**, e1001191 (2010).

56. Bannert, N. & Kurth, R. The Evolutionary Dynamics of Human Endogenous Retroviral Families. *Annu Rev Genomics Hum Genet* **7**, 149–173 (2006).

57. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* **13**, 283–296 (2012).

58. Stetson, D. B. & Medzhitov, R. Type I Interferons in Host Defense. *Immunity* vol. 25 373–381 Preprint at https://doi.org/10.1016/j.immuni.2006.08.007 (2006).

59. Ding, S. W. & Voinnet, O. Antiviral Immunity Directed by Small RNAs. *Cell* vol. 130 413–426 Preprint at https://doi.org/10.1016/j.cell.2007.07.039 (2007).

60. Barton, G. M., Kagan, J. C. & Medzhitov, R. Intracellular localization of Toll-like receptor 9 prevents recognition of self DNA but facilitates access to viral DNA. *Nat Immunol* **7**, 49–56 (2006).

61. Yoshida, H., Okabe, Y., Kawane, K., Fukuyama, H. & Nagata, S. Lethal anemia caused by interferon-β produced in mouse embryos carrying undigested DNA. *Nat Immunol* **6**, 49–56 (2005).

62. Rapicetta, M., Ferrari, C. & Levrero, M. Viral determinants and host immune responses in the pathogenesis of HBV infection. in *Journal of Medical Virology* vol. 67 454–457 (2002).

63. Roossinck, M. J. The good viruses: Viral mutualistic symbioses. *Nat Rev Microbiol* **9**, 99–108 (2011).

64. Roossinck, M. J. Move Over, Bacteria! Viruses Make Their Mark as Mutualistic Microbial Symbionts. *J Virol* **89**, 6532–6535 (2015).

65. Metcalf, J. A. & Bordenstein, S. R. The complexity of virus systems: The case of endosymbionts. *Curr Opin Microbiol* **15**, 546–552 (2012).

66. Hamelin, F. M. *et al.* The evolution of parasitic and mutualistic plant–virus symbioses through transmission-virulence trade-offs. *Virus Res* **241**, 77–87 (2017).

67. Roossinck, M. *Virus: An Illustrated Guide to 101 Incredible Microbes - Marilyn J. Roossinck - Google Books*. (The Ivy Press Limited, 2016).

68. Ames, B. N. Identifying environmental chemicals causing mutations and cancer. *Science (1979)* **204**, 587–593 (1979).

69. Danaei, G., Vander Hoorn, S., Lopez, A. D., Murray, C. J. & Ezzati, M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *The Lancet* **366**, 1784–1793 (2005).

70. Costa, N. R., Gil da Costa, R. M. & Medeiros, R. A viral map of gastrointestinal cancers. *Life Sci* **199**, 188–200 (2018).

71. Zhang, D.-H. *et al.* Prevalence and association of human papillomavirus 16, Epstein-Barr virus, herpes simplex virus-1 and cytomegalovirus infection with human esophageal carcinoma: A case-control study. *Oncol Rep* **25**, 1731–1738 (2011).

72. zur Hausen, H. Viruses in human cancers. *Curr Sci* **81**, 523–527 (2001).

73. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int J Cancer* **118**, 3030–3044 (2006).

74. Krump, N. A. & You, J. Molecular mechanisms of viral oncogenesis in humans. *Nat Rev Microbiol* **16**, 684–698 (2018).

75. Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170–173 (1976).

76. Griffiths, D. J. Endogenous retroviruses in the human genome sequence. *Genome Biol* **2**, 1–5 (2001).

77. Löwer, R., Löwer, J. & Kurth, R. The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A* **93**, 5177–5184 (1996).

78. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annu Rev Genet* **42**, 709–732 (2008).

79. Nelson, P. N. *et al.* Human endogenous retroviruses. *Molecular Pathology* **56**, 11 (2003).

80. Vincendeau, M. *et al.* Modulation of human endogenous retrovirus (HERV) transcription during persistent and de novo HIV-1 infection. *Retrovirology* **12**, 1–17 (2015).

81. Markovitz, D. M. & Arbor, A. “Reverse Genomics” and Human Endogenous Retroviruses. *Trans Am Clin Climatol Assoc* **125**, 57 (2014).

82. Sauter, M. *et al.* Human endogenous retrovirus K10: expression of Gag protein and detection of antibodies in patients with seminomas. *J Virol* **69**, 414–421 (1995).

83. Wang-Johanning, F. *et al.* Quantitation of HERV-K env gene expression and splicing in human breast cancer. *Oncogene 2003 22:10* **22**, 1528–1535 (2003).

84. Kidwell, M. G. Transposable Elements. in *The Evolution of the Genome* 165–221 (Academic Press, 2005). doi:10.1016/B978-012301463-4/50005-X.

85. Bourque, G. *et al.* Ten things you should know about transposable elements 06 Biological Sciences 0604 Genetics. *Genome Biol* **19**, (2018).

86. Burns, K. H. Transposable elements in cancer. *Nat Rev Cancer* **17**, 415–424 (2017).

87. Mao, J., Zhang, Q. & Cong, Y. S. Human endogenous retroviruses in development and disease. *Comput Struct Biotechnol J* **19**, 5978–5986 (2021).

88. Moyes, D. L. *et al.* The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease. *Genomics* **86**, 337–341 (2005).

89. Contreras-Galindo, R. *et al.* Characterization of Human Endogenous Retroviral Elements in the Blood of HIV-1-Infected Individuals. *J Virol* **86**, 262–276 (2012).

90. Contreras-Galindo, R. *et al.* Human Endogenous Retrovirus Type K (HERV-K) Particles Package and Transmit HERV-K–Related Sequences. *J Virol* **89**, 7187–7201 (2015).

91. Boller, K. *et al.* Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *Journal of General Virology* **89**, 567–572 (2008).

92. Xue, B., Sechi, L. A. & Kelvin, D. J. Human Endogenous Retrovirus K (HML-2) in Health and Disease. *Front Microbiol* **11**, (2020).

93. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**, (2015).

94. Antony, J. M., DesLauriers, A. M., Bhat, R. K., Ellestad, K. K. & Power, C. Human endogenous retroviruses and multiple sclerosis: Innocent bystanders or disease determinants? *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1812**, 162–176 (2011).

95. Gonzalez-Cao, M. *et al.* Human endogenous retroviruses and cancer. *Cancer Biol Med* **13**, 483 (2016).

96. Wang-Johanning, F. *et al.* Human Endogenous Retrovirus K Triggers an Antigen-Specific Immune Response in Breast Cancer Patients. *Cancer Res* **68**, 5869–5877 (2008).

97. Kahyo, T. *et al.* Identification and association study with lung cancer for novel insertion polymorphisms of human endogenous retrovirus. *Carcinogenesis* **34**, 2531–2538 (2013).

98. Ishida, T. *et al.* Identification of the HERV-K gag antigen in prostate cancer by SEREX using autologous patient serum and its immunogenicity. *Cancer Immun* **8**, (2008).

99. Ma, W. *et al.* Human Endogenous retroviruses-k (HML-2) expression is correlated with prognosis and progress of hepatocellular carcinoma. *Biomed Res Int* **2016**, (2016).

100. Serafino, A. *et al.* The activation of human endogenous retrovirus K (HERV-K) is implicated in melanoma cell malignant transformation. *Exp Cell Res* **315**, 849–862 (2009).

101. Kleiman, A. *et al.* HERV-K(HML-2) GAG/ENV antibodies as indicator for therapy effect in patients with germ cell tumors. *Int J Cancer* **110**, 459–461 (2004).

102. Fischer, S. *et al.* Human endogenous retrovirus np9 gene is over expressed in chronic lymphocytic leukemia patients. *Leuk Res Rep* **3**, 70–72 (2014).

103. Gitlin, S. D., Galindo, R. C.-, Kaplan, M. H. & Markovitz, D. M. Role of Human Endogenous Retroviruses in Lymphoma Pathogenesis and a Possible Biomarker of Disease. *Blood* **112**, 3751–3751 (2008).

104. Ma, W. *et al.* Human Endogenous retroviruses-k (HML-2) expression is correlated with prognosis and progress of hepatocellular carcinoma. *Biomed Res Int* **2016**, (2016).

105. Pérot, P. *et al.* Expression of young HERV-H loci in the course of colorectal carcinoma and correlation with molecular subtypes. *Oncotarget* **6**, 40095 (2015).

106. Wallace, T. A. *et al.* Elevated HERV-K mRNA expression in PBMC is associated with a prostate cancer diagnosis particularly in older men and smokers. *Carcinogenesis* **35**, 2074–2083 (2014).

107. Li, Z. *et al.* Expression of HERV-K Correlates With Status of MEK–ERK and p16INK4A–CDK4 Pathways in Melanoma Cells. *http://dx.doi.org/10.3109/07357907.2010.512604* **28**, 1031–1037 (2010).

108. Matteucci, C., Balestrieri, E., Argaw-Denboba, A. & Sinibaldi-Vallebona, P. Human endogenous retroviruses role in cancer cell stemness. *Semin Cancer Biol* **53**, 17–30 (2018).

109. Jin, L. & Su, B. Natives or immigrants: modern human origin in east asia. *Nature Reviews Genetics 2000 1:2* **1**, 126–133 (2000).

110. Stringer, C. B. & Andrews, P. Genetic and Fossil Evidence for the Origin of Modern Humans. *Science (1979)* **239**, 1263–1268 (1988).

111. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90 (2011).

112. Nishikawa, J. *et al.* Clinical Importance of Epstein–Barr Virus-Associated Gastric Cancer. *Cancers 2018, Vol. 10, Page 167* **10**, 167 (2018).

113. Al-Haddad, S. *et al.* Infection and esophageal cancer. *Ann N Y Acad Sci* **1325**, 187–196 (2014).

114. Li, S. *et al.* Oesophageal carcinoma: The prevalence of DNA tumour viruses and therapy. *Tumour Virus Res* **13**, 200231 (2022).

115. van Doorslaer, K. *et al.* The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res* **41**, D571–D578 (2013).

116. de Villiers, E. M., Fauquet, C., Broker, T. R., Bernard, H. U. & zur Hausen, H. Classification of papillomaviruses. *Virology* **324**, 17–27 (2004).

117. Farhadi, M. *et al.* Human papillomavirus in squamous cell carcinoma of esophagus in a high-risk population. *World J Gastroenterol* **11**, 1200–1203 (2005).

118. Serrano, B., Brotons, M., Bosch, F. X. & Bruni, L. Epidemiology and burden of HPV-related disease. *Best Pract Res Clin Obstet Gynaecol* **47**, 14–26 (2018).

119. Syrjänen, K. J. HPV infections and oesophageal cancer. *J Clin Pathol* **55**, 721–728 (2002).

120. Samoff, E. *et al.* Association of Chlamydia trachomatis with Persistence of High-Risk Types of Human Papillomavirus in a Cohort of Female Adolescents. *Am J Epidemiol* **162**, 668–675 (2005).

121. Pett, M. R. *et al.* Acquisition of High-Level Chromosomal Instability Is Associated with Integration of Human Papillomavirus Type 16 in Cervical Keratinocytes. *Cancer Res* **64**, 1359–1368 (2004).

122. Moody, C. A. & Laimins, L. A. Human papillomavirus oncoproteins: pathways to transformation. *Nature Reviews Cancer 2010 10:8* **10**, 550–560 (2010).

123. Syrjanen, K., Pyrhonen, S., Aukee, S. & Koskela, E. Squamous cell papilloma of the esophagus: a tumour probably caused by human papilloma virus (HPV). *Diagn Histopathol* **5**, 291–296 (1982).

124. Ludmir, E. B., Stephens, S. J., Palta, M., Willett, C. G. & Czito, B. G. Human papillomavirus tumor infection in esophageal squamous cell carcinoma. *J Gastrointest Oncol* **6**, 287 (2015).

125. Koh, J. S., Lee, S. S., Baek, H. J. & Kim, Y. I. No association of high-risk human papillomavirus with esophageal squamous cell carcinomas among Koreans, as determined by polymerase chain reaction. *Diseases of the Esophagus* **21**, 114–117 (2008).

126. Li, T. *et al.* Human papillomavirus type 16 is an important infectious factor in the high incidence of esophageal cancer in Anyang area of China. *Carcinogenesis* **22**, 929–934 (2001).

127. Clarke, B. & Chetty, R. Cell Cycle Aberrations in the Pathogenesis of Squamous Cell Carcinoma of the Uterine Cervix. *Gynecol Oncol* **82**, 238–246 (2001).

128. Pinto, À. P., Degen, M., Villa, L. L. & Cibas, E. S. Immunomarkers in Gynecologic Cytology: The Search for the Ideal 'Biomolecular Papanicolaou Test'. *Acta Cytol* **56**, 109–121 (2012).

129. Boon, S. S. *et al.* Human papillomavirus type 18 oncoproteins exert their oncogenicity in esophageal and tongue squamous cell carcinoma cell lines distinctly. *BMC Cancer* **19**, 1–12 (2019).

130. Wang, Y., Zeng, G. & Jiang, Y. The Emerging Roles of miR-125b in Cancers. *Cancer Manag Res* **12**, 1079–1088 (2020).

131. Wang, B., Li, X., Liu, L. & Wang, M. β-Catenin: Oncogenic role and therapeutic target in cervical cancer. *Biol Res* **53**, 1–11 (2020).

132. Alamoud, K. A. & Kukuruzinska, M. A. Emerging Insights into Wnt/β-catenin Signaling in Head and Neck Cancer. *J Dent Res* **97**, 665–673 (2018).

133. Smatti, M. K. *et al.* Epstein-barr virus epidemiology, serology, and genetic variability of LMP-1 oncogene among healthy population: An update. *Front Oncol* **8**, 211 (2018).

134. Chakravorty, S. *et al.* Integrated pan-cancer map of EBV-associated neoplasms reveals functional host–virus interactions. *Cancer Res* **79**, 6010–6023 (2019).

135. Phan, A. T., Fernandez, S. G., Somberg, J. J., Keck, K. M. & Miranda, J. J. L. Epstein–Barr virus latency type and spontaneous reactivation predict lytic induction levels. *Biochem Biophys Res Commun* **474**, 71–75 (2016).

136. Bakkalci, D. *et al.* Risk factors for Epstein Barr virus-associated cancers: a systematic review, critical appraisal, and mapping of the epidemiological evidence. *J Glob Health* **10**, (2020).

137. Wang, L. Detection of Epstein-Barr virus in esophageal squamous cell carcinoma in Taiwan. *Am J Gastroenterol* **94**, 2834–2839 (1999).

138. Sunpaweravong, S., Mitarnun, W. & Puttawibul, P. Absence of Epstein-Barr virus in esophageal squamous cell carcinoma. *Diseases of the Esophagus* **18**, 398–399 (2005).

139. Zhang, D. H. *et al.* Prevalence and association of human papillomavirus 16, Epstein-Barr virus, herpes simplex virus-1 and cytomegalovirus infection with human esophageal carcinoma: A case-control study. *Oncol Rep* **25**, 1731–1738 (2011).

140. Wu, M. Y., Wu, X. Y. & Zhuang, C. X. Detection of HSV and EBV in esophageal carcinomas from a high-incidence area in Shantou China. *Diseases of the Esophagus* **18**, 46–50 (2005).

141. Jenkins, T. D., Nakagawa, H. & Rustgi, A. K. The association of Epstein-Barr virus DNA with esophageal squamous cell carcinoma. *Oncogene* **13**, 1809–1813 (1996).

142. Awerkiew, S. *et al.* Esophageal cancer in germany is associated with Epstein-Barr-virus but not with papillomaviruses. *Med Microbiol Immunol* **192**, 137–140 (2003).

143. Awerkiew, S. *et al.* Presence of Epstein-Barr virus in esophageal cancer is restricted to tumor infiltrating lymphocytes. *Med Microbiol Immunol* **194**, 187–191 (2005).

144. Chang, F. *et al.* Evaluation of HPV, CMV, HSV and EBV in esophageal squamous cell carcinomas from a high-incidence area of China. *Anticancer Res* **20**, 3935–3940 (2000).

145. Yahyapour, Y. *et al.* Prevalence and association of human papillomavirus, Epstein-Barr virus and Merkel Cell polyomavirus with neoplastic esophageal lesions in northern Iran. *Caspian J Intern Med* **9**, 353–360 (2018).

146. Lyronis, I. D., Baritaki, S., Bizakis, I., Tsardi, M. & Spandidos, D. A. Evaluation of the Prevalence of Human Papillomavirus and Epstein-Barr Virus in Esophageal Squamous Cell Carcinomas. *International Journal of Biological Markers* **20**, 5–10 (2018).

147. Yanai, H. *et al.* Epstein-barr virus association is rare in esophageal squamous cell carcinoma. *International Journal of Gastrointestinal Cancer 2003 33:2* **33**, 165–170 (2003).

148. Rajendra, K. & Sharma, P. Viral Pathogens in Oesophageal and Gastric Cancer. *Pathogens 2022, Vol. 11, Page 476* **11**, 476 (2022).

149. Goodgame, R. W. Gastrointestinal cytomegalovirus disease. *Ann Intern Med* **119**, 924–935 (1993).

150. Murakami, D., Harada, H., Yamato, M. & Amano, Y. Cytomegalovirus-associated esophagitis on early esophageal cancer in immunocompetent host: a case report. *Gut Pathog* **13**, 1–8 (2021).

151. Han, C. P., Tsao, Y. P., Sun, C. A., Ng, H. T. & Chen, S. L. Human papillomavirus, cytomegalovirus and herpes simplex virus infections for cervical cancer in Taiwan. *Cancer Lett* **120**, 217–221 (1997).

152. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature 2009 458:7239* **458**, 719–724 (2009).

153. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

154. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *New England Journal of Medicine* **377**, 111–121 (2017).

155. Anglesio, M. S. *et al.* Cancer-Associated Mutations in Endometriosis without Cancer. *New England Journal of Medicine* **376**, 1835–1848 (2017).

156. Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* **68**, 6863–6872 (2008).

157. Olivier, M., Hussain, S. P., Caron de Fromentel, C., Hainaut, P. & Harris, C. C. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci Publ* 247–270 (2004).

158. Miller, D. G. On the Nature of Susceptibility to Cancer The Presidential Address. *Cancer* **46**, 1307–1318 (1980).

159. Schinzel, A. C. & Hahn, W. C. Oncogenic transformation and experimental models of human cancer. *Front Biosci* **13**, 71–84 (2008).

160. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183 (2004).

161. Touw, I. P. & Erkeland, S. J. Retroviral Insertion Mutagenesis in Mice as a Comparative Oncogenomics Tool to Identify Disease Genes in Human Leukemia. *Molecular Therapy* **15**, 13–19 (2007).

162. Johnson, G. L. & Lapadat, R. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **298**, 1911–1912 (2002).

163. Simba, H., Kuivaniemi, H., Lutje, V., Tromp, G. & Sewram, V. Systematic review of genetic factors in the etiology of esophageal squamous cell carcinoma in African populations. *Front Genet* **10**, 642 (2019).

164. Qin, H. De *et al.* Genomic Characterization of Esophageal Squamous Cell Carcinoma Reveals Critical Genes Underlying Tumorigenesis and Poor Prognosis. *The American Journal of Human Genetics* **98**, 709–727 (2016).

165. Sawada, G. *et al.* Genomic Landscape of Esophageal Squamous Cell Carcinoma in a Japanese Population. *Gastroenterology* **150**, 1171–1182 (2016).

166. Lin, D. C., Wang, M. R. & Koeffler, H. P. Genomic and Epigenomic Aberrations in Esophageal Squamous Cell Carcinoma and Implications for Patients. *Gastroenterology* **154**, 374–389 (2018).

167. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519 (2012).

168. Lawrence, M. S. *et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576 (2015).

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

170. Cheung, W. Y. & Liu, G. Genetic Variations in Esophageal Cancer Risk and Prognosis. *Gastroenterol Clin North Am* **38**, 75–91 (2009).

171. Hiyama, T., Yoshihara, M., Tanaka, S. & Chayama, K. Genetic polymorphisms and esophageal cancer risk. *Int J Cancer* **121**, 1643–1658 (2007).

172. Baba, Y. *et al.* Genetic and epigenetic characteristics of esophageal cancer tissues with microbiome fusobacterium nucleatum. *Cancer Res* **77**, 4930–4930 (2017).

173. Liu, W. *et al.* Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI Insight* **2**, (2017).

174. Patel, K. *et al.* TP53 mutations, human papilloma virus DNA and inflammation markers in esophageal squamous cell carcinoma from the Rift Valley, a high-incidence area in Kenya. *BMC Res Notes* **4**, 1–6 (2011).

175. Dietzsch, E. & Parker, M. I. Infrequent somatic deletion of the 5′ region of the COL1A2 gene in oesophageal squamous cell cancer patients. *Clin Chem Lab Med* **40**, 941–945 (2002).

176. Dietzsch, E., Laubscher, R. & Parker, M. I. Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *Int J Cancer* **107**, 38–45 (2003).

177. Naidoo, R., Ramburan, A., Reddi, A. & Chetty, R. Aberrations in the mismatch repair genes and the clinical impact on oesophageal squamous carcinomas from a high incidence area in South Africa. *J Clin Pathol* **58**, 281–284 (2005).

178. Dhanisha, S. S., Guruvayoorappan, C., Drishya, S. & Abeesh, P. Mucins: Structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. *Crit Rev Oncol Hematol* **122**, 98–122 (2018).

179. Forstner, J. F. Intestinal Mucins in Health and Disease. *Digestion* **17**, 234–263 (1978).

180. Guillem, P. *et al.* Mucin gene expression and cell differentiation in human normal, premalignant and malignant esophagus. *Int J Cancer* **88**, 856–861 (2000).

181. Hollingsworth, M. A. & Swanson, B. J. Mucins in cancer: protection and control of the cell surface. *Nature Reviews Cancer 2004 4:1* **4**, 45–60 (2004).

182. Lau, S. K., Weiss, L. M. & Chu, P. G. Differential Expression of MUC1, MUC2, and MUC5AC in Carcinomas of Various SitesAn Immunohistochemical Study. *Am J Clin Pathol* **122**, 61–69 (2004).

183. Singh, P. K. & Hollingsworth, M. A. Cell surface-associated mucins in signal transduction. *Trends Cell Biol* **16**, 467–476 (2006).

184. Van Putten, J. P. M. & Strijbis, K. Transmembrane Mucins: Signaling Receptors at the Intersection of Inflammation and Cancer. *J Innate Immun* **9**, 281–299 (2017).

185. Bansil, R. & Turner, B. S. Mucin structure, aggregation, physiological functions and biomedical applications. *Curr Opin Colloid Interface Sci* **11**, 164–170 (2006).

186. Jonckheere, N. & Van Seuningen, I. The membrane-bound mucins: From cell signalling to transcriptional regulation and expression in epithelial cancers. *Biochimie* **92**, 1–11 (2010).

187. Zheng, F., Yu, H. & Lu, J. High expression of MUC20 drives tumorigenesis and predicts poor survival in endometrial cancer. *J Cell Biochem* **120**, 11859–11866 (2019).

188. Cummings, R. D. The repertoire of glycan determinants in the human glycome. *Mol Biosyst* **5**, 1087–1104 (2009).

189. Basbaum, C. *et al.* Control of Mucin Transcription by Diverse Injury-induced Signaling Pathways. *Am J Respir Crit Care Med* **160**, 44–48 (2012).

190. Duraisamy, S., Ramasamy, S., Kharbanda, S. & Kufe, D. Distinct evolution of the human carcinoma-associated transmembrane mucins, MUC1, MUC4 AND MUC16. *Gene* **373**, 28–34 (2006).

191. Bork, P. & Patthy, L. The SEA module: A new extracellular domain associated with O-glycosylation. *Protein Science* **4**, 1421–1425 (1995).

192. Levitin, F. *et al.* The MUC1 SEA module is a self-cleaving domain. *Journal of Biological Chemistry* **280**, 33374–33386 (2005).

193. Macao, B., Johansson, D. G. A., Hansson, G. C. & Härd, T. Autoproteolysis coupled to protein folding in the SEA domain of the membrane-bound MUC1 mucin. *Nat Struct Mol Biol* **13**, 71–76 (2006).

194. Palmai-Pallag, T. *et al.* The role of the SEA (sea urchin sperm protein, enterokinase and agrin) module in cleavage of membrane-tethered mucins. *FEBS J* **272**, 2901–2911 (2005).

195. Pelaseyed, T. *et al.* Unfolding dynamics of the mucin SEA domain probed by force spectroscopy suggest that it acts as a cell-protective device. *FEBS J* **280**, 1491–1501 (2013).

196. Vadaie, N. *et al.* Cleavage of the signaling mucin Msb2 by the aspartyl protease Yps1 is required for MAPK activation in yeast. *Journal of Cell Biology* **181**, 1073–1081 (2008).

197. Funes, M., Miller, J. K., Lai, C., Carraway, K. L. & Sweeney, C. The mucin Muc4 potentiates neuregulin signaling by increasing the cell-surface populations of ErbB2 and ErbB3. *Journal of Biological Chemistry* **281**, 19310–19319 (2006).

198. Jepson, S. *et al.* Muc4/sialomucin complex, the intramembrane ErbB2 ligand, induces specific phosphorylation of ErbB2 and enhances expression of p27kip, but does not activate mitogen-activated kinase or protein kinaseB/Akt pathways. *Oncogene 2002 21:49* **21**, 7524–7532 (2002).

199. Komatsu, M., Jepson, S., Arango, M. E., Carothers Carraway, C. A. & Carraway, K. L. Muc4/sialomucin complex, an intramembrane modulator of ErbB2/HER2/Neu, potentiates primary tumor growth and suppresses apoptosis in a xenotransplanted tumor. *Oncogene 2001 20:4* **20**, 461–470 (2001).

200. Carraway, K. L., Ramsauer, V. P., Haq, B. & Carothers Carraway, C. A. Cell signaling through membrane mucins. *BioEssays* **25**, 66–71 (2003).

201. Hattrup, C. L. & Gendler, S. J. Structure and Function of the Cell Surface (Tethered) Mucins. *Annual Reviews* **70**, 431–457 (2008).

202. Malmberg, E. K. *et al.* The C-terminus of the transmembrane mucin MUC17 binds to the scaffold protein PDZK1 that stably localizes it to the enterocyte apical membrane in the small intestine. *Biochemical Journal* **410**, 283–289 (2008).

203. Lamprecht, G. & Seidler, U. The emerging role of PDZ adapter proteins for regulation of intestinal ion transport. *Am J Physiol Gastrointest Liver Physiol* **291**, 766–777 (2006).

204. Rhodes, J. M. Usefulness of novel tumour markers. *Annals of Oncology* **10**, S118–S121 (1999).

205. Mejías-Luque, R. *et al.* Inflammation modulates the expression of the intestinal mucins MUC2 and MUC4 in gastric tumors. *Oncogene* **29**, 1753–1762 (2010).

206. Lan, M. S., Batra, S. K., Qi, W. N., Metzgar, R. S. & Hollingsworth, M. A. Cloning and sequencing of a human pancreatic tumor mucin cDNA. *Journal of Biological Chemistry* **265**, 15294–15299 (1990).

207. Lan, Michael. S., Hollingsworth, Michael. A. & Metzgar, Richard. S. Polypeptide Core of a Human Pancreatic Tumor Mucin Antigen1 | Cancer Research | American Association for Cancer Research. *Cancer Res* **50**, 2997–3001 (1990).

208. Hanisch, F. G. & Müller, S. MUC1: the polymorphic appearance of a human mucin. *Glycobiology* **10**, 439–449 (2000).

209. Schroeder, J. A., Thompson, M. C., Gardner, M. M. & Gendler, S. J. Transgenic MUC1 Interacts with Epidermal Growth Factor Receptor and Correlates with Mitogen-activated Protein Kinase Activation in the Mouse Mammary Gland. *Journal of Biological Chemistry* **276**, 13057–13064 (2001).

210. Smorodinsky, N. *et al.* Detection of a Secreted MUC1/SEC Protein by MUC1 Isoform Specific Monoclonal Antibodies. *Biochem Biophys Res Commun* **228**, 115–121 (1996).

211. Wild, C. P. & Hardie, L. J. Reflux, Barrett's oesophagus and adenocarcinoma: burning questions. *Nature Reviews Cancer* **3**, 676–684 (2003).

212. Sagara, M. *et al.* Expression of mucin 1 (MUC1) in esophageal squamous-cell carcinoma: Its relationship with prognosis. *Int. J. Cancer (Pred. Oncol.)* **84**, 251–257 (1999).

213. Osako, M. *et al.* lmmunohistochemical Study of Mucin Carbohydrates and Core Proteins in Human Pancreatic Tumors. *Cancer* **71**, 2191–2199 (1993).

214. Yamashita, K. *et al.* Immunohistochemical study of mucin carbohydrates and core proteins in hepatolithiasis and cholangiocarcinoma. *Int J Cancer* **55**, 82–91 (1993).

215. Yonezawa, S., Tachikawa, T., Shin, S. & Sato, E. Sialosyl-Tn Antigen: Its Distribution in Normal Human Tissues and Expression in Adenocarcinomas. *Am J Clin Pathol* **98**, 167–174 (1992).

216. Aubert, S. *et al.* MUC1, a new hypoxia inducible factor target gene, is an actor in clear renal cell carcinoma tumor progression. *Cancer Res* **69**, 5707–5715 (2009).

217. Costa, N. R., Paulo, P., Caffrey, T., Hollingsworth, M. A. & Santos-Silva, F. Impact of MUC1 Mucin Downregulation in the Phenotypic Characteristics of MKN45 Gastric Carcinoma Cell Line. *PLoS One* **6**, e26970 (2011).

218. Hattrup, C. L. & Gendler, S. J. MUC1 alters oncogenic events and transcription in human breast cancer cells. *Breast Cancer Research* **8**, 1–10 (2006).

219. Yin, L., Li, Y., Ren, J., Kuwahara, H. & Kufe, D. Human MUC1 carcinoma antigen regulates intracellular oxidant levels and the apoptotic response to oxidative stress. *Journal of Biological Chemistry* **278**, 35458–35464 (2003).

220. Kufe, D. W. Mucins in cancer: function, prognosis and therapy. *Nature Reviews Cancer 2009 9:12* **9**, 874–885 (2009).

221. Togami, S. *et al.* Expression of mucin antigens (MUC1 and MUC16) as a prognostic factor for mucinous adenocarcinoma of the uterine cervix. *Journal of Obstetrics and Gynaecology Research* **36**, 588–597 (2010).

222. Macha, M. A. *et al.* MUC4 regulates cellular senescence in head and neck squamous cell carcinoma through p16/Rb pathway. *Oncogene* **34**, 1698–1708 (2014).

223. Senapati, S. *et al.* Deregulation of MUC4 in gastric adenocarcinoma: potential pathobiological implication in poorly differentiated non-signet ring cell type gastric cancer. *British Journal of Cancer 2008 99:6* **99**, 949–956 (2008).

224. Chaturvedi, P., Singh, A. P. & Batra, S. K. Structure, evolution, and biology of the MUC4 mucin. *The FASEB Journal* **22**, 966–981 (2008).

225. Saitou, M. *et al.* MUC4 expression is a novel prognostic factor in patients with invasive ductal carcinoma of the pancreas. *J Clin Pathol* **58**, 845–852 (2005).

226. Hinoda, Y. *et al.* Increased expression of MUC1 in advanced pancreatic cancer. *J Gastroenterol* **38**, 1162–1166 (2003).

227. Ligtenberg, M. J. L., Buijs, F., Vos, H. L. & Hilkens, J. Suppression of Cellular Aggregation by High Levels of Episialin | Cancer Research | American Association for Cancer Research. *Cancer Res* **52**, 2318–2324 (1992).

228. Makiguchi, Y., Hinoda, Y. & Imai, K. Effect of MUC1 Mucin, an Anti-adhesion Molecule, on Tumor Cell Growth. *Japanese Journal of Cancer Research* **87**, 505–511 (1996).

229. Song, Z. B. *et al.* Expression of MUC1 in esophageal squamous-cell carcinoma and its relationship with prognosis of patients from Linzhou city, a high incidence area of northern China. *World J Gastroenterol* **9**, 404–407 (2003).

230. Kanwal, M., Ding, X. J., Song, X., Zhou, G. B. & Cao, Y. MUC16 overexpression induced by gene mutations promotes lung cancer cell growth and invasion. *Oncotarget* **9**, 12239 (2018).

231. Sheng, Y. *et al.* MUC13 overexpression in renal cell carcinoma plays a central role in tumor progression and drug resistance. *Int J Cancer* **140**, 2351–2363 (2017).

232. Valque, H., Gouyer, V., Gottrand, F. & Desseyn, J. L. MUC5B Leads to Aggressive Behavior of Breast Cancer MCF7 Cells. *PLoS One* **7**, e46699 (2012).

233. Astashchanka, A., Shroka, T. M. & Jacobsen, B. M. Mucin 2 (MUC2) modulates the aggressiveness of breast cancer. *Breast Cancer Res Treat* **173**, 289–299 (2019).

234. Chauhan, S. C. *et al.* Aberrant expression of MUC4 in ovarian carcinoma: diagnostic significance alone and in combination with MUC1 and MUC16 (CA125). *Modern Pathology 2006 19:10* **19**, 1386–1394 (2006).

235. Reynolds, I. S. *et al.* Mucin glycoproteins block apoptosis; promote invasion, proliferation, and migration; and cause chemoresistance through diverse pathways in epithelial cancers. *Cancer and Metastasis Reviews* **38**, 237–257 (2019).

236. Komatsu, M., Carothers Carraway, C. A., Fregien, N. L. & Carraway, K. L. Reversible disruption of cell-matrix and cell-cell interactions by overexpression of sialomucin complex. *Journal of Biological Chemistry* **272**, 33245–33254 (1997).

237. Komatsu, M., Tatum, L., Altman, N. H., Carraway, C. A. C. & Carraway, K. L. Potentiation of metastasis by cell surface sialomucin complex (rat MUC4), a multifunctional anti-adhesive glycoprotein. *Int J Cancer* **87**, 480–486 (2000).

238. Price-Schiavi, S. A. *et al.* Rat Muc4 (sialomucin complex) reduces binding of anti-ErbB2 antibodies to tumor cell surfaces, a potential mechanism for herceptin resistance. *Int J Cancer* **99**, 783–791 (2002).

239. Nagy, P. *et al.* Decreased Accessibility and Lack of Activation of ErbB2 in JIMT-1, a Herceptin-Resistant, MUC4-Expressing Breast Cancer Cell Line. *Cancer Res* **65**, 473–482 (2005).

240. Komatsu, M., Yee, L. & Carraway, K. L. Overexpression of Sialomucin Complex, a Rat Homologue of MUC4, Inhibits Tumor Killing by Lymphokine-activated Killer Cells1 | Cancer Research | American Association for Cancer Research. *Cancer Reseach* **59**, 2229–2236 (1999).

241. Fegelman, E. *et al.* Squamous cell carcinoma of the esophagus with mucin-secreting component: Mucoepidermoid carcinoma. *Journal of Thoracic and Cardiovascular Surgery* **107**, 62–67 (1994).

242. Karaki, Y. *et al.* Histogenesis of Adenosquamous Carcinoma of the Esophagus. in *Diseases of the Esophagus* 60–63 (Springer, Berlin, Heidelberg, 1988). doi:10.1007/978-3-642-86432-2_13.

243. Ikeda, Y. *et al.* Expression of Sialyl-Tn Antigens in Normal Squamous Epithelium, Dysplasia, and Squamous Cell Carcinoma in the Esophagus | Cancer Research | American Association for Cancer Research. *Cancer Res* **53**, 1706–1708 (1993).

244. Hirasawa, Y. *et al.* Natural Autoantibody to MUC1 Is a Prognostic Indicator for Non–Small Cell Lung Cancer. *Am J Respir Crit Care Med* **161**, 589–594 (2012).

245. Baldus, S. E. *et al.* Immunoreactivity of Monoclonal Antibody BW835 Represents a Marker of Progression and Prognosis in Early Gastric Cancer. *Oncology* **61**, 147–155 (2001).

246. Ye, Q. *et al.* MUC1 induces metastasis in esophageal squamous cell carcinoma by upregulating matrix metalloproteinase 13. *Laboratory Investigation* **91**, 778–787 (2011).

247. Jonckheere, N. & van Seuningen, I. The Membrane-Bound Mucins: How Large O-Glycoproteins Play Key Roles in Epithelial Cancers and Hold Promise as Biological Tools for Gene-Based and Immunotherapies. *Crit Rev Oncog* **14**, 177–196 (2008).

248. Arul, G. S. *et al.* Mucin gene expression in Barrett's oesophagus: an in situ hybridisation and immunohistochemical study. *Gut* **47**, 753–761 (2000).

249. Bruyère, E., Jonckheere, N., Frénois, F., Mariette, C. & van Seuningen, I. The MUC4 membrane-bound mucin regulates esophageal cancer cell proliferation and migration properties: Implication for S100A4 protein. *Biochem Biophys Res Commun* **413**, 325–329 (2011).

250. Niv, Y., Ho, S. B., Fass, R. & Rokkas, T. Mucin Expression in the Esophageal Malignant and Pre-malignant States. *J Clin Gastroenterol* **52**, 91–96 (2018).

251. Gum, J. R. *et al.* MUC3 Human Intestinal Mucin. *Journal of Biological Chemistry* **272**, 26678–26686 (1997).

252. Crawley, S. C. *et al.* Genomic Organization and Structure of the 3′ Region of Human MUC3: Alternative Splicing Predicts Membrane-Bound and Soluble Forms of the Mucin. *Biochem Biophys Res Commun* **263**, 728–736 (1999).

253. Guddo, F. *et al.* MUC1 (episialin) expression in non-small cell lung cancer is independent of EGFR and c-erbB-2 expression and correlates with poor survival in node positive patients. *J Clin Pathol* **51**, 667–671 (1998).

254. Songyang, Z. *et al.* Specific motifs recognized by the SH2 domains of Csk, 3BP2, fps/fes, GRB-2, HCP, SHC, Syk, and Vav. *Mol Cell Biol* **14**, 2777–2785 (1994).

255. Williams, S. J., Munster, D. J., Quin, R. J., Gotley, D. C. & McGuckin, M. A. The MUC3 Gene Encodes a Transmembrane Mucin and Is Alternatively Spliced. *Biochem Biophys Res Commun* **261**, 83–89 (1999).

256. Kitamoto, S. *et al.* Promoter hypomethylation contributes to the expression of MUC3A in cancer cells. *Biochem Biophys Res Commun* **397**, 333–339 (2010).

257. Park, H. U. *et al.* Aberrant expression of muc3 and muc4 membrane-associated mucins and sialyl lex antigen in pancreatic intraepithelial neoplasia. *Nursing (Brux)* **26**, (1996).

258. Wang, R.-Q. Alterations of MUC1 and MUC3 expression in gastric carcinoma: relevance to patient clinicopathological features. *J Clin Pathol* **56**, 378–384 (2003).

259. Rakha, E. A. *et al.* Expression of mucins (MUC1, MUC2, MUC3, MUC4, MUC5AC and MUC6) and their prognostic significance in human breast cancer. *Modern Pathology* **18**, 1295–1304 (2005).

260. Duncan, T. J., Watson, N. F. S., Al-Attar, A. H., Scholefield, J. H. & Durrant, L. G. The role of MUC1 and MUC3 in the biology and prognosis of colorectal cancer. *World J Surg Oncol* **5**, 1–11 (2007).

261. Leroy, X. *et al.* Quantitative RT-PCR assay for MUC3 and VEGF mRNA in renal clear cell carcinoma: relationship with nuclear grade and prognosis. *Urology* **62**, 771–775 (2003).

262. Pan, Q. *et al.* Enhanced membrane-tethered mucin 3 (MUC3) expression by a tetrameric branched peptide with a conserved tflk motif inhibits bacteria adherence. *Journal of Biological Chemistry* **288**, 5407–5416 (2013).

263. Louis, N. A. *et al.* Selective induction of mucin-3 by hypoxia in intestinal epithelia. *J Cell Biochem* **99**, 1616–1627 (2006).

264. Ho, S. B. *et al.* Cysteine-Rich Domains of Muc3 Intestinal Mucin Promote Cell Migration, Inhibit Apoptosis, and Accelerate Wound Healing. *Gastroenterology* **131**, 1501–1517 (2006).

265. Peng, Z. *et al.* Autoproteolysis of the SEA module of rMuc3 C-terminal domain modulates its functional composition. *Arch Biochem Biophys* **503**, 238–247 (2010).

266. Sun, Y. *et al.* MUC3A promotes non-small cell lung cancer progression via activating the NFκB pathway and attenuates radiosensitivity. *Int J Biol Sci* **17**, 2536 (2021).

267. Niu, T. *et al.* Increased expression of MUC3A is associated with poor prognosis in localized clear-cell renal cell carcinoma. *Oncotarget* **7**, 50026 (2016).

268. Copin, M.-C. *et al.* From normal respiratory mucosa to epidermal carcinoma: Expression of human mucin genes. *Int J Cancer* **86**, 162–168 (2000).

269. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nature Genetics 2003 33:3* **33**, 238–244 (2003).

270. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

271. Nowell, P. C. The Clonal Evolution of Tumor Cell Populations. *Science (1979)* **194**, 23–28 (1976).

272. Stransky, B. & Galante, P. Application of bioinformatics in cancer research. in *An Omics Perspective on Cancer Research* 211–233 (Springer, Dordrecht, 2010). doi:10.1007/978-90-481-2675-0_12/COVER.

273. Desany, B. & Zhang, Z. Bioinformatics and cancer target discovery. *Drug Discov Today* **9**, 795–802 (2004).

274. Lu, D. Y., Lu, T. R., Chen, E. H., Ding, J. & Xu, B. Cancer Bioinformatics, its Impacts on Cancer Therapy. *Journal of Postgenomics Drug & Biomarker Development* **05**, e133 (2015).

275. Kim, B. *et al.* Clinical Validity of the Lung Cancer Biomarkers Identified by Bioinformatics Analysis of Public Expression Data. *Cancer Res* **67**, 7431–7438 (2007).

276. Wu, D., Rice, C. M. & Wang, X. Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics 2012 13:1* **13**, 1–4 (2012).

277. Couzin-Frankel, J. Cancer immunotherapy. *Science (1979)* **342**, 1432–1433 (2013).

278. Knight, Z. A. & Shokat, K. M. Chemical Genetics: Where Genetics and Pharmacology Meet. *Cell* **128**, 425–430 (2007).

279. Komarova, N. L. Mathematical modeling of tumorigenesis: Mission possible. *Curr Opin Oncol* **17**, 39–43 (2005).

280. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115 (2015).

281. Bonetta, L. Going on a Cancer Gene Hunt. *Cell* **123**, 735–737 (2005).

282. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature 2003 422:6934* **422**, 835–847 (2003).

283. Albertson, D. G. *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genetics 2000 25:2* **25**, 144–146 (2000).

284. BLAST: Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi (2022).

285. Quackenbush, J. Microarrays--Guilt by Association. *Science (1979)* **302**, 240–241 (2003).

286. Nagl, S. *Cancer Bioinformatics: From Therapy Design to Treatment*. *Cancer Bioinformatics: From Therapy Design to Treatment* vol. 30 (John Wiley and Sons, 2006).

287. Park, S. T. & Kim, J. Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. *Int Neurourol J* **20**, 76–83 (2016).

288. Ng, P. C. & Kirkness, E. F. Whole genome sequencing. in *Genetic Variation* (eds. Barnes, M. R. & Gerome, B.) vol. 628 215–226 (Humana Press Inc., 2010).

289. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11 (2005).

290. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

291. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, 1–9 (2004).

292. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res* **11**, 1725–1729 (2001).

293. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133–141 (2008).

294. Ormond, K. E. *et al.* Challenges in the clinical application of whole-genome sequencing. *The Lancet* **375**, 1749–1751 (2010).

295. Veale, R. B. & Thornley, A. L. Atypical Cytokeratins Synthesized by Human Oesophageal Carcinoma Cells in Culture. *S Afr J Sci* **80**, 260–267 (1984).

296. Shimada, Y., Lmamura, M., Wagata, T., Yamaguchi, N. & Tobe, T. Characterization of 2 1 Newly Established Esophageal Cancer Cell Lines. *Cancer* **69**, 277–284 (1992).

297. Harada, H. *et al.* Telomerase Induces Immortalization of Human Esophageal Keratinocytes Without p16 INK4a Inactivation. *Molecular Cancer Research* **1**, 729–738 (2003).

298. Lee, P. Y., Costumbrado, J., Hsu, C. Y. & Kim, Y. H. Agarose Gel Electrophoresis for the Separation of DNA Fragments. *JoVE (Journal of Visualized Experiments)* e3923 (2012) doi:10.3791/3923.

299. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

300. Picard Tools. Preprint at https://broadinstitute.github.io/picard/ (2019).

301. GATK . Preprint at https://github.com/broadinstitute/gatk (2022).

302. Jones, D. *et al.* cgpCaVEManWrapper: Simple execution of caveman in order to detect somatic single nucleotide variants in NGS data. *Curr Protoc Bioinformatics* **2016**, 15.10.1-15.10.18 (2016).

303. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15.7.1-15.7.12 (2015).

304. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910–16915 (2010).

305. Strauss, W. M. Preparation of Genomic DNA from Mammalian Tissue . *Curr Protoc Mol Biol* **42**, 2.2.1-2.2.3 (1998).

306. Canene-Adams, K. General PCR. in *laboratory Methods in Enzymology* (ed. Lorsch, J.) vol. 529 291–298 (Academic Press Inc., 2013).

307. Turner, G. *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology* **11**, 1531–1535 (2001).

308. Burmeister, T. *et al.* Insertional polymorphisms of endogenous HERV-K113 and HERV-K115 retroviruses in breast cancer patients and age-matched controls. *AIDS Res Hum Retroviruses* **20**, 1223–1229 (2004).

309. McLaughlin-Drubin, M. E. & Munger, K. Viruses associated with human cancer. *Biochim Biophys Acta Mol Basis Dis* **1782**, 127–150 (2008).

310. Khoury, J. D. *et al.* Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J Virol* **87**, 8916–26 (2013).

311. Delecluse, H. J. & Hammerschmidt, W. Status of Marek's disease virus in established lymphoma cell lines: herpesvirus integration is common. *J Virol* **67**, 82–92 (1993).

312. Delecluse, H. J., Schüller, S. & Hammerschmidt, W. Latent Marek's disease virus can be activated from its chromosomally integrated state in herpesvirus-transformed lymphoma cells. *EMBO J* **12**, 3277–3286 (1993).

313. Flippot, R., Malouf, G. G., Su, X., Khayat, D. & Spano, J. P. Oncogenic viruses: Lessons learned using next-generation sequencing technologies. *Eur J Cancer* **61**, 61–68 (2016).

314. Hurley, E. A. *et al.* When Epstein-Barr virus persistently infects B-cell lines, it frequently integrates. *J Virol* **65**, 1245–54 (1991).

315. Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. *J Virol* **84**, 12100–9 (2010).

316. Forster, M. *et al.* Vy-PER: Eliminating false positive detection of virus integration events in next generation sequencing data. *Sci Rep* **5**, (2015).

317. Braunagel, S. C. & Summers, M. D. Autographa californica Nuclear Polyhedrosis Virus, PDV, and ECV Viral Envelopes and Nucleocapsids: Structural Proteins, Antigens, Lipid and Fatty Acid Profiles. *Virology* **202**, 315–328 (1994).

318. McDougal, V. v. & Guarino, L. A. The Autographa californica Nuclear Polyhedrosis Virus p143 Gene Encodes a DNA Helicase . *J Virol* **74**, 5273–5279 (2000).

319. van Loo, N.-D. *et al.* Baculovirus Infection of Nondividing Mammalian Cells: Mechanisms of Entry and Nuclear Transport of Capsids. *J Virol* **75**, 961–970 (2001).

320. Whittaker, G. R. & Helenius, A. Nuclear Import and Export of Viruses and Virus Genomes. *Virology* **246**, 1–23 (1998).

321. Bannert, N., Hofmann, H., Block, A. & Hohn, O. HERVs New Role in Cancer: From Accused Perpetrators to Cheerful Protectors. *Front Microbiol* **9**, (2018).

322. Zhang, M., Liang, J. Q. & Zheng, S. Expressional activation and functional roles of human endogenous retroviruses in cancers. *Rev Med Virol* **29**, (2019).

323. Li, W. *et al.* A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-k in human populations. *PLoS Comput Biol* **15**, (2019).

324. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

325. Chen, X. & Li, D. ERVcaller: Identify polymorphic endogenous retrovirus (ERV) and other transposable element (TE) insertions using whole-genome sequencing data. *bioRxiv* 332833 (2018) doi:10.1101/332833.

326. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Oxford University Press* 1–3 (2013).

327. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics Applications Note* **26**, 841–842 (2010).

328. bedtools: closest. https://bedtools.readthedocs.io/en/latest/content/tools/closest.html?highlight=closest (2022).

329. Chen, N., Zhang, G., Fu, J. & Wu, Q. Identification of Key Modules and Hub Genes Involved in Esophageal Squamous Cell Carcinoma Tumorigenesis Using WCGNA. *Cancer Control* **27**, (2020).

330. Xian, Q. & Zhu, D. The Involvement of WDHD1 in the Occurrence of Esophageal Cancer as a Downstream Target of PI3K/AKT Pathway. *J Oncol* **2022**, (2022).

331. Zare, M., Hadi, F. & Alivand, M. R. Considering the downregulation of Tpm1.6 and Tpm1.7 in squamous cell carcinoma of esophagus as a potent biomarker. *Future Medicine* **15**, 361–370 (2018).

332. Li, C. *et al.* A four-DNA methylation signature as a novel prognostic biomarker for survival of patients with gastric cancer. *Cancer Cell Int* **20**, 1–10 (2020).

333. Mao, C. *et al.* SMARCA6-LINC00559-ZBTB18 Axis Accelerates Cancer Progression Depending on LINC00559. *SSRN Electronic Journal* (2019) doi:10.2139/SSRN.3498577.

334. Anwar, S. L., Wulaningsih, W. & Lehmann, U. Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *Int J Mol Sci* **18**, 974 (2017).

335. Roy-Engel, A. M. *et al.* Human retroelements may introduce intragenic polyadenylation signals. *Cytogenet Genome Res* **110**, 365–371 (2005).

336. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**, 21–42 (2012).

337. Wang, Y. *et al.* Identification of GGT5 as a Novel Prognostic Biomarker for Gastric Cancer and its Correlation With Immune Cell Infiltration. *Front Genet* **13**, 599 (2022).

338. Sisu, C. Pseudogenes as Biomarkers and Therapeutic Targets in Human Cancers. *Methods in Molecular Biology* **2324**, 319–337 (2021).

339. Li, J., Xu, W. & Zhu, Y. Mammaglobin B may be a prognostic biomarker of uterine corpus endometrial cancer. *Oncol Lett* **20**, 1–1 (2020).

340. Guo, W. *et al.* Loss of SUSD2 expression correlates with poor prognosis in patients with surgically resected lung adenocarcinoma. *J Cancer* **11**, 1648–1656 (2020).

341. Song, J. *et al.* A Novel Ferroptosis-Related Biomarker Signature to Predict Overall Survival of Esophageal Squamous Cell Carcinoma. *Front Mol Biosci* **8**, 607 (2021).

342. Maccarty, W. C. Identification of the cancer cell. *J Am Med Assoc* **107**, 844–845 (1936).

343. von Hansemann, D. Ueber asymmetrische Zelltheilung in epithel Krebsen und deren biologische Bedeutung. *Virchows Arch. Path. Anat* **119**, 299 (1890).

344. Boveri, T. Zur Frage der Entstehung Maligner Tumoren. in (Fischer, G, 1914).

345. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

346. Avery, O. T., MacLeod, C. M. & McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumoccal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumoccus type III. *J Exp Med* **79**, 37–58 (1944).

347. Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290**, 261–264 (1981).

348. Krontiris, T. G. & Cooper, G. M. Transforming activity of human tumor DNAs. *Proc Natl Acad Sci U S A* **78**, 1181–1184 (1981).

349. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).

350. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science (1979)* **349**, 1483–1489 (2015).

351. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820–823 (1971).

352. Chapman, B. *et al.* bcbio/bcbio-nextgen: Preprint at https://doi.org/10.5281/ZENODO.5781867 (2021).

353. Guimera, R. V. bcbio-nextgen: Automated, distributed next-gen sequencing pipeline. *EMBnet J* **17**, 30 (2011).

354. bcbio-nextgen 1.2.9 documentation. https://bcbio-nextgen.readthedocs.io/en/latest/index.html (2022).

355. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

356. Mark PCR duplicates and sequencing platform / optical duplicates in single-end SAM files. https://github.com/afontenot/mark-duplicates (2019).

357. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012) doi:10.48550/arxiv.1207.3907.

358. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108–e108 (2016).

359. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).

360. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* **15**, 1–19 (2014).

361. Kelly, B. J. *et al.* Churchill: An ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol* **16**, 1–14 (2015).

362. Faust, G. G. & Hall, I. M. SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

363. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol* **9**, e1003153 (2013).

364. GEMINI. GEMINI: gemini 0.20.1 documentation. https://gemini.readthedocs.io/en/latest/index.html (2022).

365. SnpEff & SnpSift Documentation. https://pcingola.github.io/SnpEff/se_introduction/ (2022).

366. SUSD2 - UniProt. https://www.uniprot.org/uniprotkb/Q9UGT4/entry (2022).

367. DUX4 - UniProt. https://www.uniprot.org/uniprotkb/Q9UBX2/entry (2022).

368. LINC00559 - NCBI. https://www.ncbi.nlm.nih.gov/gene/100874187 (2022).

369. WDHD1 - UniProt. https://www.uniprot.org/uniprotkb/O75717/entry (2022).

370. CAPN8 - UniProt. https://www.uniprot.org/uniprotkb/A6NHC0/entry (2022).

371. Janco, M. *et al.* α-Tropomyosin with a D175N or E180G mutation in only one chain differs from tropomyosin with mutations in both chains. *Biochemistry* **51**, 9880–9890 (2012).

372. MUC3A Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=MUC3A (2022).

373. TP53 Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=TP53 (2022).

374. Bellini, M. F., Cadamuro, A. C. T., Succi, M., Proença, M. A. & Silva, A. E. Alterations of the TP53 gene in gastric and esophageal carcinogenesis. *J Biomed Biotechnol* **2012**, (2012).

375. Yao, L. *et al.* Investigation on the Potential Correlation Between TP53 and Esophageal Cancer. *Front Cell Dev Biol* **9**, 730337 (2021).

376. HEATR9 Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=HEATR9 (2022).

377. VPS52 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=VPS52 (2022).

378. NCOR1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NCOR1 (2022).

379. AHNAK Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=AHNAK (2022).

380. Mai, Z. *et al.* Inactivation of Hippo pathway characterizes a poor-prognosis subtype of esophageal cancer. *JCI Insight* **7**, (2022).

381. OR4D10 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=OR4D10 (2022).

382. CDKN2A Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDKN2A (2022).

383. Zhou, C., Li, J. & Li, Q. CDKN2A methylation in esophageal cancer: a meta-analysis. *Oncotarget* **8**, 50083 (2017).

384. Suzuki, H. *et al.* Intragenic mutations of CDKN2B and CDKN2A in primary human esophageal cancers. *Hum Mol Genet* **4**, 1883–1887 (1995).

385. Hu, N. *et al.* High frequency of CDKN2A alterations in esophageal squamous cell carcinoma from a high-risk Chinese population. *Genes Chromosomes Cancer* **39**, 205–216 (2004).

386. FIGN Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FIGN (2022).

387. OR4D11 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=OR4D11 (2022).

388. KMT2D Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KMT2D (2022).

389. CST3 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CST3 (2022).

390. Yan, Y. *et al.* LncRNA Snhg1, a non-degradable sponge for miR-338, promotes expression of proto-oncogene CST3 in primary esophageal cancer cells. *Oncotarget* **8**, 35760 (2017).

391. GRIN2A Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=GRIN2A (2022).

392. DEF8 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=DEF8 (2022).

393. FGFR1 Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=FGFR1 (2022).

394. Lin, D. C., Wang, M. R. & Koeffler, H. P. Targeting genetic lesions in esophageal cancer. *Cell Cycle* **13**, 2013–2014 (2014).

395. Takase, N. *et al.* NCAM- and FGF-2-mediated FGFR1 signaling in the tumor microenvironment of esophageal cancer regulates the survival and migration of tumor-associated macrophages and cancer cells. *Cancer Lett* **380**, 47–58 (2016).

396. Von Loga, K. *et al.* FGFR1 Amplification Is Often Homogeneous and Strongly Linked to the Squamous Cell Carcinoma Subtype in Esophageal Carcinoma. *PLoS One* **10**, e0141867 (2015).

397. TBL1X Gene - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=TBL1X (2022).

398. Liu, L. *et al.* TBL1XR1 promotes lymphangiogenesis and lymphatic metastasis in esophageal squamous cell carcinoma. *Gut* **64**, 26–36 (2015).

399. FAM135A - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=FAM135A (2022).

400. TDG Gene - GeneCards | TDG Protein | TDG Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TDG (2022).

401. Zhou, W., Zhang, L., Chen, P., Li, S. & Cheng, Y. Thymine DNA glycosylase-regulated TAZ promotes radioresistance by targeting nonhomologous end joining and tumor progression in esophageal cancer. *Cancer Sci* **111**, 3613–3625 (2020).

402. CST5 - GeneCards . https://www.genecards.org/cgi-bin/carddisp.pl?gene=CST5&keywords=CST5 (2022).

403. SLC4A3 - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SLC4A3 (2022).

404. Gum, J. R. *et al.* Initiation of Transcription of the MUC3A Human Intestinal Mucin from a TATA-less Promoter and Comparison with the MUC3B Amino Terminus *. *Journal of Biological Chemistry* **278**, 49600–49609 (2003).

405. Volin, M. V., Shahrara, S., Haines, G. K., Woods, J. M. & Koch, A. E. Expression of mucin 3 and mucin 5AC in arthritic synovial tissue. *Arthritis Rheum* **58**, 46–52 (2008).

406. Wang, J. *et al.* Serum mucin 3A as a potential biomarker for extrahepatic cholangiocarcinoma. *Saudi Journal of Gastroenterology* **26**, 129 (2020).

407. Genome Data Viewer - NCBI. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/gene/?id=4584 (2022).

408. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res* **77**, e31–e34 (2017).

409. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013).

410. CreateSomaticPanelOfNormals (BETA) – GATK. https://gatk.broadinstitute.org/hc/en-us/articles/9570531313563-CreateSomaticPanelOfNormals-BETA-.

411. GATK. Mutect2 – GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2 (2022).

412. GATK. https://gatk.broadinstitute.org/hc/en-us (2023).

413. Bian, X. *et al.* Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19**, 1–11 (2018).

414. Shipanga, H. Unpublished work. *PhD Candidate in the Iqbal Parker laboratory* Preprint at (2022).

415. Pratt, W. S. *et al.* Multiple transcripts of MUC3: Evidence for two genes, MUC3A and MUC3B. *Biochem Biophys Res Commun* **275**, 916–923 (2000).

416. Lu, P. *et al.* Genome-Wide Gene Expression Profile Analyses Identify CTTN as a Potential Prognostic Marker in Esophageal Cancer. *PLoS One* **9**, e88918 (2014).

417. Visser, E., Franken, I. A., Brosens, L. A. A., Ruurda, J. P. & van Hillegersberg, R. Prognostic gene expression profiling in esophageal cancer: a systematic review. *Oncotarget* **8**, 5566 (2017).

418. Mimura, K. *et al.* Lapatinib inhibits receptor phosphorylation and cell growth and enhances antibody-dependent cellular cytotoxicity of EGFR- and HER2-overexpressing esophageal cancer cell lines. *Int J Cancer* **129**, 2408–2416 (2011).

419. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11* **20**, 631–656 (2019).

420. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).

421. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).

422. Berger, M. F. *et al.* Integrative analysis of the melanoma transcriptome. *Genome Res* **20**, 413–427 (2010).

423. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).

424. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).

425. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**, 1–13 (2013).

426. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12**, e0190152 (2017).

427. Li, P., Piao, Y., Shon, H. S. & Ryu, K. H. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* **16**, 1–9 (2015).

428. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol* **11**, 1–10 (2010).

429. Steinbaugh, M. J. *et al.* bcbioRNASeq: R package for bcbio RNA-seq analysis. *F1000Research 2018 6:1976* **6**, 1976 (2018).

430. STAR: RNA-seq aligner. https://github.com/alexdobin/STAR (2022).

431. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15 (2013).

432. hbc/bcbioRNASeq. https://github.com/hbc/bcbioRNASeq (2022).

433. R Core Team: A language and environment for statistical computing. . Preprint at https://www.R-project.org/ (2020).

434. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292 (2016).

435. Jolliffe, I. T. *Principal Component Analysis*. (Springer-Verlag, 2002). doi:10.1007/B98835.

436. acidgenomics/r-deseqanalysis. https://github.com/acidgenomics/r-deseqanalysis/ (2022).

437. Dündar, F., Skrabanek, L. & Zumbo, P. Introduction to differential gene expression analysis using RNA-seq. *Applied Bioinformatics Core* 1–97 Preprint at (2015).

438. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

439. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc* **58**, 236–244 (1963).

440. Li, Y., Wang, X., Shi, L., Xu, J. & Sun, B. Predictions for high COL1A1 and COL10A1 expression resulting in a poor prognosis in esophageal squamous cell carcinoma by bioinformatics analyses. *Transl Cancer Res* **9**, 85 (2020).

441. Song, Y. *et al.* Identification of four genes and biological characteristics of esophageal squamous cell carcinoma by integrated bioinformatics analysis. *Cancer Cell Int* **21**, 1–12 (2021).

442. Liu, M. *et al.* MMP1 promotes tumor growth and metastasis in esophageal squamous cell carcinoma. *Cancer Lett* **377**, 97–104 (2016).

443. Li, H. *et al.* CCAAT/Enhancer Binding Protein β-Mediated MMP3 Upregulation Promotes Esophageal Squamous Cell Cancer Invasion In Vitro and Is Associated with Metastasis in Human Patients. *https://home.liebertpub.com/gtmb* **23**, 304–309 (2019).

444. Wang, D. *et al.* PAI-1 overexpression promotes invasion and migration of esophageal squamous carcinoma cells. *Yi Chuan* **42**, 287–295 (2020).

445. Ozawa, D. *et al.* TGFBI Expression in Cancer Stromal Cells is Associated with Poor Prognosis and Hematogenous Recurrence in Esophageal Squamous Cell Carcinoma. *Ann Surg Oncol* **23**, 282–289 (2016).

446. Liu, S. *et al.* FTO promotes cell proliferation and migration in esophageal squamous cell carcinoma through up-regulation of MMP13. *Exp Cell Res* **389**, (2020).

447. Wang, S., You, L., Dai, M. & Zhao, Y. Mucins in pancreatic cancer: A well-established but promising family for diagnosis, prognosis and therapy. *J Cell Mol Med* **24**, 10279–10289 (2020).

448. Li, X., Xiao, X., Chang, R. & Zhang, C. Comprehensive bioinformatics analysis identifies lncRNA HCG22 as a migration inhibitor in esophageal squamous cell carcinoma. *J Cell Biochem* **121**, 468–481 (2020).

449. Dozmorov, M. G. & Valencia, A. Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. *Bioinformatics* **33**, 3323–3330 (2017).

450. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

451. Li, J. *et al.* The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. *PeerJ* **2019**, e7705 (2019).

452. Li, G. *et al.* High expression of collagen 1A2 promotes the proliferation and metastasis of esophageal cancer cells. *Ann Transl Med* **8**, 1672–1672 (2020).

453. Kang, Z. *et al.* COL11A1 promotes esophageal squamous cell carcinoma proliferation and metastasis and is inversely regulated by miR-335-5p. *Ann Transl Med* **9**, 1577–1577 (2021).

454. Senthebane, D. A. *et al.* The Role of Tumor Microenvironment in Chemoresistance: 3D Extracellular Matrices as Accomplices. *International Journal of Molecular Sciences 2018, Vol. 19, Page 2861* **19**, 2861 (2018).

455. Palumbo, A. *et al.* Esophageal Cancer Development: Crucial Clues Arising from the Extracellular Matrix. *Cells* **9**, 455 (2020).

456. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).

457. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267–273 (2003).

458. Pappu, R., Ramirez-Carrozzi, V. & Sambandam, A. The interleukin-17 cytokine family: critical players in host defence and inflammatory diseases. *Immunology* **134**, 8–16 (2011).

459. Lu, L. *et al.* IL-17A promotes immune cell recruitment in human esophageal cancers and the infiltrating dendritic cells represent a positive prognostic marker for patient survival. *Journal of Immunotherapy* **36**, 451–458 (2013).

460. Wang, Y., Wu, H., Wu, X., Bian, Z. & Gao, Q. Interleukin 17A Promotes Gastric Cancer Invasiveness via NF-κB Mediated Matrix Metalloproteinases 2 and 9 Expression. *PLoS One* **9**, e96678 (2014).

461. Ren, H. *et al.* IL-17A Promotes the Migration and Invasiveness of Colorectal Cancer Cells Through NF-κB-Mediated MMP Expression. *Oncol Res* **23**, 249 (2016).

462. Liu, D. *et al.* Interleukin-17A promotes esophageal adenocarcinoma cell invasiveness through ROS-dependent, NF-κB-mediated MMP-2/9 activation. *Oncol Rep* **37**, 1779–1785 (2017).

463. Alcami, A. Viral mimicry of cytokines, chemokines and their receptors. *Nat Rev Immunol* **3**, 36–50 (2003).

464. KEGG PATHWAY: hsa04061. https://www.genome.jp/dbget-bin/www_bget?hsa04061 (2002).

465. Michaelis, M., Doerr, H. W. & Cinatl, J. The story of human cytomegalovirus and cancer: Increasing evidence and open questions. *Neoplasia* **11**, 1–9 (2009).

466. KEGG PATHWAY: hsa05163. https://www.genome.jp/dbget-bin/www_bget?pathway+hsa05163 (2022).

467. Ramos-Vara, J. A. Principles and methods of immunohistochemistry. in *Methods in molecular biology* (ed. Gautier, J.-C.) vol. 691 83–96 (Humana Press, 2011).

468. Yong, W. H., Dry, S. M. & Shabihkhani, M. A practical approach to clinical and research biobanking. in *Methods in Molecular Biology* vol. 1180 137–162 (Humana Press Inc., 2014).

469. Fitzgibbons, P. L. *et al.* Principles of Analytic Validation of Immunohistochemical Assays: Guideline From the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med* **138**, 1432–1443 (2014).

470. Cregger, M., Berger, A. J. & Rimm, D. L. Immunohistochemistry and Quantitative Analysis of Protein Expression. *Arch Pathol Lab Med* **130**, 1026–1030 (2006).

471. D'Amico, F., Skarmoutsou, E. & Stivala, F. State of the art in antigen retrieval for immunohistochemistry. *J Immunol Methods* **341**, 1–18 (2009).

472. Lin, F. & Chen, Z. Standardization of Diagnostic Immunohistochemistry: Literature Review and Geisinger Experience. *Arch Pathol Lab Med* **138**, 1564–1577 (2014).

473. Higashi, M. *et al.* Immunohistochemical study of mucin expression in periampullary adenomyoma. *J Hepatobiliary Pancreat Sci* **17**, 275–283 (2010).

474. Park, H. U. *et al.* Aberrant expression of muc3 and muc4 membrane-associated mucins and sialyl lex antigen in pancreatic intraepithelial neoplasia. *Pancreas* **26**, e48–e54 (2003).

475. Cui, J. *et al.* Comprehensive characterization of the genomic alterations in human gastric cancer. *Int J Cancer* **137**, 86–95 (2015).

Appendix 1

Table A1.1: Quality control metrics of WGS data reads performed by the Wellcome Sanger Institute. % Map indicates the percentage of reads per sample that mapped to the reference genome (Hg37). %R1 UM and % R2 UM indicate the percentage of unmapped reads for read 1 and read 2 respectively. R1 GC dev and R2 GC dev indicate the deviation of GC content from whole genome average read 1 and read 2 respectively. GC dist indicates the difference in GC content between read 1 and read 2. Ins size sd indicates the standard deviation of the insert size. Edit dist indicates the fold change of mismatched bases in read 1 vs read 2. Map dist indicates the difference between percentage unmapped read 1 and read 2 expressed as a fraction.

UCT Samples

Sample		% Map	% R1 UM	% R2 UM	R1 GC dev	R2 GC dev	GC dist	Ins size sd	Edit dist	Map dist
PD39445a	Tumour	99,38375	0,11	1,11875	2,835	2,88125	0,04875	217,39875	1,63875	0,01
PD39445b	Normal	99,6675	0,07	0,59875	0,035	0,33625	0,30125	139,7225	1,90875	0,0075
PD39446a	Tumour	99,55125	0,1325	0,7625	2,82125	2,9325	0,11125	570,40375	1,64625	0,01
PD39446b	Normal	99,5775	0,10125	0,74125	2,9025	3,265	0,365	136,6125	2,0625	0,01
PD39447a	Tumour	99,77375	0,03125	0,41875	0,185	0,4025	0,21875	127,86	1,74875	0
PD39447b	Normal	99,58	0,03	0,81	-0,17625	0,24875	0,42375	139,16625	2,09875	0,01
PD39448a	Tumour	99,5925	0,2125	0,5975	-0,06125	0,16625	0,225	125,79	1,79375	0
PD39448b	Normal	99,6025	0,055	0,74	0,02875	0,3975	0,36875	141,4275	1,97125	0,01
PD39449a	Tumour	99,4575	0,36	0,74	0,2925	0,53625	0,2425	127,78	1,77	0
PD39449b	Normal	99,61375	0,06	0,7125	0,00375	0,38125	0,3775	135,28375	2,07375	0,01
PD39450a	Tumour	99,80375	0,03125	0,35625	0,27625	0,45125	0,1825	122,88875	1,7025	0
PD39450b	Normal	99,6575	0,05	0,635	0,0375	0,38875	0,35	133,63	2,0225	0,01
PD39451a	Tumour	99,81375	0,03125	0,33875	0,11875	0,28375	0,16625	123,8675	1,65	0
PD39451b	Normal	99,66125	0,1025	0,57375	0,14625	0,42375	0,2775	124,1425	1,875	0,00375
PD39452a	Tumour	99,78375	0,03125	0,39375	0,2025	0,39875	0,195	124,745	1,705	0
PD39452b	Normal	99,71625	0,04875	0,51875	-0,15125	0,1075	0,2625	127,47625	1,88375	0,00375
PD39453a	Tumour	99,8125	0,0625	0,3125	0,0025	0,17875	0,17625	123,71375	1,56125	0
PD39453b	Normal	99,70875	0,0625	0,5425	-0,1325	0,18	0,31375	126,95375	1,90625	0,00375
PD39454a	Tumour	99,7275	0,14	0,405	-0,665	-0,4875	0,18125	123,7575	1,58375	0
PD39454b	Normal	99,69125	0,07	0,55125	-0,02625	0,2275	0,255	125,4325	1,91125	0,00375
PD39455a	Tumour	99,52	0,29	0,6725	-0,12375	0,09125	0,21	127,48875	1,75125	0

PD39455b	Normal	99,73	0,05875	0,48125	-0,04625	0,20375	0,2525	128,485	1,82	0
PD39456a	Tumour	99,78	0,07	0,3675	-1,13	-0,96	0,1725	125,52875	1,605	0
PD39456b	Normal	99,67625	0,0625	0,58625	0,01375	0,3075	0,29375	129,29375	1,905	0,0075
PD39457c	Tumour	99,8075	0,03875	0,35125	-1,675	-1,505	0,17	123,39	1,6425	0
PD39457b	Normal	99,55125	0,04125	0,85375	-0,11625	0,08375	0,2	121,78	1,96	0,01
PD39458c	Tumour	99,7425	0,09	0,4275	-1,37125	-1,20375	0,17	128,07	1,6475	0
PD39458b	Normal	99,58625	0,03	0,80125	-0,27375	-0,03	0,24375	120,49125	2	0,01
PD39459a	Tumour	99,28333	0,57	0,866666667	-1,046666667	-0,95	0,096666667	107,0633333	1,523333	0
PD39459b	Normal	99,55125	0,06125	0,8375	-0,17125	0,02875	0,19875	117,30375	1,965	0,01
PD39460a	Tumour	99,73667	0,046666667	0,483333333	0,163333333	0,303333333	0,143333333	112,2533333	1,71	0
PD39460b	Normal	99,5775	0,06125	0,78	-0,14875	0,04625	0,19375	117,795	1,9725	0,01
PD50649a	Tumour	96,2225	1,345	6,21	-0,5025	-0,2725	0,2275	154,5825	1,5625	0,0525
PD50649b	Normal	96,0575	1,395	6,485	-0,49	-0,2275	0,255	166,065	1,5075	0,0525
PD50650a	Tumour	96,2	1,465	6,14	-0,4675	-0,22	0,2475	153,715	1,5375	0,0525
PD50650b	Normal	96,045	1,395	6,5125	-0,2275	0,04	0,27	147,4975	1,53	0,0525
PD50651a	Tumour	96,3075	1,2425	6,14	-0,205	0,0525	0,2575	154,49	1,325	0,0525
PD50651b	Normal	96,925	1,2825	4,8675	-0,44	-0,2075	0,2325	138,95	1,2975	0,04
PD50653a	Tumour	96,0825	1,2425	6,5925	-0,335	-0,0275	0,305	158,405	1,3675	0,055
PD50653b	Normal	96,2525	1,31	6,185	-0,33	-0,0775	0,25	148,7225	1,325	0,0525
PD51372a	Tumour	93,57	1,51	11,345	-0,145	0,52	0,67	165,165	1,79	0,11
PD51372b	Normal	94,72	1,53	9,03	0,125	0,62	0,495	140,74	1,735	0,085

WITS Samples

Sample		% Map	% R1 UM	% R2 UM	R1 GC dev	R2 GC dev	GC dist	Ins size sd	Edit dist	Map dist
PD44691a	Tumour	97,77667	1,011666667	3,435	-0,343333333	-0,245	0,098333333	136,455	1,311667	0,023333
PD44691b	Normal	96,30167	1,243333333	6,153333333	-0,056666667	0,206666667	0,261666667	151,4766667	1,476667	0,051667
PD44692a	Tumour	84,57	14,22333333	16,63833333	0,066666667	0,188333333	0,12	133,71	1,311667	0,031667
PD44692b	Normal	95,955	1,26	6,83	0,775	1,068333333	0,293333333	149,4916667	1,523333	0,06
PD44693a	Tumour	97,225	1,115	4,438333333	-0,388333333	-0,258333333	0,13	139,02	1,343333	0,033333
PD44693b	Normal	96,135	1,356666667	6,373333333	0,201666667	0,466666667	0,265	149,3983333	1,446667	0,055

PD44694a	Tumour	97,27833	1,21	4,23	-0,483333333	-0,356666667	0,126666667	138,1733333	1,331667	0,033333
PD44694b	Normal	95,71667	1,293333333	7,268333333	0,206666667	0,545	0,336666667	145,59	1,526667	0,065
PD44695a	Tumour	97,32333	1,238333333	4,111666667	-0,585	-0,511666667	0,075	137,4133333	1,333333	0,031667
PD44695b	Normal	96,43833	1,125	5,998333333	0,435	0,661666667	0,225	146,26	1,545	0,051667
PD44696a	Tumour	97,11833	1,275	4,488333333	-0,21	-0,098333333	0,113333333	139,1733333	1,333333	0,033333
PD44696b	Normal	96,27	1,16	6,3	0,043333333	0,305	0,261666667	161,3333333	1,531667	0,055
PD44697a	Tumour	95,675	1,91	6,741666667	-0,226666667	-0,028333333	0,255	144,3833333	1,295	0,053333
PD44697b	Normal	95,39333	1,431666667	7,781666667	-0,268333333	-0,208333333	0,476666667	155,56	1,48	0,068333
PD44698a	Tumour	95,935	1,746666667	6,381666667	-0,411666667	-0,198333333	0,216666667	143,4266667	1,311667	0,05
PD44698b	Normal	95,335	1,488333333	7,84	-0,175	0,25	0,425	152,9116667	1,483333	0,068333
PD44699a	Tumour	95,715	1,788333333	6,785	-0,225	0,043333333	0,268333333	144,6033333	1,316667	0,053333
PD44699b	Normal	95,92333	1,36	6,795	-0,37	-0,015	0,358333333	151,7016667	1,471667	0,056667
PD44700a	Tumour	96,596	1,298	5,51	-1,304	-1,138	0,166	145,126	1,358	0,046
PD44700b	Normal	94,57333	1,456666667	9,4	-0,148333333	-0,458333333	0,608333333	149,1183333	1,575	0,088333
PD44701a	Tumour	96,996	1,2	4,81	-0,3	-0,172	0,126	146,864	1,324	0,036
PD44701b	Normal	95,26667	1,473333333	7,993333333	-0,135	0,338333333	0,471666667	153,2533333	1,486667	0,071667
PD44702a	Tumour	96,714	1,254	5,312	-0,256	-0,13	0,128	143,704	1,342	0,044
PD44702b	Normal	96,18333	1,366666667	6,265	-0,52	-0,2	0,32	149,7	1,425	0,055
PD44703a	Tumour	96,64	1,246	5,468	-0,182	0,012	0,192	138,324	1,386	0,046
PD44703b	Normal	95,965	1,838333333	6,23	-0,533333333	-0,318333333	0,218333333	138,7733333	1,301667	0,048333
PD44704a	Tumour	96,686	1,266	5,36	-0,23	-0,09	0,144	141,422	1,34	0,044
PD44704b	Normal	95,88667	1,861666667	6,366666667	-0,34	-0,135	0,203333333	142,26	1,291667	0,05

Appendix 2

Table A2.1 GEMINI Impact and Impact_severity term mappings of the variant/variant_impacts tables within the output database files (GEMINI 2022).

GEMINI terms (impact)	Impact severity	GEMINI terms (impact)	Impact severity
splice_acceptor	HIGH	inframe_codon_loss	MED
splice_donor	HIGH	inframe_codon_change	MED
stop_gain	HIGH	codon_change_del	MED
stop_loss	HIGH	codon_change_ins	MED
frame_shift	HIGH	UTR_5_del	MED
start_loss	HIGH	UTR_3_del	MED
exon_deleted	HIGH	splice_region	MED
non_synonymous_start	HIGH	mature_miRNA	MED
transcript_codon_change	HIGH	regulatory_region	MED
chrom_large_del	HIGH	TF_binding_site	MED
rare_amino_acid	HIGH	regulatory_region_ablation	MED
		regulatory_region_amplification	MED
		TFBS_ablation	MED
		non_syn_coding	MED
		inframe_codon_gain	MED
		TFBS_amplification	MED

GEMINI terms (impact)	Impact severity	GEMINI terms (impact)	Impact severity
synonymous_stop	LOW	start_gain	LOW
synonymous_coding	LOW	synonymous_start	LOW
UTR_5_prime	LOW	intron_conserved	LOW
UTR_3_prime	LOW	nc_transcript	LOW
intron	LOW	NMD_transcript	LOW
CDS	LOW	incomplete_terminal_codon	LOW
upstream	LOW	nc_exon	LOW
downstream	LOW	transcript_ablation	LOW
intergenic	LOW	transcript_amplification	LOW
intergenic_conserved	LOW	feature elongation	LOW
intragenic	LOW	feature truncation	LOW
gene	LOW	transcript	LOW
exon	LOW		

MED = Medium impact severity

Appendix 3

Table A3.1: All HIGH impact severity *MUC3A* mutations detected through bioinformatics analysis of WGS data using the bcbio-nextgen (Chapman et al. 2021) pipeline. Start positions correspond to reference genome GRCh38 and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Start Position	Gene	Ref	Alt	Impact	Impact Severity
PD39445	100952610	MUC3A	C	T	stop_gained	HIGH
PD39445	100952646	MUC3A	AGG	A	frameshift_variant	HIGH
PD39445	100952649	MUC3A	AC	A	frameshift_variant	HIGH
PD39445	100958132	MUC3A	G	GAAGAC	frameshift_variant	HIGH
PD39445	100958143	MUC3A	A	ACTTCTTTGATAGCCAC	frameshift_variant	HIGH
PD39447	100957681	MUC3A	C	GCTCACACAGTACCCTCGGCTTAAGTTCTTCG	frameshift_variant	HIGH
PD39447	100957684	MUC3A	C	CGTCACCA	frameshift_variant	HIGH
PD39447	100957690	MUC3A	A	ACCTCACACAGTACTGCCAGCTTCACTTCTTCGATCACTACCACCGAGACACTATCACATAG	stop_gained	HIGH
PD39447	100957699	MUC3A	G	GCTTGACTCCTGTGAACACCACCACTGAAACCACGTCAAACAGTATT	frameshift_variant	HIGH
PD39447	100957701	MUC3A	C	CAGGCT	frameshift_variant	HIGH
PD39448	100957775	MUC3A	C	CG	frameshift_variant	HIGH
PD39448	100957776	MUC3A	A	AGT	frameshift_variant	HIGH
PD39448	100957779	MUC3A	AC	A	frameshift_variant	HIGH
PD39448	100957784	MUC3A	CT	C	frameshift_variant	HIGH
PD39448	100957787	MUC3A	AC	A	frameshift_variant	HIGH
PD39448	100958184	MUC3A	A	ACC	frameshift_variant	HIGH
PD39448	100958186	MUC3A	C	CTGA	stop_gained	HIGH
PD39448	100958187	MUC3A	G	GA	frameshift_variant	HIGH
PD39448	100958216	MUC3A	T	TGCCACCGAGA	frameshift_variant	HIGH
PD39448	100958219	MUC3A	A	AT	frameshift_variant	HIGH
PD39448	100958224	MUC3A	C	CA	frameshift_variant	HIGH
PD39448	100958230	MUC3A	C	CTT	frameshift_variant	HIGH
PD39448	100958234	MUC3A	G	GC	frameshift_variant	HIGH
PD39448	100958806	MUC3A	C	CCAAGACCACCTCAACCAGTCCTCCCAGATTCACCT	frameshift_variant	HIGH
PD39448	100958808	MUC3A	T	TG	frameshift_variant	HIGH
PD39448	100958856	MUC3A	TC	T	frameshift_variant	HIGH
PD39448	100958858	MUC3A	T	TGG	frameshift_variant	HIGH
PD39448	100958859	MUC3A	TC	T	frameshift_variant	HIGH
PD39451	100957775	MUC3A	C	CG	frameshift_variant	HIGH
PD39451	100957776	MUC3A	A	AGT	frameshift_variant	HIGH
PD39451	100957779	MUC3A	AC	A	frameshift_variant	HIGH
PD39451	100957784	MUC3A	CT	C	frameshift_variant	HIGH
PD39451	100957787	MUC3A	AC	A	frameshift_variant	HIGH
PD39452	100955243	MUC3A	TTC	T	frameshift_variant	HIGH
PD39452	100955247	MUC3A	TC	T	frameshift_variant	HIGH
PD39452	100958302	MUC3A	G	GCTTAA	frameshift_variant	HIGH
PD39452	100958303	MUC3A	G	GTTCT	frameshift_variant	HIGH
PD39452	100958352	MUC3A	A	ATCACCACCACTGGCAACACCTCACACAGT	frameshift_variant	HIGH
PD39452	100958354	MUC3A	T	TGCCGAGCTTC	frameshift_variant	HIGH

PD39452	100958356	MUC3A	C	CTTCTT	frameshift_variant	HIGH
PD39452	100958357	MUC3A	C	CG	frameshift_variant	HIGH
PD39452	100958809	MUC3A	C	CAATTGCCACCTCCA AGACCACCTCAACCA GTCCTCCCA	frameshift_variant	HIGH
PD39452	100958811	MUC3A	A	AT	frameshift_variant	HIGH
PD39453	100953731	MUC3A	CTC	T	frameshift_variant	HIGH
PD39453	100953737	MUC3A	T	TGG	frameshift_variant	HIGH
PD39453	100953808	MUC3A	A	AG	frameshift_variant	HIGH
PD39453	100953812	MUC3A	AG	A	frameshift_variant	HIGH
PD39453	100953834	MUC3A	A	AT	frameshift_variant	HIGH
PD39453	100953838	MUC3A	TC	T	frameshift_variant	HIGH
PD39453	100958653	MUC3A	C	CAAACAGTA	frameshift_variant	HIGH
PD39453	100958654	MUC3A	T	TTCCAGGC	frameshift_variant	HIGH
PD39453	100958656	MUC3A	T	TC	frameshift_variant	HIGH
PD39453	100958657	MUC3A	A	ACTTCTTCG	frameshift_variant	HIGH
PD39453	100958829	MUC3A	AAC	A	frameshift_variant	HIGH
PD39453	100958836	MUC3A	A	AAC	frameshift_variant	HIGH
PD39454	100958011	MUC3A	C	CAT	frameshift_variant	HIGH
PD39454	100958015	MUC3A	A	ATAGT	stop_gained	HIGH
PD39454	100958016	MUC3A	A	AC	frameshift_variant	HIGH
PD39454	100958018	MUC3A	C	CCTAGCTTG	stop_gained	HIGH
PD39454	100958995	MUC3A	CT	C	frameshift_variant	HIGH
PD39454	100958999	MUC3A	T	TC	frameshift_variant	HIGH
PD39455	100957959	MUC3A	TC	T	frameshift_variant	HIGH
PD39455	100957963	MUC3A	T	TC	frameshift_variant	HIGH
PD39455	100958184	MUC3A	A	ACCACCTC	frameshift_variant	HIGH
PD39455	100958188	MUC3A	C	CAGTACTCCTGGTTT	frameshift_variant	HIGH
PD39455	100958190	MUC3A	A	ACTTCTTCGATTGT	frameshift_variant	HIGH
PD39455	100958192	MUC3A	A	AC	frameshift_variant	HIGH
PD39455	100958195	MUC3A	C	CCG	frameshift_variant	HIGH
PD39455	100958197	MUC3A	G	GACCACCTCACCCCA	frameshift_variant	HIGH
PD39455	100958302	MUC3A	G	GCTTAA	frameshift_variant	HIGH
PD39455	100958303	MUC3A	G	GTTCT	frameshift_variant	HIGH
PD39455	100958352	MUC3A	A	ATCACCACCACTGGC AACACCTCACACAGT	frameshift_variant	HIGH
PD39455	100958354	MUC3A	T	TGCCGAGCTTC	frameshift_variant	HIGH
PD39455	100958356	MUC3A	C	CTGCCAGCTTCACTT CTTCAATCACCACCA CTGGCAACACCTCAC ACAGTATGCCGAGCT TCACTTCTT	frameshift_variant	HIGH
PD39455	100958356	MUC3A	C	CTTCTT	frameshift_variant	HIGH
PD39455	100958357	MUC3A	C	CG	frameshift_variant	HIGH
PD39455	100958357	MUC3A	C	CG	frameshift_variant	HIGH
PD39456	100953731	MUC3A	CTC	T	frameshift_variant	HIGH
PD39456	100953737	MUC3A	T	TGG	frameshift_variant	HIGH
PD39456	100953752	MUC3A	T	TA	frameshift_variant	HIGH
PD39456	100953758	MUC3A	TA	T	frameshift_variant	HIGH
PD39456	100953774	MUC3A	AT	A	frameshift_variant	HIGH
PD39456	100953777	MUC3A	C	CA	frameshift_variant	HIGH
PD39456	100957834	MUC3A	CAT	C	frameshift_variant	HIGH
PD39456	100957842	MUC3A	A	AC	frameshift_variant	HIGH
PD39456	100957843	MUC3A	A	AG	frameshift_variant	HIGH
PD39456	100958184	MUC3A	A	ACC	frameshift_variant	HIGH
PD39456	100958186	MUC3A	C	CTGA	stop_gained	HIGH
PD39456	100958187	MUC3A	G	GA	frameshift_variant	HIGH
PD39456	100958216	MUC3A	T	TGCCACCGAGA	frameshift_variant	HIGH
PD39456	100958219	MUC3A	A	AT	frameshift_variant	HIGH
PD39456	100958224	MUC3A	C	CA	frameshift_variant	HIGH

PD39456	100958230	MUC3A	C	CTT	frameshift_variant	HIGH
PD39456	100958234	MUC3A	G	GC	frameshift_variant	HIGH
PD39457	100953731	MUC3A	CTC	T	frameshift_variant	HIGH
PD39457	100953737	MUC3A	T	TGG	frameshift_variant	HIGH
PD39457	100953808	MUC3A	A	AG	frameshift_variant	HIGH
PD39457	100953812	MUC3A	AG	A	frameshift_variant	HIGH
PD39457	100953834	MUC3A	A	AT	frameshift_variant	HIGH
PD39457	100953838	MUC3A	TC	T	frameshift_variant	HIGH
PD39457	100958619	MUC3A	G	GTCTCCATATCACAC	frameshift_variant	HIGH
PD39457	100958621	MUC3A	G	GT	frameshift_variant	HIGH
PD39457	100958623	MUC3A	C	CT	frameshift_variant	HIGH
PD39457	100958626	MUC3A	C	CAGCATCACTACGTCAATCGCCACCACTGAGATCAT	frameshift_variant	HIGH
PD39457	100958646	MUC3A	T	TTGAGTTCTGCGATCACCACCACTGAGGCCATGTCACATAGTATTCCAGGCTTCACTTCTTTGATCACCACTGCTGAGGCTACCTCACACCTTCCTCC	frameshift_variant	HIGH
PD39457	100958647	MUC3A	C	CAGGTT	frameshift_variant	HIGH
PD39457	100958763	MUC3A	A	ACCACCACCTGTTCACTTCTTCTGTCGCCACCATGGAGACCACCTCACACAGTACTCCCAGCATCGCTACGTCAATCGCCACCACTGAGATCATCTCACACAGCACTCCCAGCTATGCTTCTTCAATTG	frameshift_variant	HIGH
PD39458	100958011	MUC3A	C	CAT	frameshift_variant	HIGH
PD39458	100958015	MUC3A	A	ATGGT	frameshift_variant	HIGH
PD39458	100958016	MUC3A	A	AC	frameshift_variant	HIGH
PD39458	100958018	MUC3A	C	CCTAGCTTG	stop_gained	HIGH
PD39458	100958184	MUC3A	A	AC	frameshift_variant	HIGH
PD39458	100958188	MUC3A	C	CT	frameshift_variant	HIGH
PD39458	100958198	MUC3A	TAC	T	frameshift_variant	HIGH
PD39459	100958250	MUC3A	AC	A	frameshift_variant	HIGH
PD39459	100958257	MUC3A	G	GT	frameshift_variant	HIGH
PD39460	100958183	MUC3A	GAACGCCACAC	G	frameshift_variant	HIGH
PD39460	100958196	MUC3A	AGTACT	A	frameshift_variant	HIGH
PD39460	100958204	MUC3A	CAACTTCACTTCTTCAA	C	frameshift_variant	HIGH
PD39460	100958224	MUC3A	CCACCACCGAG	C	frameshift_variant	HIGH
PD39460	100958235	MUC3A	AC	A	frameshift_variant	HIGH
PD39460	100958239	MUC3A	CA	C	frameshift_variant	HIGH
PD39460	100958249	MUC3A	TAC	T	frameshift_variant	HIGH
PD44691	100957885	MUC3A	CATCCCATAGTACTCCCAGCTTCACTT	C	frameshift_variant	HIGH
PD44691	100957913	MUC3A	TTCGATCACCACCGAGACCA	T	frameshift_variant	HIGH
PD44691	100957978	MUC3A	AG	A	frameshift_variant	HIGH
PD44691	100957982	MUC3A	CACCT	C	frameshift_variant	HIGH
PD44691	100957989	MUC3A	CACAGTACTCTCAGCT	C	frameshift_variant	HIGH

			ACACTACC TCA			
PD44691	100958018	MUC3A	CACCA	C	frameshift_variant	HIGH
PD44691	100958026	MUC3A	CCG	C	frameshift_variant	HIGH
PD44691	100958042	MUC3A	CAG	C	frameshift_variant	HIGH
PD44691	100958183	MUC3A	GAACGCCA CAC	G	frameshift_variant	HIGH
PD44691	100958196	MUC3A	AGTACT	A	frameshift_variant	HIGH
PD44691	100958204	MUC3A	CAACTTCC TTCTTCAA	C	frameshift_variant	HIGH
PD44691	100958224	MUC3A	CCACCACC GAG	C	frameshift_variant	HIGH
PD44691	100958235	MUC3A	AC	A	frameshift_variant	HIGH
PD44691	100958239	MUC3A	CA	C	frameshift_variant	HIGH
PD44691	100958249	MUC3A	TAC	T	frameshift_variant	HIGH
PD44692	100958586	MUC3A	A	ACTCCCAGCTTCACTT ACCACAGAGTACTCC AGCTTCACTTCTTTGA TCACTACCAGCGAGA TGACATCCCACAGTA CTCCCAGCTTGACTTT TTCAATCACCACCAC CAAAAGCACATCCTA CAGTC	stop_gained	HIGH
PD44693	100958184	MUC3A	A	AC	frameshift_variant	HIGH
PD44693	100958188	MUC3A	C	CT	frameshift_variant	HIGH
PD44693	100958198	MUC3A	TAC	T	frameshift_variant	HIGH
PD44693	100958904	MUC3A	A	ACTTCTGC	frameshift_variant	HIGH
PD44693	100958905	MUC3A	G	GA	frameshift_variant	HIGH
PD44693	100958907	MUC3A	T	TGCCAGCACCAAGAT CA	frameshift_variant	HIGH
PD44693	100958944	MUC3A	CT	C	frameshift_variant	HIGH
PD44693	100958948	MUC3A	T	TC	frameshift_variant	HIGH
PD44693	100958963	MUC3A	G	A	stop_gained	HIGH
PD44695	100955009	MUC3A	C	CG	frameshift_variant	HIGH
PD44695	100957642	MUC3A	C	CTCCCAGCTTCGGAC ATCCCACAGCTCTCT CAGCTACACTTCTTC GATCGCCACCAGAGA GACCCCCTCACACAC TGT	frameshift_variant	HIGH
PD44696	100954996	MUC3A	TC	T	frameshift_variant	HIGH
PD44696	100955002	MUC3A	C	CG	frameshift_variant	HIGH
PD44696	100957770	MUC3A	A	AT	frameshift_variant	HIGH
PD44696	100957773	MUC3A	A	AGT	frameshift_variant	HIGH
PD44696	100958298	MUC3A	A	AGTACTGCT	frameshift_variant	HIGH
PD44696	100958300	MUC3A	G	GCTTCACTTCTTCAA	frameshift_variant	HIGH
PD44696	100958304	MUC3A	G	GCTTCACTTCTTCAAT CA	frameshift_variant	HIGH
PD44696	100958581	MUC3A	A	ACAGTATTC	frameshift_variant	HIGH
PD44696	100958582	MUC3A	T	TAG	frameshift_variant	HIGH
PD44696	100958583	MUC3A	G	GCTTCACTCCTTCTAT CATCTCACCT	frameshift_variant	HIGH
PD44696	100958584	MUC3A	A	AG	frameshift_variant	HIGH
PD44697	100954996	MUC3A	TC	T	frameshift_variant	HIGH
PD44697	100955009	MUC3A	C	CG	frameshift_variant	HIGH
PD44697	100958920	MUC3A	C	CTT	frameshift_variant	HIGH
PD44697	100958927	MUC3A	G	GT	frameshift_variant	HIGH
PD44697	100958929	MUC3A	C	CT	frameshift_variant	HIGH
PD44697	100958932	MUC3A	C	CAGTTTGA	stop_gained	HIGH
PD44697	100958935	MUC3A	T	TCTCTGATC	frameshift_variant	HIGH

PD44697	100958936	MUC3A	A	AC	frameshift_variant	HIGH
PD44697	100958938	MUC3A	A	AC	frameshift_variant	HIGH
PD44697	100958943	MUC3A	A	ACCAG	frameshift_variant	HIGH
PD44697	100958947	MUC3A	C	CAG	frameshift_variant	HIGH
PD44697	100958955	MUC3A	A	AG	frameshift_variant	HIGH
PD44697	100958956	MUC3A	C	CTTAAG	stop_gained	HIGH
PD44697	100958996	MUC3A	T	TG	frameshift_variant	HIGH
PD44697	100959001	MUC3A	GC	G	frameshift_variant	HIGH
PD44698	100958183	MUC3A	GAACGCCACAC	G	frameshift_variant	HIGH
PD44698	100958196	MUC3A	AGTACT	A	frameshift_variant	HIGH
PD44698	100958204	MUC3A	CAACTTCACTTCTTCAA	C	frameshift_variant	HIGH
PD44698	100958224	MUC3A	CCACCACCGAG	C	frameshift_variant	HIGH
PD44698	100958235	MUC3A	AC	A	frameshift_variant	HIGH
PD44698	100958239	MUC3A	CA	C	frameshift_variant	HIGH
PD44698	100958249	MUC3A	TAC	T	frameshift_variant	HIGH
PD44699	100958183	MUC3A	GAACGCCACAC	G	frameshift_variant	HIGH
PD44699	100958196	MUC3A	AGTACT	A	frameshift_variant	HIGH
PD44699	100958204	MUC3A	CAACTTCACTTCTTCAA	C	frameshift_variant	HIGH
PD44699	100958224	MUC3A	CCACCACCGAG	C	frameshift_variant	HIGH
PD44699	100958235	MUC3A	AC	A	frameshift_variant	HIGH
PD44699	100958239	MUC3A	CA	C	frameshift_variant	HIGH
PD44699	100958249	MUC3A	TAC	T	frameshift_variant	HIGH
PD44700	100954996	MUC3A	TC	T	frameshift_variant	HIGH
PD44700	100957775	MUC3A	C	CG	frameshift_variant	HIGH
PD44700	100957776	MUC3A	A	AGT	frameshift_variant	HIGH
PD44700	100957779	MUC3A	AC	A	frameshift_variant	HIGH
PD44700	100957784	MUC3A	CT	C	frameshift_variant	HIGH
PD44700	100957787	MUC3A	AC	A	frameshift_variant	HIGH
PD44701	100954996	MUC3A	TC	T	frameshift_variant	HIGH
PD44701	100955009	MUC3A	C	CG	frameshift_variant	HIGH
PD44702	100955173	MUC3A	CG	C	frameshift_variant	HIGH
PD44702	100955175	MUC3A	ATG	A	frameshift_variant	HIGH
PD44702	100955244	MUC3A	TC	T	frameshift_variant	HIGH
PD44702	100955246	MUC3A	TTC	T	frameshift_variant	HIGH
PD44702	100958132	MUC3A	G	GAAGAC	frameshift_variant	HIGH
PD44702	100958143	MUC3A	A	ACTTCTTTGATAGCCAC	frameshift_variant	HIGH
PD44703	100953731	MUC3A	CTC	T	frameshift_variant	HIGH
PD44703	100953737	MUC3A	T	TGG	frameshift_variant	HIGH
PD44703	100953808	MUC3A	A	AG	frameshift_variant	HIGH
PD44703	100953812	MUC3A	AG	A	frameshift_variant	HIGH
PD44703	100953834	MUC3A	A	AT	frameshift_variant	HIGH
PD44703	100953838	MUC3A	TC	T	frameshift_variant	HIGH
PD44703	100955009	MUC3A	C	CG	frameshift_variant	HIGH
PD44704	100955243	MUC3A	TTC	T	frameshift_variant	HIGH
PD44704	100955247	MUC3A	TC	T	frameshift_variant	HIGH
PD44704	100958216	MUC3A	T	TTTGA	frameshift_variant	HIGH
PD44704	100958219	MUC3A	A	ACC	frameshift_variant	HIGH
PD44704	100958619	MUC3A	G	GTCTCCATATCACAC	frameshift_variant	HIGH
PD44704	100958621	MUC3A	G	GT	frameshift_variant	HIGH
PD44704	100958623	MUC3A	C	CT	frameshift_variant	HIGH
PD44704	100958626	MUC3A	C	CAGCATCACTACGTCAATCGCCACCACTGAGATCAT	frameshift_variant	HIGH

PD50649	100954996	MUC3A	TC	T	frameshift_variant	HIGH
PD50649	100955009	MUC3A	C	CG	frameshift_variant	HIGH
PD50649	100957774	MUC3A	C	CTGAGACCG	frameshift_variant	HIGH
PD50649	100957776	MUC3A	A	AT	frameshift_variant	HIGH
PD50649	100957784	MUC3A	C	CA	frameshift_variant	HIGH
PD50649	100957788	MUC3A	C	CCGAGTCCACATCCA	frameshift_variant	HIGH
PD50649	100958132	MUC3A	G	GAAGAC	frameshift_variant	HIGH
PD50649	100958143	MUC3A	A	ACTTCTTTGATAGCCAC	frameshift_variant	HIGH
PD50649	100958184	MUC3A	A	AC	frameshift_variant	HIGH
PD50649	100958188	MUC3A	C	CT	frameshift_variant	HIGH
PD50649	100958198	MUC3A	TAC	T	frameshift_variant	HIGH
PD50650	100953774	MUC3A	AT	A	frameshift_variant	HIGH
PD50650	100953777	MUC3A	C	CA	frameshift_variant	HIGH
PD50650	100958586	MUC3A	A	ACTCCCAGCTTCACTTACCACAGAGTACTCCCACCTTCACTTCTTTGATCACTACCAGCGAGATGACATCCCACAGTACTCCCAGCTTGACTTTTTCAATCACCACCACCAAAAGCACATCCTAAGTC	stop_gained	HIGH
PD50650	100958619	MUC3A	G	GAGACCCCCTCACACAGTACTACACCAAGACTACCTCACACCTTCCTCCCAGGTTCACTTGTTTGATCACCACCACCA	stop_gained	HIGH
PD50650	100958932	MUC3A	C	TATCACATAGTACTCCCAGCTTGA	frameshift_variant	HIGH
PD50650	100958935	MUC3A	T	TCTGCAATCACCAATACTGAGACCATGTC	stop_gained	HIGH
PD50650	100958949	MUC3A	G	GCCAGCACCGA	frameshift_variant	HIGH
PD50650	100958950	MUC3A	G	GACCA	frameshift_variant	HIGH
PD50650	100958956	MUC3A	C	CACAGTACTCCCAGTTTGAA	frameshift_variant	HIGH
PD50650	100958959	MUC3A	C	CTCTGATCACCACCACCGGGACCAGCTCACACA	frameshift_variant	HIGH
PD50650	100958961	MUC3A	T	TACCCTCGGCTTAA	frameshift_variant	HIGH
PD50653	100958302	MUC3A	G	GCTTAA	frameshift_variant	HIGH
PD50653	100958303	MUC3A	G	GTTCT	frameshift_variant	HIGH
PD50653	100958352	MUC3A	A	ATCACCACCACTGGCAACACCTCACACAGT	frameshift_variant	HIGH
PD50653	100958354	MUC3A	T	TGCCGAGCTTC	frameshift_variant	HIGH
PD50653	100958356	MUC3A	C	CTTCTT	frameshift_variant	HIGH
PD50653	100958357	MUC3A	C	CG	frameshift_variant	HIGH
PD50653	100958806	MUC3A	C	CCAAGACCACCTCAACCAGTCCTCCCAGATTCACCT	frameshift_variant	HIGH
PD50653	100958808	MUC3A	T	TG	frameshift_variant	HIGH
PD50653	100958856	MUC3A	TC	T	frameshift_variant	HIGH
PD50653	100958858	MUC3A	T	TGG	frameshift_variant	HIGH
PD50653	100958859	MUC3A	TC	T	frameshift_variant	HIGH
PD51372	100953731	MUC3A	CTC	T	frameshift_variant	HIGH
PD51372	100953737	MUC3A	T	TGG	frameshift_variant	HIGH
PD51372	100953808	MUC3A	A	AG	frameshift_variant	HIGH
PD51372	100953812	MUC3A	AG	A	frameshift_variant	HIGH
PD51372	100953834	MUC3A	A	AT	frameshift_variant	HIGH
PD51372	100953838	MUC3A	TC	T	frameshift_variant	HIGH

Appendix 4

Table A4.1: Clustering of locations of *MUC3A* variants detected using the bcbio-nextgen (Chapman et al. 2021) pipeline. Start and End Positions correspond to reference genome GRCh38.

Cluster Number	Patient	Start Position	End Position	Cluster Length (bp)
1	PD39456	100953731	100953734	109
	PD39457	100953731	100953734	
	PD44703	100953731	100953734	
	PD39456	100953737	100953738	
	PD39457	100953737	100953738	
	PD44703	100953737	100953738	
	PD39456	100953752	100953753	
	PD39456	100953758	100953760	
	PD39456	100953774	100953776	
	PD39456	100953777	100953778	
	PD39457	100953808	100953809	
	PD44703	100953808	100953809	
	PD39457	100953812	100953814	
	PD44703	100953812	100953814	
	PD39457	100953834	100953835	
	PD44703	100953834	100953835	
	PD39457	100953838	100953840	
	PD44703	100953838	100953840	
2	PD44696	100954996	100954998	253
	PD44697	100954996	100954998	
	PD44700	100954996	100954998	
	PD44701	100954996	100954998	
	PD44696	100955002	100955003	
	PD44697	100955009	100955010	
	PD44701	100955009	100955010	
	PD44702	100955173	100955175	
	PD44702	100955175	100955178	
	PD39452	100955243	100955246	
	PD39456	100955243	100955246	
	PD44701	100955243	100955246	
	PD44704	100955243	100955246	
	PD44702	100955244	100955246	
	PD44702	100955246	100955249	
	PD39456	100955247	100955249	
	PD44701	100955247	100955249	
	PD44704	100955247	100955249	
3	PD44695	100957642	100957643	502
	PD44696	100957770	100957771	

	PD44696	100957773	100957774	
	PD39448	100957775	100957776	
	PD39451	100957775	100957776	
	PD44700	100957775	100957776	
	PD39448	100957776	100957777	
	PD39451	100957776	100957777	
	PD44700	100957776	100957777	
	PD39448	100957779	100957781	
	PD39451	100957779	100957781	
	PD44700	100957779	100957781	
	PD39448	100957784	100957786	
	PD39451	100957784	100957786	
	PD44700	100957784	100957786	
	PD39448	100957787	100957789	
	PD39451	100957787	100957789	
	PD44700	100957787	100957789	
	PD39456	100957834	100957837	
	PD39456	100957842	100957843	
	PD39456	100957843	100957844	
	PD44703	100957884	100957890	
	PD44691	100957885	100957912	
	PD44703	100957894	100957902	
	PD44703	100957909	100957912	
	PD44691	100957913	100957933	
	PD44703	100957913	100957931	
	PD44703	100957935	100957941	
	PD39455	100957959	100957961	
	PD39455	100957963	100957964	
	PD44691	100957978	100957980	
	PD44691	100957982	100957987	
	PD44691	100957989	100958016	
	PD39458	100958011	100958012	
	PD39458	100958015	100958016	
	PD39458	100958016	100958017	
	PD39458	100958018	100958019	
	PD44691	100958018	100958023	
	PD44691	100958026	100958029	
	PD44691	100958042	100958045	
	PD44702	100958132	100958144	
	PD44702	100958143	100958144	

	PD39460	100958183	100958194	
	PD44691	100958183	100958194	
4	PD44699	100958183	100958194	475
	PD39448	100958184	100958185	
	PD39455	100958184	100958185	

	PD39458	100958184	100958185	
	PD44693	100958184	100958185	
	PD39448	100958186	100958187	
	PD39448	100958187	100958188	
	PD39455	100958188	100958189	
	PD39458	100958188	100958189	
	PD44693	100958188	100958189	
	PD39455	100958190	100958191	
	PD39458	100958190	100958191	
	PD39455	100958192	100958193	
	PD39458	100958192	100958193	
	PD39455	100958195	100958196	
	PD39458	100958195	100958196	
	PD39460	100958196	100958202	
	PD44691	100958196	100958202	
	PD44699	100958196	100958202	
	PD39455	100958197	100958198	
	PD39458	100958197	100958198	
	PD44693	100958198	100958201	
	PD39460	100958204	100958221	
	PD44691	100958204	100958221	
	PD44699	100958204	100958221	
	PD39448	100958216	100958217	
	PD44704	100958216	100958217	
	PD44704	100958216	100958217	
	PD39448	100958219	100958220	
	PD44704	100958219	100958220	
	PD44704	100958219	100958220	
	PD39448	100958224	100958225	
	PD39460	100958224	100958235	
	PD44691	100958224	100958235	
	PD44699	100958224	100958235	
	PD44704	100958224	100958225	
	PD39448	100958230	100958231	
	PD44704	100958230	100958231	
	PD39448	100958234	100958235	
	PD44704	100958234	100958235	
	PD39460	100958235	100958237	
	PD44691	100958235	100958237	
	PD44699	100958235	100958237	
	PD39460	100958239	100958241	
	PD44691	100958239	100958241	
	PD44699	100958239	100958241	
	PD39460	100958249	100958252	
	PD44691	100958249	100958252	
	PD44699	100958249	100958252	
	PD44696	100958298	100958299	

	PD44696	100958300	100958301	
	PD39455	100958302	100958303	
	PD39455	100958303	100958304	
	PD44696	100958304	100958305	
	PD39455	100958352	100958353	
	PD39455	100958354	100958355	
	PD39455	100958356	100958357	
	PD39455	100958356	100958357	
	PD39455	100958357	100958358	
	PD39455	100958357	100958358	
	PD44692	100958586	100958587	
	PD39457	100958619	100958620	
	PD44704	100958619	100958620	
	PD39457	100958621	100958622	
	PD44704	100958621	100958622	
	PD39457	100958623	100958624	
	PD44704	100958623	100958624	
	PD44691	100958624	100958625	
	PD39457	100958626	100958627	
	PD44704	100958626	100958627	
	PD44691	100958629	100958630	
	PD39457	100958646	100958647	
	PD39457	100958647	100958648	
	PD39453	100958653	100958654	
	PD39453	100958654	100958655	
	PD39453	100958656	100958657	
	PD39453	100958657	100958658	

	PD39457	100958763	100958764	
	PD39448	100958806	100958807	
	PD39448	100958808	100958809	
	PD39448	100958856	100958858	
	PD39448	100958858	100958859	
	PD39448	100958859	100958861	
5	PD44693	100958904	100958905	237
	PD44693	100958905	100958906	
	PD44693	100958907	100958908	
	PD44693	100958944	100958946	
	PD44693	100958948	100958949	
	PD44693	100958963	100958964	
	PD39454	100958995	100958997	
	PD39454	100958999	100959000	

Appendix 5

Table A5.1: All HIGH impact severity *MUC3A* mutations detected through bioinformatics analysis of WGS data using Mutect2 Variant caller and the PON approach with bcbio-nextgen (Chapman et al. 2021) pipeline. Start positions correspond to reference genome GRCh38 and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	start	dbSNP Reference	ref	alt	Impact	Impact Severity
PD39445	100952716	rs773095268;	C	CGT	frameshift	HIGH
PD39445	100952718	.	CCT	C	frameshift	HIGH
PD39445	100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH
PD39445	100954987	.	AC	A	frameshift	HIGH
PD39445	100954991	.	C	CGGCG	frameshift	HIGH
PD39445	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD39445	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD39445	100955296	.	GC	G	frameshift	HIGH
PD39445	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD39445	100956808	.	ACC	A	frameshift	HIGH
PD39445	100956813	.	G	GAA	frameshift	HIGH
PD39446	100955537	.	G	GTA	frameshift	HIGH
PD39446	100955539	.	ATG	A	frameshift	HIGH
PD39446	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD39446	100955571	.	A	ATG	frameshift	HIGH
PD39446	100956690	.	TG	T	frameshift	HIGH
PD39446	100956694	.	G	GA	frameshift	HIGH
PD39447	100954960	rs1584801081;	AAC	GAC,A	frameshift	HIGH
PD39447	100954965	.	C	CTT	frameshift	HIGH
PD39447	100954987	.	AC	A	frameshift	HIGH
PD39447	100954991	.	C	CGGCG	frameshift	HIGH
PD39447	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD39447	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD39447	100955296	.	GC	G	frameshift	HIGH
PD39447	100955302	rs1419100339;	C	CA,A	frameshift	HIGH
PD39447	100955537	rs1305667129;	G	A,GTA	frameshift	HIGH
PD39447	100955539	.	ATG	A	frameshift	HIGH
PD39447	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD39447	100955565	.	G	GAA	frameshift	HIGH

PD39447	100955585	.	CTG	C	frameshift	HIGH
PD39447	100955595	.	TC	T	frameshift	HIGH
PD39447	100955606	.	C	CAG	frameshift	HIGH
PD39447	100955607	.	CCT	C	frameshift	HIGH
PD39447	100955774	.	CA	C	frameshift	HIGH
PD39447	100956690	.	TG	T	frameshift	HIGH
PD39448	100953281	rs1229668308;	G	GTA,A	frameshift	HIGH
PD39448	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD39448	100953342	.	T	TTG	frameshift	HIGH
PD39448	100953350	.	C	CAG	frameshift	HIGH
PD39448	100953351	.	CCT	C	frameshift	HIGH
PD39448	100953359	rs986103920;	C	CAA,T	frameshift	HIGH
PD39448	100953361	.	ACC	A	frameshift	HIGH
PD39448	100955010	.	C	CG	frameshift	HIGH
PD39448	100955296	.	GC	G	frameshift	HIGH
PD39448	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD39448	100955606	.	C	CAG	frameshift	HIGH
PD39448	100955607	.	CCT	C	frameshift	HIGH
PD39448	100956694	.	G	GA	frameshift	HIGH
PD39449	100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH
PD39449	100953283	.	ATG	A	frameshift	HIGH
PD39449	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD39449	100953342	.	T	TTG	frameshift	HIGH
PD39449	100953350	.	C	CAG	frameshift	HIGH
PD39449	100953351	.	CCT	C	frameshift	HIGH
PD39449	100953359	rs986103920;	C	CAA,T	frameshift	HIGH
PD39449	100953361	.	ACC	A	frameshift	HIGH
PD39449	100953775	.	AT	A	frameshift	HIGH
PD39449	100953778	.	CC	AT,AC,CAC	frameshift	HIGH
PD39449	100954987	.	AC	A	frameshift	HIGH
PD39449	100954991	.	C	CGGCG	frameshift	HIGH
PD39449	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD39449	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD39449	100955296	.	GC	G	frameshift	HIGH
PD39449	100955302	rs1584801296;	C	CA	frameshift	HIGH

PD39449	100956808	.	ACC	A	frameshift	HIGH
PD39449	100956813	.	G	GAA	frameshift	HIGH
PD39450	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD39450	100952650	rs765973058;	AC	A	frameshift	HIGH
PD39450	100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH
PD39450	100955010	.	C	CG	frameshift	HIGH
PD39450	100955296	.	GC	G	frameshift	HIGH
PD39450	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD39450	100955595	.	TC	T	frameshift	HIGH
PD39450	100955606	.	C	CAG	frameshift	HIGH
PD39450	100955607	.	CCT	C	frameshift	HIGH
PD39450	100955615	.	C	CAA	frameshift	HIGH
PD39450	100955617	.	AC	A	frameshift	HIGH
PD39450	100956690	.	TG	T	frameshift	HIGH
PD39450	100956694	.	G	GA	frameshift	HIGH
PD39450	100956808	.	ACC	A	frameshift	HIGH
PD39450	100956813	.	G	GAA	frameshift	HIGH
PD39450	100957507	.	TCA	T	frameshift	HIGH
PD39450	100957511	.	T	C,TGATA	frameshift	HIGH
PD39450	100957512	.	C	CG	frameshift	HIGH
PD39451	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD39451	100952650	rs765973058;	AC	A	frameshift	HIGH
PD39451	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD39451	100955126	.	CTA	C	frameshift	HIGH
PD39451	100955133	.	C	CTT	frameshift	HIGH
PD39451	100955149	.	T	TTG	frameshift	HIGH
PD39451	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD39452	100953281	rs1229668308;	G	GTA,A	frameshift	HIGH
PD39452	100953283	.	ATG	A	frameshift	HIGH
PD39452	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD39452	100953350	.	C	CAG	frameshift	HIGH
PD39452	100953351	.	CCT	C	frameshift	HIGH
PD39452	100953359	rs986103920;	C	CAA,T	frameshift	HIGH
PD39452	100953361	.	ACC	A	frameshift	HIGH
PD39452	100953730	.	AC	A	frameshift	HIGH

PD39452	100953733	.	TC	T	frameshift	HIGH
PD39452	100953738	.	T	TGG	frameshift	HIGH
PD39452	100953752	.	AT	A	frameshift	HIGH
PD39452	100953760	.	A	AT	frameshift	HIGH
PD39452	100953949	.	AC	A	frameshift	HIGH
PD39452	100953954	.	A	AG	frameshift	HIGH
PD39452	100954960	rs1584801081;	AAC	GAC,A	frameshift	HIGH
PD39452	100954965	.	C	CTT	frameshift	HIGH
PD39452	100954983	.	GGTGA	G	frameshift	HIGH
PD39452	100954987	.	AC	A	frameshift	HIGH
PD39452	100954989	rs1390720829;	T	C,TTGCC	frameshift	HIGH
PD39452	100954993	.	CTCCTCACTAC CAT	C	frameshift	HIGH
PD39452	100954996	.	CT	C	frameshift	HIGH
PD39452	100955004	.	CAT	C	frameshift	HIGH
PD39452	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD39452	100955126	.	CTA	C	frameshift	HIGH
PD39452	100955133	.	C	CTT	frameshift	HIGH
PD39452	100955149	.	T	TTG	frameshift	HIGH
PD39452	100955153	.	TCC	AGC,T	frameshift	HIGH
PD39452	100955245	.	TC	T	frameshift	HIGH
PD39452	100955247	.	TTC	T	frameshift	HIGH
PD39452	100955296	.	GC	G	frameshift	HIGH
PD39452	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD39452	100955539	.	ATG	A	frameshift	HIGH
PD39452	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD39452	100955565	.	G	GAA	frameshift	HIGH
PD39452	100955585	.	CTG	C	frameshift	HIGH
PD39452	100955595	.	TC	T	frameshift	HIGH
PD39452	100955606	.	C	CAG	frameshift	HIGH
PD39452	100955607	.	CCT	C	frameshift	HIGH
PD39452	100955615	.	C	CAA	frameshift	HIGH
PD39452	100955617	.	ACC	A	frameshift	HIGH
PD39452	100956690	.	TG	T	frameshift	HIGH
PD39452	100956694	.	G	GA	frameshift	HIGH
PD39453	100953155	.	CCGGT	C	frameshift	HIGH

PD39453	100953160	.	T	TGAA	frameshift	HIGH
PD39453	100953163	.	C	CT	frameshift	HIGH
PD39453	100954987	.	AC	A	frameshift	HIGH
PD39453	100954991	.	C	CGGCG	frameshift	HIGH
PD39453	100954993	.	CTCCTCACTAC CAT	C	frameshift	HIGH
PD39453	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD39453	100955571	.	A	ATG,G	frameshift	HIGH
PD39453	100955606	.	C	CAG	frameshift	HIGH
PD39453	100955607	.	CCT	C	frameshift	HIGH
PD39454	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD39454	100952650	rs765973058;	AC	A	frameshift	HIGH
PD39454	100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH
PD39454	100955247	.	TTC	T	frameshift	HIGH
PD39454	100955774	.	CA	C	frameshift	HIGH
PD39454	100955779	rs1313584776;	C	CA,A	frameshift	HIGH
PD39454	100956694	.	G	GA	frameshift	HIGH
PD39455	100954991	.	C	CGGCG	frameshift	HIGH
PD39455	100955010	.	C	CG	frameshift	HIGH
PD39455	100955126	.	CTA	C	frameshift	HIGH
PD39455	100955133	.	C	CTT	frameshift	HIGH
PD39455	100955149	.	T	TTG	frameshift	HIGH
PD39455	100955153	.	TCC	T	frameshift	HIGH
PD39455	100955247	.	TTC	T	frameshift	HIGH
PD39455	100955606	.	C	CAG	frameshift	HIGH
PD39455	100955607	.	CCT	C	frameshift	HIGH
PD39455	100955774	.	CA	C	frameshift	HIGH
PD39455	100955779	rs1313584776;	C	CA,A	frameshift	HIGH
PD39455	100956694	.	G	GA	frameshift	HIGH
PD39456	100955004	.	CAT	C	frameshift	HIGH
PD39456	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD39456	100955169	.	AG	A	frameshift	HIGH
PD39456	100955171	.	CCA	C	frameshift	HIGH
PD39457	100954998	.	CACTA	C	frameshift	HIGH
PD39457	100955004	.	CAT	C	frameshift	HIGH
PD39457	100955010	.	C	CG	frameshift	HIGH

PD39457	100955539	.	ATG	A	frameshift	HIGH
PD39457	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD39457	100955565	.	G	GAA	frameshift	HIGH
PD39457	100955585	.	CTG	C	frameshift	HIGH
PD39457	100955595	.	TC	T	frameshift	HIGH
PD39457	100956808	.	ACC	A	frameshift	HIGH
PD39457	100956813	.	G	GAA	frameshift	HIGH
PD39458	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD39458	100952650	rs765973058;	AC	A	frameshift	HIGH
PD39458	100955004	.	CAT	C	frameshift	HIGH
PD39458	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD39458	100955537	.	G	GTA	frameshift	HIGH
PD39458	100955539	.	ATG	A	frameshift	HIGH
PD39458	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD39458	100955565	.	G	GAA	frameshift	HIGH
PD39458	100955585	.	CTG	C	frameshift	HIGH
PD39458	100955595	.	TC	T	frameshift	HIGH
PD39458	100955606	.	C	CAG	frameshift	HIGH
PD39458	100955607	.	CCT	C	frameshift	HIGH
PD39458	100956690	.	TG	T	frameshift	HIGH
PD39458	100956694	.	G	GA	frameshift	HIGH
PD39459	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD39459	100952650	rs765973058;	AC	A	frameshift	HIGH
PD39459	100952833	rs1584799673;	GCC	G,ACC	frameshift	HIGH
PD39459	100955126	.	CTA	C	frameshift	HIGH
PD39459	100955133	.	C	CTT	frameshift	HIGH
PD39459	100955149	.	T	TTG	frameshift	HIGH
PD39459	100955153	.	TCC	AGC,T	frameshift	HIGH
PD39459	100955606	.	C	CAG	frameshift	HIGH
PD39459	100955607	.	CCT	C	frameshift	HIGH
PD39459	100955615	.	C	CAA	frameshift	HIGH
PD39459	100955617	.	ACC	A	frameshift	HIGH
PD39459	100956690	.	TG	T	frameshift	HIGH
PD39459	100956694	.	G	GA	frameshift	HIGH
PD39460	100953281	rs1229668308;	G	GTA,A	frameshift	HIGH
PD39460	100953283	.	ATG	A	frameshift	HIGH

PD39460	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD39460	100953350	.	C	CAG	frameshift	HIGH
PD39460	100953351	.	CCT	C	frameshift	HIGH
PD39460	100953949	.	AC	A	frameshift	HIGH
PD39460	100953954	.	A	AG	frameshift	HIGH
PD39460	100954987	.	AC	A	frameshift	HIGH
PD39460	100954991	.	C	CGGCG	frameshift	HIGH
PD39460	100954993	.	CTCCTCACTAC CAT	C	frameshift	HIGH
PD39460	100955010	.	C	CG	frameshift	HIGH
PD39460	100955169	.	AG	A	frameshift	HIGH
PD39460	100955171	.	CCA	C	frameshift	HIGH
PD39460	100956690	.	TG	T	frameshift	HIGH
PD39460	100956694	.	G	GA	frameshift	HIGH
PD44691	100954987	.	AC	A	frameshift	HIGH
PD44691	100954991	.	C	CGGCG	frameshift	HIGH
PD44691	100955010	.	C	CG	frameshift	HIGH
PD44691	100955537	rs1305667129;	G	A,GTA	frameshift	HIGH
PD44691	100955539	.	ATG	A	frameshift	HIGH
PD44691	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD44691	100955571	.	A	ATG	frameshift	HIGH
PD44691	100955586	.	TG	AC,TACCCCG	frameshift	HIGH
PD44691	100963159	.	G	T	frameshift	HIGH
PD44691	100963171	.	G	GTC	frameshift	HIGH
PD44692	100955296	.	GC	G	frameshift	HIGH
PD44692	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD44692	100955585	.	CTG	C	frameshift	HIGH
PD44692	100955595	.	TC	T	frameshift	HIGH
PD44692	100955606	.	C	CAG	frameshift	HIGH
PD44692	100955607	.	CCT	C	frameshift	HIGH
PD44692	100955774	.	CA	C	frameshift	HIGH
PD44692	100956690	.	TG	T	frameshift	HIGH
PD44692	100956694	.	G	GA	frameshift	HIGH
PD44692	100957605	.	TTCTTCAATCA CCAATACCAA GACCACCTCA CACAGCTCTC CCAGCTTCACT TCTTCGATCAC	T	frameshift	HIGH

			CACCACCGAGACCACATCCCACAATACTCCCAGCC			
PD44693	100956690	.	TG	T	frameshift	HIGH
PD44693	100956694	.	G	GA	frameshift	HIGH
PD44694	100956690	.	TG	T	frameshift	HIGH
PD44694	100956694	.	G	GA	frameshift	HIGH
PD44694	100956808	.	ACC	A	frameshift	HIGH
PD44694	100956813	.	G	GAA	frameshift	HIGH
PD44695	100953221	.	C	CTG	frameshift	HIGH
PD44695	100953223	.	ACG	A	frameshift	HIGH
PD44695	100954987	.	AC	A	frameshift	HIGH
PD44695	100954991	.	C	CGGCG	frameshift	HIGH
PD44695	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD44695	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD44695	100955606	.	C	CAG	frameshift	HIGH
PD44695	100955607	.	CCT	C	frameshift	HIGH
PD44695	100956690	.	TG	T	frameshift	HIGH
PD44695	100956694	.	G	GA	frameshift	HIGH
PD44696	100953281	.	G	GTA	frameshift	HIGH
PD44696	100953283	.	ATG	A	frameshift	HIGH
PD44696	100954987	.	AC	A	frameshift	HIGH
PD44696	100954991	.	C	CGGCG	frameshift	HIGH
PD44696	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD44696	100955008	rs1273949401;	A	G,AT	frameshift	HIGH
PD44696	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD44696	100955571	.	A	ATG,G	frameshift	HIGH
PD44696	100955585	.	CTG	C	frameshift	HIGH
PD44696	100955595	.	TC	T	frameshift	HIGH
PD44696	100955774	.	CA	C	frameshift	HIGH
PD44696	100955779	rs1313584776;	C	CA,G,A	frameshift	HIGH
PD44696	100956690	.	TG	T	frameshift	HIGH
PD44696	100956694	.	G	GA	frameshift	HIGH
PD44696	100956808	.	ACC	A	frameshift	HIGH
PD44696	100956813	.	G	GAA	frameshift	HIGH

PD44697	100953281	rs1229668308;	G	GTA,A	frameshift	HIGH
PD44697	100953283	.	ATG	A	frameshift	HIGH
PD44697	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD44697	100954991	.	C	CGGCG	frameshift	HIGH
PD44697	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD44697	100955585	.	CTG	C	frameshift	HIGH
PD44697	100955595	.	TC	T	frameshift	HIGH
PD44697	100955606	.	C	CAG	frameshift	HIGH
PD44697	100955607	.	CCT	C	frameshift	HIGH
PD44697	100956690	.	TG	T	frameshift	HIGH
PD44697	100956694	.	G	GA	frameshift	HIGH
PD44697	100956808	.	ACC	A	frameshift	HIGH
PD44697	100956813	.	G	GAA	frameshift	HIGH
PD44698	100953155	.	CCGGT	C	frameshift	HIGH
PD44698	100953160	.	T	TGAA	frameshift	HIGH
PD44698	100954993	.	CTCCTCACTAC CAT	C	frameshift	HIGH
PD44698	100955774	.	CA	C	frameshift	HIGH
PD44698	100955779	rs1313584776;	C	CA,A	frameshift	HIGH
PD44698	100956690	.	TG	T	frameshift	HIGH
PD44698	100956694	.	G	GA	frameshift	HIGH
PD44698	100956808	.	ACC	A	frameshift	HIGH
PD44698	100956813	.	G	GAA	frameshift	HIGH
PD44699	100954987	.	AC	A	frameshift	HIGH
PD44699	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD44699	100956808	.	ACC	A	frameshift	HIGH
PD44699	100956813	.	G	GAA	frameshift	HIGH
PD44700	100954993	.	CTCCTCACTAC CAT	C	frameshift	HIGH
PD44700	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD44700	100955174	.	CG	C	frameshift	HIGH
PD44700	100955176	rs1584801207;	ATG	A	frameshift	HIGH
PD44700	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD44700	100955585	.	CTG	C	frameshift	HIGH
PD44700	100955595	.	TC	T	frameshift	HIGH
PD44700	100955606	.	C	CAG	frameshift	HIGH

PD44700	100955607	.	CCT	C	frameshift	HIGH
PD44700	100955916	rs1584801772;	AG	A	frameshift	HIGH
PD44700	100955920	rs1584801776;	A	AG	frameshift	HIGH
PD44700	100955959	.	ACC	A	frameshift	HIGH
PD44700	100956690	.	TG	T	frameshift	HIGH
PD44700	100956694	.	G	GA	frameshift	HIGH
PD44701	100952716	rs773095268;	C	CGT	frameshift	HIGH
PD44701	100952718	.	CCT	C,ACT	frameshift	HIGH
PD44701	100954960	rs1584801081;	AAC	A	frameshift	HIGH
PD44701	100954965	.	C	CTT	frameshift	HIGH
PD44701	100954991	.	C	CGGCG	frameshift	HIGH
PD44701	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD44701	100955010	rs1344850949;	C	T,CG	frameshift	HIGH
PD44701	100955916	rs1584801772;	AG	A	frameshift	HIGH
PD44701	100955920	rs1584801776;	A	AG	frameshift	HIGH
PD44701	100955959	.	ACC	A	frameshift	HIGH
PD44701	100956690	.	TG	T	frameshift	HIGH
PD44701	100956694	.	G	GA	frameshift	HIGH
PD44701	100963159	.	G	T	frameshift	HIGH
PD44701	100963171	.	G	GTC	frameshift	HIGH
PD44701	100963173	.	C	CCGTATCATTAA	frameshift	HIGH
PD44701	100963179	.	AGTGTCTGTGG	A	frameshift	HIGH
PD44702	100952647	rs773185111;	AGG	A	frameshift	HIGH
PD44702	100952650	rs765973058;	AC	A	frameshift	HIGH
PD44702	100953312	rs1196620613;	ACC	A	frameshift	HIGH
PD44702	100953315	.	A	ATG,G	frameshift	HIGH
PD44702	100953335	rs1237264217;	G	GAA,A	frameshift	HIGH
PD44702	100953359	rs986103920;	C	CAA,T	frameshift	HIGH
PD44702	100955010	rs1344850949;	C	CG,T	frameshift	HIGH
PD44702	100955126	.	CTA	C	frameshift	HIGH
PD44702	100955133	.	C	CTT	frameshift	HIGH
PD44702	100955149	.	T	TTG	frameshift	HIGH
PD44702	100955153	.	TCC	T,AGC	frameshift	HIGH
PD44702	100955173	.	ACG	A	frameshift	HIGH

PD44702	100955296	.	GC	G	frameshift	HIGH
PD44702	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD44702	100956690	.	TG	T	frameshift	HIGH
PD44702	100956691	.	G	GT	frameshift	HIGH
PD44702	100956693	.	AG	GA,A	frameshift	HIGH
PD44702	100956694	.	G	GA	frameshift	HIGH
PD44702	100956808	.	ACC	A	frameshift	HIGH
PD44702	100956813	.	G	GAA	frameshift	HIGH
PD44702	100963159	.	G	T	frameshift	HIGH
PD44702	100963171	.	G	GTC	frameshift	HIGH
PD44702	100963173	.	C	CCGTATCATTAA	frameshift	HIGH
PD44702	100963179	.	AGTGTCTGTGG	A	frameshift	HIGH
PD44703	100953949	.	AC	A	frameshift	HIGH
PD44703	100953954	.	A	AG	frameshift	HIGH
PD44703	100954987	.	AC	A	frameshift	HIGH
PD44703	100954991	.	C	CGGCG	frameshift	HIGH
PD44703	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD44703	100955010	.	C	CG	frameshift	HIGH
PD44703	100955126	.	CTA	C	frameshift	HIGH
PD44703	100955133	.	C	CTT	frameshift	HIGH
PD44703	100955149	.	T	TTG	frameshift	HIGH
PD44703	100955153	.	TCC	T	frameshift	HIGH
PD44703	100955537	.	G	GTA	frameshift	HIGH
PD44703	100955539	.	ATG	A	frameshift	HIGH
PD44703	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD44703	100955571	.	A	ATG	frameshift	HIGH
PD44703	100956691	.	G	GT	frameshift	HIGH
PD44703	100956693	.	AG	A	frameshift	HIGH
PD44703	100956694	.	G	GA	frameshift	HIGH
PD44703	100956808	.	ACC	A	frameshift	HIGH
PD44703	100956813	.	G	GAA	frameshift	HIGH
PD44703	100957504	.	TC	T	frameshift	HIGH
PD44703	100957507	.	TC	T	frameshift	HIGH
PD44703	100957511	.	T	TGATA	frameshift	HIGH

PD44703	100957512	.	C	CG	frameshift	HIGH
PD44704	100954987	.	ACTACAC	ATACAC,A	frameshift	HIGH
PD44704	100955126	.	CTA	C	frameshift	HIGH
PD44704	100955133	.	C	CTT	frameshift	HIGH
PD44704	100955149	.	T	TTG	frameshift	HIGH
PD44704	100955153	.	TCC	T	frameshift	HIGH
PD44704	100955245	.	TC	T	frameshift	HIGH
PD44704	100955247	.	TTC	T	frameshift	HIGH
PD44704	100955585	.	CTG	C	frameshift	HIGH
PD44704	100955595	.	TC	CC,T	frameshift	HIGH
PD44704	100955606	.	C	A,CAG	frameshift	HIGH
PD44704	100955607	.	CCT	C	frameshift	HIGH
PD44704	100956690	.	TG	T	frameshift	HIGH
PD44704	100956694	.	G	GA	frameshift	HIGH
PD44704	100963159	.	G	T	frameshift	HIGH
PD44704	100963171	.	G	GTC	frameshift	HIGH
PD44704	100963179	.	AGTGTCTGTGG	A	frameshift	HIGH
PD50649	100955920	rs1584801776;	A	AG	frameshift	HIGH
PD50649	100956694	.	G	GA	frameshift	HIGH
PD50650	100954991	.	C	CGGCG	frameshift	HIGH
PD50650	100954993	.	CTCCTCACTACCAT	C	frameshift	HIGH
PD50650	100955537	rs1305667129;	G	GTA,A	frameshift	HIGH
PD50650	100955539	.	ATG	A	frameshift	HIGH
PD50650	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD50650	100955571	.	A	ATG	frameshift	HIGH
PD50651	100953949	.	AC	A	frameshift	HIGH
PD50651	100953954	.	A	AG	frameshift	HIGH
PD50651	100955302	rs1584801296;	C	CA	frameshift	HIGH
PD50651	100955539	.	ATG	A	frameshift	HIGH
PD50651	100955560	rs1584801446;	AGT	A	frameshift	HIGH
PD50651	100955565	.	G	GAA	frameshift	HIGH
PD50651	100955585	.	CTG	C	frameshift	HIGH
PD50651	100955595	.	TC	T	frameshift	HIGH
PD50651	100955606	.	C	CAG	frameshift	HIGH

PD50651	100955607	.	CCT	C	frameshift	HIGH
PD50653	100953283	.	ATG	A	frameshift	HIGH
PD50653	100953330	rs1237518715;	ACT	A	frameshift	HIGH
PD50653	100953350	.	C	CAG	frameshift	HIGH
PD50653	100953351	.	CCT	C	frameshift	HIGH
PD50653	100953359	rs986103920;	C	CAA,T	frameshift	HIGH
PD50653	100953361	.	ACC	A	frameshift	HIGH
PD50653	100954987	.	AC	A	frameshift	HIGH
PD50653	100955010	.	C	CG	frameshift	HIGH
PD50653	100956690	.	TG	T	frameshift	HIGH
PD50653	100956694	.	G	GA	frameshift	HIGH
PD50653	100956808	.	ACC	A	frameshift	HIGH
PD50653	100956813	.	G	GAA	frameshift	HIGH
PD51372	100953221	.	C	CTG	frameshift	HIGH
PD51372	100953223	.	ACG	A	frameshift	HIGH
PD51372	100955010	.	C	CG	frameshift	HIGH
PD51372	100955606	.	C	CAG	frameshift	HIGH
PD51372	100955607	.	CCT	C	frameshift	HIGH
PD51372	100955774	.	CA	C	frameshift	HIGH
PD51372	100955779	.	C	CA	frameshift	HIGH
PD51372	100956808	.	ACC	A	frameshift	HIGH
PD51372	100956813	.	G	GAA	frameshift	HIGH
PD51372	100963159	.	G	T	frameshift	HIGH
PD51372	100963171	.	G	GTC	frameshift	HIGH
PD51372	100963173	.	C	CCGTATCATTAA	frameshift	HIGH
PD51372	100963179	.	AGTGTCTGTGG	A	frameshift	HIGH
PD51372	100973055	rs796345412;	G	A	frameshift	HIGH

Appendix 6

Table A6.1: All medium and high impact *MUC3A* variants detected though variant calling of RNA-seq data from 15 patients. Start Positions correspond to reference genome GRCh38 and Ref and Alt refer to the reference and alternate alleles respectively.

Patient	Start Position	Gene	Ref	Alt	Impact	Impact Severity
PR50551	100952327	*MUC3A*	C	T	missense_variant	MED
PR44703	100952756	*MUC3A*	G	C	missense_variant	MED
PR44703	100952788	*MUC3A*	A	C	missense_variant	MED
PR44703	100952792	*MUC3A*	C	T	missense_variant	MED
PR44703	100952797	*MUC3A*	C	T	missense_variant	MED
PR44703	100954123	*MUC3A*	G	A	missense_variant	MED
PR44703	100954136	*MUC3A*	C	G	missense_variant	MED
PR44703	100955384	*MUC3A*	C	G	missense_variant	MED
PR44703	100955486	*MUC3A*	T	A	missense_variant	MED
PR44703	100955962	*MUC3A*	G	A	missense_variant	MED
PR44703	100955968	*MUC3A*	A	G	missense_variant	MED
PR44703	100955981	*MUC3A*	G	C	missense_variant	MED
PR44703	100956223	*MUC3A*	C	T	missense_variant	MED
PR44703	100956249	*MUC3A*	TAC	CAG	missense_variant	MED
PR44703	100956404	*MUC3A*	C	A	missense_variant	MED
PR44703	100956443	*MUC3A*	C	A	missense_variant	MED
PR50554	100957015	*MUC3A*	T	C	missense_variant	MED
PR44697	100957015	*MUC3A*	T	C	missense_variant	MED
PR44705	100957568	*MUC3A*	CT	GC	missense_variant	MED
PR44705	100957573	*MUC3A*	G	A	missense_variant	MED
PR50554	100957768	*MUC3A*	A	C	missense_variant	MED
PR50554	100957777	*MUC3A*	ACC	GAG	missense_variant	MED
PR50552	100957857	*MUC3A*	G	C	missense_variant	MED
PR44705	100958000	*MUC3A*	T	C	missense_variant	MED
PR50551	100958051	*MUC3A*	T	C	missense_variant	MED
PR50551	100958135	*MUC3A*	TG	CC	missense_variant	MED
PR50552	100958135	*MUC3A*	TG	CC	missense_variant	MED
PR44699	100958135	*MUC3A*	TG	CC	missense_variant	MED
PR44703	100958247	*MUC3A*	C	G	missense_variant	MED
PR44704	100958251	*MUC3A*	A	C	missense_variant	MED
PR50551	100958270	*MUC3A*	TG	CA	missense_variant	MED
PR50551	100958276	*MUC3A*	C	G	missense_variant	MED
PR44704	100958277	*MUC3A*	CA	TG	missense_variant	MED
PR44704	100958290	*MUC3A*	A	G	missense_variant	MED
PR44702	100958352	*MUC3A*	AAAT	GATC	missense_variant	MED
PR50551	100958396	*MUC3A*	T	C	missense_variant	MED
PR50554	100958396	*MUC3A*	T	C	missense_variant	MED
PR44702	100958396	*MUC3A*	T	C	missense_variant	MED
PR50548	100958426	*MUC3A*	C	T	missense_variant	MED
PR50554	100958428	*MUC3A*	AG	CA	missense_variant	MED

PR50551	100958437	*MUC3A*	C	G	missense_variant	MED
PR50548	100958447	*MUC3A*	G	C	missense_variant	MED
PR50551	100958447	*MUC3A*	G	C	missense_variant	MED
PR50548	100958459	*MUC3A*	C	G	missense_variant	MED
PR50551	100958491	*MUC3A*	G	A	missense_variant	MED
PR50551	100958744	*MUC3A*	A	G	missense_variant	MED
PR44697	100958795	*MUC3A*	G	C	missense_variant	MED
PR44697	100958831	*MUC3A*	ACTCT	CATCC	missense_variant	MED
PR44703	100958845	*MUC3A*	A	G	missense_variant	MED
PR44703	100958870	*MUC3A*	C	G	missense_variant	MED
PR44703	100958882	*MUC3A*	C	T	missense_variant	MED
PR50548	100958912	*MUC3A*	CAATC	TGATA	missense_variant	MED
PR50551	100958912	*MUC3A*	CAATC	TGATA	missense_variant	MED
PR44703	100958912	*MUC3A*	CA	TG	missense_variant	MED
PR50548	100958923	*MUC3A*	A	T	missense_variant	MED
PR44699	100958977	*MUC3A*	A	G	missense_variant	MED
PR44699	100959091	*MUC3A*	G	C	missense_variant	MED
PR44699	100959497	*MUC3A*	T	C	missense_variant	MED
PR44699	100959868	*MUC3A*	G	T	missense_variant	MED
PR44703	100959868	*MUC3A*	G	T	missense_variant	MED
PR44701	100964807	*MUC3A*	G	C	missense_variant	MED
PR44703	100964820	*MUC3A*	T	C	missense_variant	MED
PR50552	100964838	*MUC3A*	C	T	missense_variant	MED
PR44703	100965345	*MUC3A*	C	A	missense_variant	MED
PR50548	100966478	*MUC3A*	C	T	missense_variant	MED
PR44701	100966478	*MUC3A*	C	T	missense_variant	MED
PR44703	100966478	*MUC3A*	C	T	missense_variant	MED
PR50548	100966916	*MUC3A*	T	C	missense_variant	MED
PF50549	100966916	*MUC3A*	T	C	missense_variant	MED
PR50550	100966916	*MUC3A*	T	C	missense variant	MED
PR50551	100966916	*MUC3A*	T	C	missense_variant	MED
PR50552	100966916	*MUC3A*	T	C	missense_variant	MED
PR50554	100966916	*MUC3A*	T	C	missense_variant	MED
PR44697	100966916	*MUC3A*	T	C	missense_variant	MED
PR44699	100966916	*MUC3A*	T	C	missense_variant	MED
PR44703	100966916	*MUC3A*	T	C	missense_variant	MED
PR50554	100966942	*MUC3A*	CA	C	frameshift_variant	HIGH
PR50549	100967142	*MUC3A*	AT	A	frameshift_variant	HIGH
PR50552	100967146	*MUC3A*	C	A	missense_variant	MED
PR44699	100967146	*MUC3A*	C	A	missense_variant	MED

Appendix 7

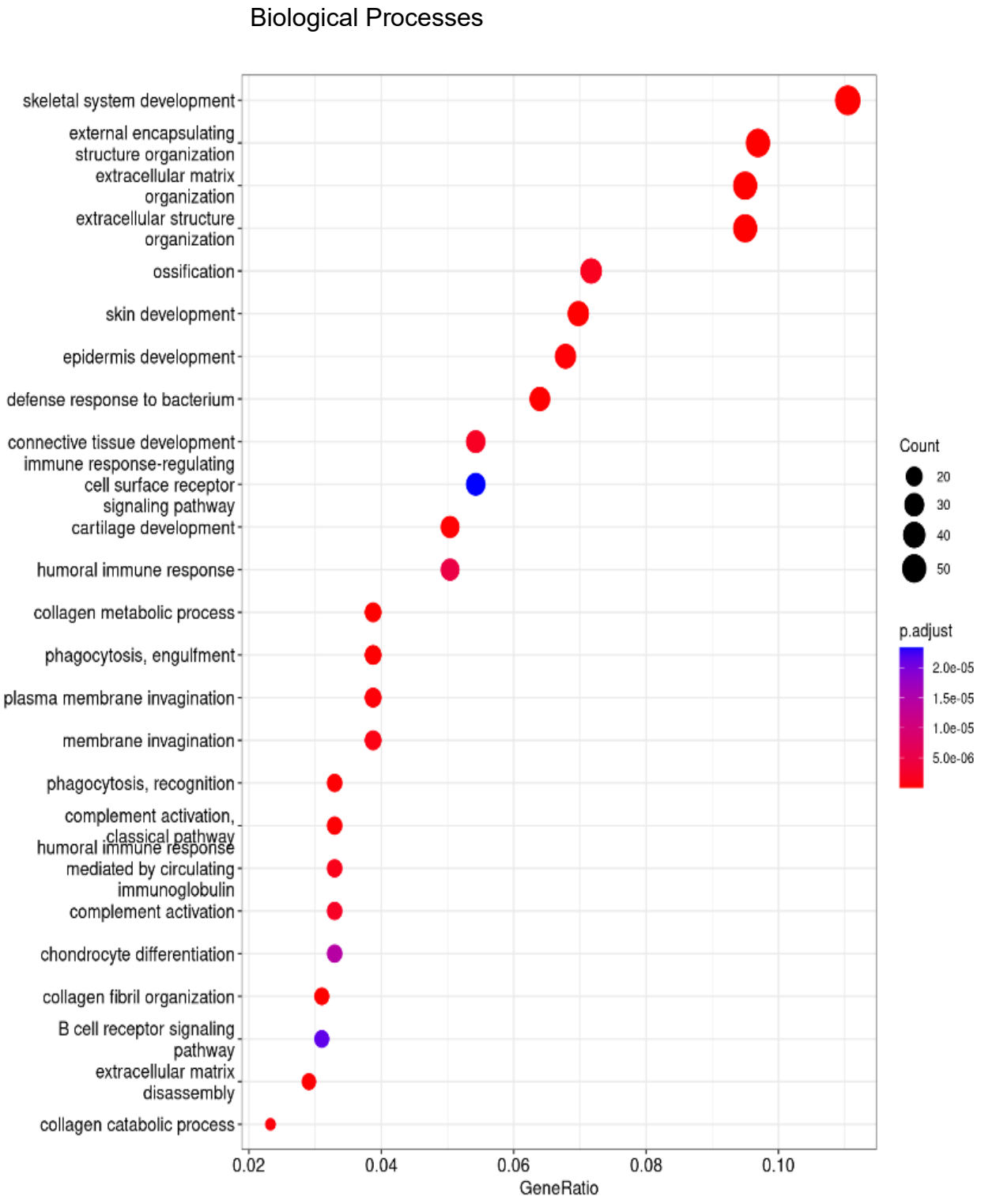


Figure A7.1: Dot plots showing the top 25 enriched GO terms for Biological Processes, where the LFC threshold was set at 2. The largest gene ratios are plotted in order of gene ratio while the size of the dots represents the number of genes in the significant DE genes list. Dot colour represents the *P*-adjusted values (BH).

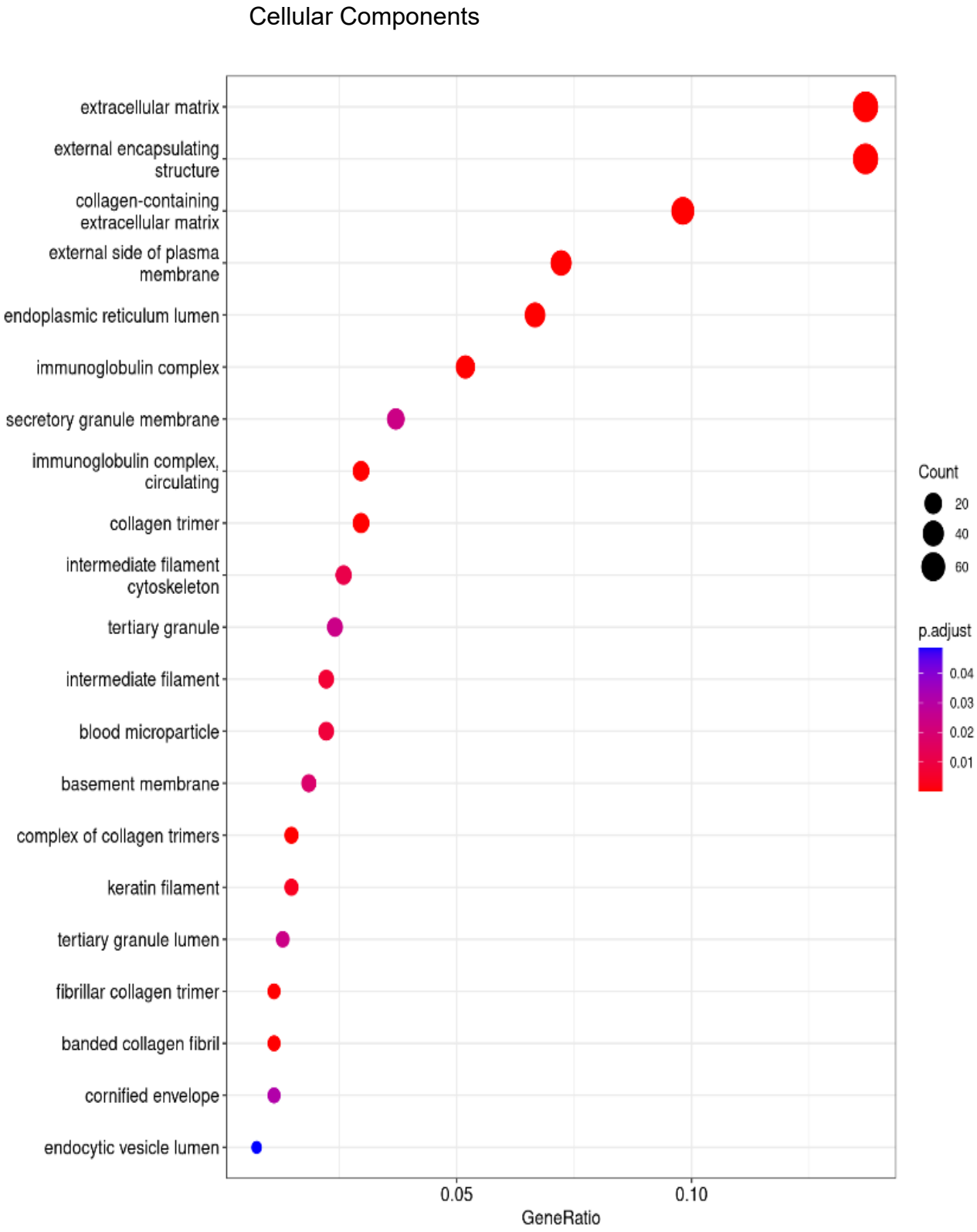


Figure A7.2: Dot plots showing the top 25 enriched GO terms for Cellular Components, where the LFC threshold was set at 2. The largest gene ratios are plotted in order of gene ratio while the size of the dots represents the number of genes in the significant DE genes list. Dot colour represents the *P*-adjusted values (BH).

Biological Processes

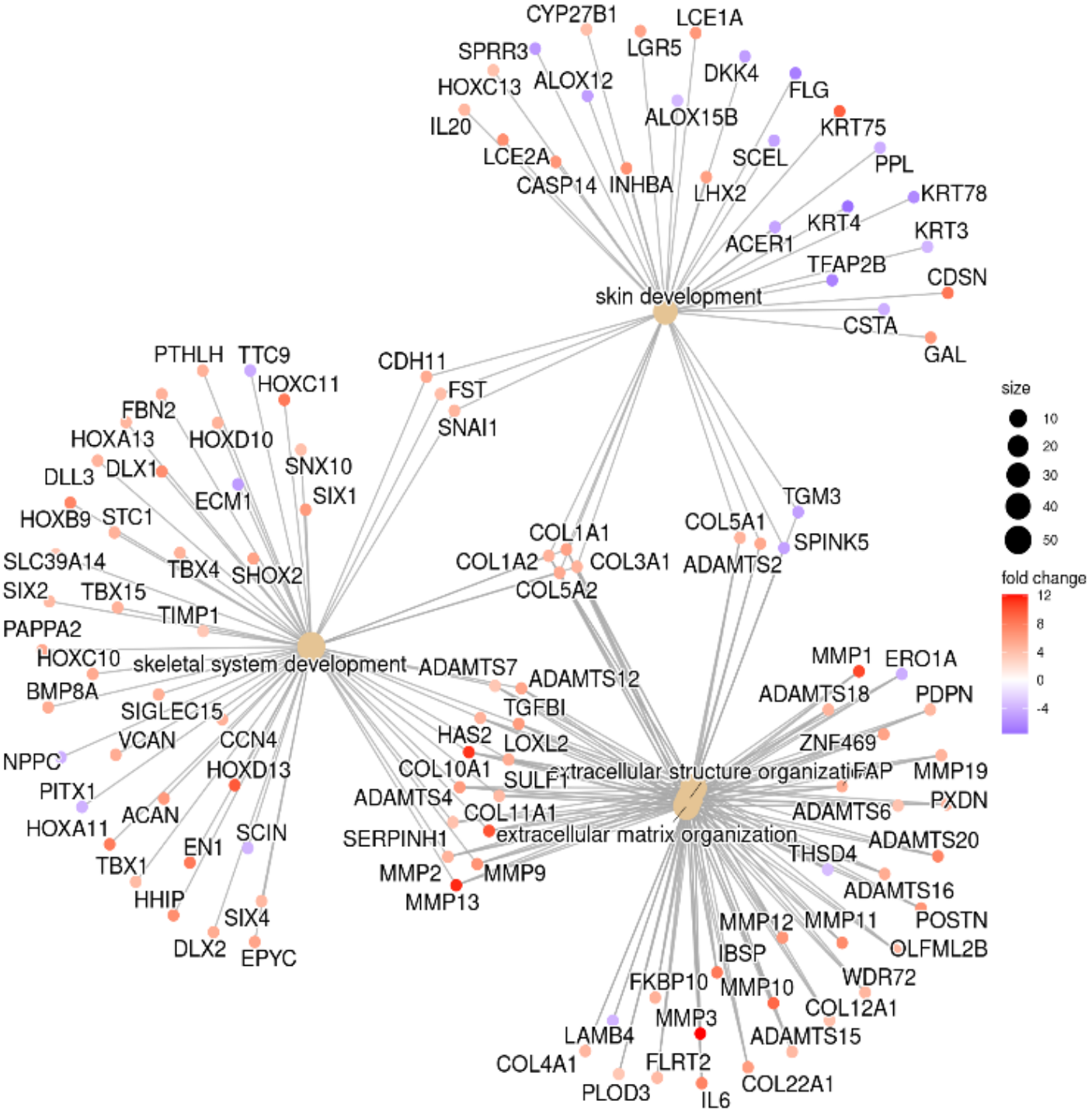


Figure A7.3: Category net-plots of the top most significant GO terms (P-adjusted values) plotted and connected with lines to associate DE genes for Biological processes where the LFC threshold was set at 2. Colours represent the fold changes of the significant genes associated with the GO terms while the size of the terms reflects the *P* values, with the more significant terms being larger.

Cellular Components

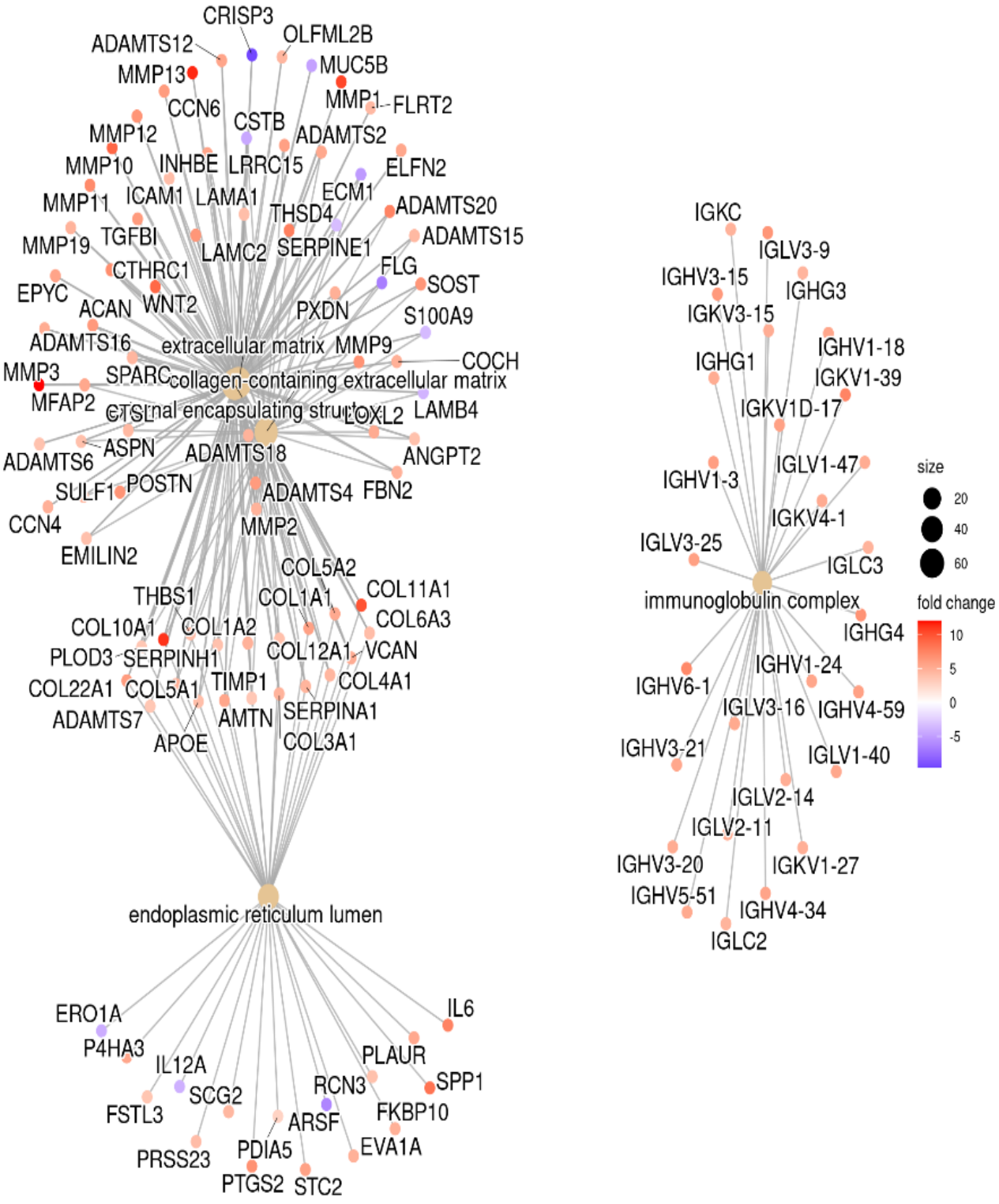


Figure A7.4: Category net-plots of the top most significant GO terms (P-adjusted values) plotted and connected with lines to associate DE genes for Cellular Components, where the LFC threshold was set at 2. Colours represent the fold changes of the significant genes associated with the GO terms while the size of the terms reflects the *P* values, with the more significant terms being larger.