

https://dspace.org/

Integrated methylome and transcriptome analysis of esophageal squamous cell carcinoma in the South African cohort

Moodley, Mishalan

2024

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/47340

Moodley, Mishalan . (2024). Integrated methylome and transcriptome analysis of esophageal squamous cell carcinoma in the South African cohort [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/47340

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

INTEGRATED METHYLOME AND TRANSCRIPTOME ANALYSIS OF ESOPHAGEAL SQUAMOUS CELL CARCINOMA IN THE SOUTH AFRICAN COHORT

Mishalan Moodley



A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg in fulfilment of the requirements for the degree of Doctor of Philosophy in Molecular Medicine and Haematology

Johannesburg, 2024

Declaration

I, Mishalan Moodley, declare that this thesis is of my own, unaided work. Where others have contributed and assisted, they have been duly acknowledged. It is being submitted for the Doctor of Philosophy Degree at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



Signed on the 30 day of October 2024

Dedications

I dedicate this work to Sadhu Ram Ji.

Peer-Reviewed Publications Arising from the Research Project

The research findings arising from this study were drafted and presented in three separate manuscripts. One of these has been published in international peer-reviewed scientific journals. The second and third manuscripts have been submitted for publication. The signed co-author agreement/supervisor declaration for the use of each manuscript can be found in Appendix A.

Title: Processing and Analysis of Tissue Samples from Esophageal Cancer Patients in an African Setting

Lucien Ferndale,[1,2,*] **Mishalan Moodley**,[3,*] Wenlong C. Chen,[4,5] Reubina Wadee,[6] Colleen A. Wright,[6,7] Mohamed Iqbal Parker,[8] Pascale Willem,[3] and Christopher G. Mathew[5,9]

*Equal contribution.

Abstract

Esophageal squamous cell carcinoma is highly prevalent in Eastern and Southern Africa, characterized by a high incidence and mortality. The etiology of Esophageal Squamous Cell Carcinoma is complex, and multiomic studies to improve our understanding of this deadly disease is severely lacking in this part of the world. DNA methylation is an epigenetic mechanism which plays an important role for gene regulation and frequently displays aberrant patterns in several cancers. This study aimed to identify aberrant methylation patterns in Black South African ESCC patients from matched tumour normal pairs.

We designed an efficient workflow for the collection, storage, and processing of high quality ESCC specimens ensuring preservation of tissue architecture and sample quality. A total of 142 ESCC samples were collected. Methylome and transcriptome was investigated in 11 paired tumour and adjacent normal ESCC cases. Differential methylation analysis revealed significant alterations in CpG methylation across the genome. Principal component analysis effectively discriminated the tumour samples from adjacent normal tissues for most samples.

The focus was on differentially methylated regions within enhancers and promoters using a methylation difference of ≥20% and an expression fold change difference of ≥2 (p values ≤0.05) between tumour and normal tissue. We identified three critical gene promoters (CDC42EP3, TNC and KRT13) and 1 enhancer (COL6A3) that exhibited significant differential methylation and a significant inverse relationship with gene expression, all of which plays a role in the ECM. From transcriptome data, there were 241 identified differentially expressed genes, of which 170 were upregulated and 71 genes that were downregulated and whose gene promoters were and hypomethylated and hypermethylated respectively. The top three candidate genes exhibited an 8-9-fold increase in expression and had the most significant FDR (<10-9 to 10-16), and are are all involved in the ECM pathway: FN1, POSTN and COL6A3. Transcription data also showed that FN1 and POSTN were also among the top 2 differentially expressed genes across all genes, with COL6A3 also making up the top 3 differentially expressed genes.

In summary this study sheds light on the methylome and gene expression changes in ESCC biased towards the ECM and highlights the importance of epigenetic modifications and their potential role in the disease. Understanding these molecular alterations is critical for developing targeted therapeutic strategies for ESCC.

Acknowledgements

I want to convey my genuine thanks and deep appreciation to the numerous individuals who have accompanied me on this incredible voyage:

To my primary supervisor, Dr Pascale Willem, my sincere gratitude for your invaluable guidance, unwavering support, and mentorship throughout the course of my thesis. Your expertise and encouragement, always delivered with a smile, have been instrumental in shaping my journey and fostering my growth as a cancer researcher. This achievement owes much to your steadfast support.

To my co-supervisor, Professor Christopher Mathew, a thanks for your support. I am grateful for the opportunity to be part of the SBIMB and the ESCC group, and your role in bringing me into these esteemed communities.

Many thanks to my family, friends and colleagues at NHLS, NICD and Separations.

I want to convey my appreciation to Dr. Tun, Dr. Khan, and all the study participants for their generous contributions. Your involvement would have not made this study possible. Finally, I extend my thanks to the funders whose support has made this study possible.

Table of Contents

Declaration i

Dedications ii

Peer-Reviewed Publications Arising from the Research Project iii

Abstract iv

Acknowledgements v

List of Figures x

List of Tables xii

List of Abbreviations xiii

Preface xv

Chapter 1: Literature Review 1

1.2 Clinical Aspect of ESCC: Diagnosis and Staging 3

1.3 DNA Methylation Studies in ESCC 8

1.3.1 DNA Repair Genes 12

1.3.2 Genes Involved in the Cell Cycle 12

1.3.3 Canonical WNT Signalling 12

1.3.4 TGF-β Pathway 12

1.3.5 TP53 and WNT Pathways 13

1.4 Study Aim and Objectives 15

Chapter 2: Summary of Methods 16

2.1 Experimental Design 16

2.2 Objective 1 16

2.3 Objectives 2 and 3 17

2.4 FastQC 17

2.5 Trimming 18

2.6 Alignment and Sequencing Depth of Coverage 19
2.8 Methylkit: QC 22
2.9 Bisulfite Conversion and PCR Bias QC 22
2.10 Secondary Analysis of DMRs Between Tumour Normal Samples 24
2.11 RNA Sequencing 25
2.11.1 Primary Analysis and Quality Control for Transcriptome Data 25
2.11.2 Distribution of A, T, G, and C Bases 25
Chapter 3 (Manuscript 1): 29
Chapter 4: The Methylome of South African Esophageal Squamous Cell Carcinoma (Manuscript 2) 50
4.1 Abstract 50
4.2 Introduction 50
4.3 Methods 52
4.3.1 Specimen Collection, Methods and Study Design 52
4.3.2 Sample Processing Using the Truseq Methyl EPIC Assay 55
4.3.3 Quality Control 55
4.3.4 Differential Methylation Analysis 56
4.3.5 RNA Sequencing 57
4.3.6 Pathway Analysis 58
4.4 Results 59
4.4.1 CpG Differential Methylation across the tumour genome 59
4.4.2 Genome-wide DMR distributions 61
4.4.3 Hypomethylated and hypermethylated pathways across gene promoters 64
4.4.4 Enhancer Methylation Analysis 64
4.4.5 Matched Transcriptome Analysis 66

4.4.6 Genomic Integration 67
4.4.7 Pathway analysis after integrating differentially methylated promoters and enhancers with matched transcriptome data 69
4.4.8 Candidate Gene Prioritisation 70
4.5 Discussion 72
Chapter 5: Investigating the Transcriptome in African Esophageal Squamous Cell Carcinoma (Manuscript 3) 78
5.1 Abstract 78
5.2 Introduction 79
5.3 Methods 84
5.4 Results 87
5.4.1 Transcriptome Profiling of South African ESCC 87
5.4.2 Gene Set Enrichment Analysis in South African ESCC 93
5.5 Discussion 98
5.5.1 ECM Genes are Differentially Expressed in South African ESCC 99
Chapter 6: Discussion and Conclusion 102
6.1 Objective 1 102
6.2 Objective 2 102
6.3 Proposed Model for ECM Remodeling in ESCC Through Aberrant Methylation and Therapeutic Implications 108
6.4 Limitations of the Study 109
6.5 Future Studies 110
6.6 Conclusion 110
List of References 111
Supplementary Information 140
Supplementary Information: Chapter 2 141

Supplementary Information: Chapter 4 142
Supplementary Information: Chapter 5 150
Appendices 152
Appendix A – Ethics Certificate 153
Appendix B – Co-author Agreement form/ Supervisor Declaration 154
Appendix C – Published Manuscripts 156
Appendix D – Plagiarism Declaration 166
Appendix E – Turn-It-In Report 167

List of Figures

Figure 2- 1: Sequencing quality distribution for sample (2 N). The base position is on the horizontal axis, and the base quality is on the vertical axis. Green represents good quality bases with Phred scores ≥28, orange indicates reasonable quality and red indicates poor quality (Phred scores ≥20). 19
Figure 2- 2: An example of a sequencing reads taken from a matched normal and tumour pairs 22
Figure 2- 3: Histogram of percent methylation of cytosine for samples N2 and T2. X-axis: percentage of methylation. Y-axis: frequency 23
Figure 2- 4: CpG coverage distribution for samples N2 and T2. X-axis: log 10 of read coverage, y-axis: Frequency 24
Figure 2-5: Distribution of A, T, G, and C contents for sample 2 N. The horizontal axis represents the position of each base, and the vertical axis represents the percentage of each base..... 26
Figure 2- 6: Structural outline of the ESCC project and list of bioinformatics tools u..... 28

Figure 3- 1: Protocols followed by the two sites for collection, storage, and processing of samples..... 33
Figure 3- 2: Composite photomicrograph of esophageal invasive squamous cell carcinomas..... 40
Figure 3- 3: DNA samples extracted from tumour tissues on a 1% agarose gel 42
Figure 3- 4: Recommended operational workflow for tissue collection in a low-resource setting..... 48

Figure 4- 1: First genome-wide methylation study on 3.3 million CpG targets with matched transcriptome data in ESCC patients 52
Figure 4- 2: Overview of the bioinformatics approach used for this study 59
Figure 4- 3: Pie chart illustrating hypo (blue) and hypermethylated (red) DMCs in tumour vs normal samples 60

Figure 4- 4; Volcano plot of DMCs 60
Figure 4- 5: Illustrations of differential methylation across the genome and PCA analysis between our 2 groups..... 61
Figure 4- 6: Distribution of DMRs relative to CpG islands and gene features 63
Figure 4- 7: Circos plot illustrating differentially methylated promoters 68

Figure 5- 1: Normalisation of RNA transcripts 89
Figure 5-2: PCA, volcano and heatmap between tumour and normal samples..... 92
Figure 5-3: Pathways enriched after transcriptome analysis with BPs (A) and Reactome (B). “Count” = number of genes..... 95
Figure 5-4: GSEA and Volcano plots illustrating over representation of the ECM in our cohort 97

Figure 6-1: The figure illustrates the proposed model for ECM remodeling in ESCC…109

Figure S4- 1: Pathways analysis from methylome profiling 144
Figure S4- 2: Enhancer Pathway Analysis 145
Figure S4- 3: Pathways enriched after transcriptome analysis..... 147

Figure S5- 1: Pathways enriched from transcriptome data across KEGG 150

List of Tables

Table 1- 1: Risk Factors associated with ESCC 4
Table 1- 2: List of all known methylation studies carried out globally in ESCC 9

Table 2- 1: Main sequencing statistics generated from methylome sequencing 20

Table 3- 1: Number of Biopsies Collected at Each Site 38
Table 3- 2: Quality Control Statistics for DNA Exome Sequencing Data from the Tissue DNA Samples 43
Table 3- 3: Cost per sample breakdown for the two protocols 44

Table 4- 1: Patient characteristics of specimens processed in this study 54
Table 4- 2: Top-ranking 20 DEGs, according to most significant p value (False Discovery Rate – FDR) 66
Table 4- 3: Promoters and enhancers demonstrating significant methylation variations (≤20%) with q-value ≥0.01 and RNA fold-changes in opposite direction (≤2 logFC) 71

Table 5- 1: Gene expression studies on ESCC over the past decade 81
Table 5- 2: Patient characteristics of specimens processed in this study 85
Table 5- 3: The number of reads after normalisation and filtering for all samples in our cohort 88

List of Abbreviations

AJCC	American Joint Committee on Cancer
BP	Biological processes
CAF	Cancer-associated fibroblasts
CD	Cluster of differentiation
CG	Cytosine Guanine
CGI	CpG islands
CHH	H is any base except G
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital
CNS	Central nervous system
COV	Coverage file
CTOB	Complementary To Original Bottom
CTOT	Complementary To Original Top
DEG	Differentially expressed genes
DMC	Differentially methylated cytosines
DMR	Differentially methylated regions
EAC	Esophageal Adenocarcinoma
EBV	Epstein Barr virus
ECM	Extracellular matrix
ESCC	Esophageal squamous cell carcinoma
FANTOM	Functional ANnoTation Of the Mammalian genome
FAP	Fibroblast activation protein
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
GEO	Gene Expression Omnibus
GH	Grey's Hospital
GO	Gene Ontology
GPRC	G protein-coupled
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-wide association studies
H&E	Hematoxylin and eosin
IARC	International Agency for Research and Cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
OB	Original Bottom
OCT	Optimal cutting temperature
OSCC	Oesophageal squamous cell carcinoma
OT	Original Top
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
PDGF	Platelet-derived growth factor
PE	Paired-End

QC	Quality control
RIN	RNA Integrity Number
RNA	Ribonucleic acid
SA	South Africa
SAM	Sequence Alignment Map
SFRP	Secreted frizzled-related proteins
SLIM	Sliding window model
SNP	Single Nucleotide Polymorphism
SPINK	Serine protease inhibitor Kazal
SSA	Sub-Saharan Africa
TCGA	The Cancer Genome Atlas
TNM	Tumour-node-metastasis
TSS	Transcription start sites
UCSC	University of California Santa Cruz Genome Browser
UKZN	University of KwaZulu-Natal
US	United States
UV	Ultra Violet

Preface

This thesis is presented in a manuscript format following the University of the Witwatersrand guidelines for thesis/dissertation submissions. This PhD thesis is presented in the "divided-block" format, which includes five chapters and three manuscripts:

Chapter 1: Introduction

This chapter describes the literature on the somatic genetics of esophageal squamous cell cancer and the study's aims and objectives.

Chapter 2: Summary of Methods

The methods are described entirely in each manuscript. This chapter provides an overview of the methodological approaches used in each manuscript.

Chapter 3: Processing and Analysis of Tissue Samples from Esophageal Cancer Patients in an African Setting

This is a published manuscript and the first results chapter (3). In this chapter, the quality and quantity of genomic DNA and RNA isolated from fresh biopsy tissue are described. High DNA integrity is essential for genomic studies. We established that the nucleic acid was of good quality and suitable for downstream processing for the planned genomic testing, which included methylome and expression analysis.

Chapter 4: Methylome of Esophageal Squamous Cell Carcinoma in a Cohort of South African Patients Reveals Dysregulation of the Extracellular Matrix Pathway

This is the second manuscript that will be submitted for publication. This chapter describes work in which we tested the methylation status of 3.3 million CpG sites across matched tumour normal specimens in South African ESCC patients, something that has not been investigated previously in an African ESCC cohort. Methylome data were also correlated with transcriptome data on the same cohort in the same manuscript.

Chapter 5: Investigating the Transcriptome in South African Esophageal Squamous Cell Carcinoma

This is the third manuscript that will be submitted for publication. This chapter describes in more detail the expression profile we detected in South African ESCC patients.

Chapter 6: Discussion and Conclusions

This chapter integrates and discusses the three manuscripts described in chapters 3-5 as well as limitations of the study and future directions and suggestions. This chapter also includes a conclusion of the entire project.

Chapter 1: Literature Review

Cancer is the second leading cause of death worldwide and accounts for 10 million deaths globally each year, with the global incidence expected to rise by 28.4 million (47%) cases by 2040 (Sung *et al.*, 2021). Cancer is a multifactorial disease with genetic, environmental and lifestyle factors contributing to the disease. From a genetic point of view, cancer can be classified into two groups: inherited predisposition and sporadic cancers. Inherited genetic mutations increase the risk of developing cancer and run in families. The most common inherited mutations involve highly penetrant germline mutations in the *BRCA1* and *BRCA2* genes, which increase the risk of developing breast and/or ovarian cancer (van der Groep *et al.*, 2006). However, 80% of cancers are sporadic with unknown causes (3). Initially, sporadic cancers were thought to have no germline genetic component. However, genome-wide association studies (GWAS) have identified several low-risk loci (mostly SNPs) distributed throughout the genome and associated with a modest risk of cancer. The combined effect of these low penetrance variants is assigned a polygenic risk score that can have a substantial impact on the risk of disease. The polygenic basis of sporadic cases explains the initial misguided belief that sporadic cancers have little or no germline component (Y. Lu *et al.*, 2014).

A mutagen is an agent that can introduce mutations and/or cause chromosome breakage in somatic cells and may lead to an increased cancer risk. A number of studies have identified mutagens that can affect DNA; they are classified into three groups: chemical, physical and biological mutagens. Chemical mutagens include smoking, high alcohol consumption, and exposure to asbestos, insecticides, herbicides and heavy metals. Physical mutagens refer to different types of radiation, such as X-rays, solar radiation and UV waves. Biological mutagens include microorganisms such as some type of viruses linked to cancer (Salem, 2016).

1.1 Anatomy of the Esophagus and Esophageal Squamous Cell Carcinoma

The adult human esophagus is a relatively straight muscular tube (18-25 cm in length) starting from the pharyngoesophageal junction in the neck (C5-6 vertebral level), extending posteriorly through the gastroesophageal junction and ending at the orifice of

the cardia of the stomach (T11 vertebral level) (Kuo and Urma, 2006). The wall of the esophagus is made up of four layers: mucosa, submucosa, muscularis propria and adventitia (Kuo and Urma, 2006). There are two main histological types of malignant esophageal tumours: esophageal squamous cell carcinoma (ESCC), which occurs mainly in the middle or upper part of the esophagus, and esophageal adenocarcinoma (EAC), which generally occurs in the lower third of the esophagus. This thesis relates to ESCC. ESCC develops from the native squamous epithelial cells that line the esophagus. ESCC might arise from inflammation, dysplasia, hyperplasia or local injury (Jain and Dhingra, 2017). The characteristics of well-differentiated ESCC include keratinocytes accompanied by intercellular bridge keratin, while poorly differentiated ESCC usually contains in situ lesions, squamous-oriented infiltration and intraepithelial neoplasia (Jain and Dhingra, 2017).

In recent decades, EAC has increased and ESCC has decreased in developed countries. However, ESCC is the predominant histological subtype globally, accounting for 87% of all esophageal cases (Arnold *et al.*, 2014). ESCC is the seventh most common cancer worldwide and is one of the most common cancers in most Eastern and Southern African countries (Huang *et al.*, 2021). ESCC is associated with geographically endemic regions, including Northern Iran, China, Japan, Central Asia, South America, African Americans in the US and Northern France (Islami, Pourshams, *et al.*, 2009). In central Asia and Sub-Saharan Africa (SSA), the incidence of OSCC is 20-fold higher than that seen in Western countries (Somdyala *et al.*, 2010). Cancer of the esophagus was an uncommon disease in South Africa until 1930; thereafter, the number of ESCC cases increased quickly (Alouna et al., 2019). The most recent review of ESCC in South Africa shows that this disease accounts for the second highest number of cancer-related deaths in South Africa (SA), with squamous cell carcinoma being the most prevalent type, with an incidence of 46.7/100,000 and 19.2/100,000 for males and females, respectively (Alouna et al., 2019). In the Eastern Cape, ESCC is the most common cancer among males and the second most common cancer among females (Bray *et al.*, 2015). Kwazulu-Natal also has a high incidence of ESCC (Ferndale *et al.*, 2019).

1.2 Clinical Aspect of ESCC: Diagnosis and Staging

Esophageal squamous dysplasia is regarded as the precursor lesion of ESCC (Wang *et al.*, 2005; Codipilly *et al.*, 2018). Progressive dysplasia is the most common symptom that is first diagnosed and occurs in 74% of ESCC cases. This is followed by involuntary weight loss (57% of ESCC patients) and pain on swallowing (odynophagia) in 17% of ESCC patients (Daly *et al.*, 2000). Other signs and symptoms include vomiting, nausea, chest pain, loss of appetite and hoarseness (Harris, Croce and Munkholm-Larsen, 2017). There are three stages of pathological changes during ESCC development. Alterations occur from normal esophagus (or low-grade intraepithelial neoplasia) to middle- or high-grade intraepithelial neoplasia and finally to invasive carcinoma (Aaltonen *et al.*, 2000). Patients with high-grade squamous dysplasia are at high risk for developing ESCC compared to patients with low-grade intraepithelial neoplasia (Wang *et al.*, 2005). Esophagogastroduodenoscopy (EGD) is the standard method for diagnosing ESCC (Sano *et al.*, 2017). After ESCC diagnosis, staging is critical for ESCC outcome. ESCC is based on the tumour-node-metastasis (TNM) subclassifications curated by the American Joint Committee on Cancer (AJCC). The TNM system is based on the depth of invasion of the primary tumour (T0-4), lymph node involvement (N0-3) and extent of metastatic disease (M0-1)(Rice *et al.*, 2017).

A number of risk factors have been associated with ESCC, including poor nutrition, tobacco smoking, alcohol consumption, smoke inhalation from cooking and heating fires, unhygienic water sources, consumption of maize using different cooking methods and the presence of an oncovirus such as herpesvirus, Epstein Barr virus and/or human papillomavirus (Liu et al., 2016). In China, Iran and South Africa, smoking and alcohol consumption are thought to contribute to ESCC (Islami, Pourshams, *et al.*, 2009); (Gao *et al.*, 1994) and (Segal *et al.*, 1988), although this is insufficient to explain the high incidence of the disease. Additional risk factors for ESCC have been proposed in Africa and China, including exposure to fumonisin, a Fusarium fungi toxin that grows on maize (Marasas *et al.*, 1988) (Sun *et al.*, 2007), and self-induced vomiting, a cultural practice among some South African black populations on the eastern coast of the country (Matsha

et al., 2006). More information about risk factors and their degree of impact in ESCC patients is presented in Table 1.1.

Table 1- 1: Risk Factors associated with ESCC

Risk factor	Risk of ESCC
Smoking	Higher (Niu *et al.*, 2021)
Alcohol consumption	Higher (Niu *et al.*, 2021)
Drinking water obtained from contaminated source	Higher (Niu *et al.*, 2021)
Hot tea consumption	Medium (Niu *et al.*, 2021)
Salty Diet	Higher (Niu *et al.*, 2021)
Diet including poultry meat, fruits, milk, acid suppressants, vitamins, antibiotics and purified water	Lower (Niu *et al.*, 2021)
Physical exercise and/or housework	Lower (Niu *et al.*, 2021)
Family history of cancer	Higher (Niu *et al.*, 2021)
Refrigerator use	Lower (Niu *et al.*, 2021)
Betel Quid	Higher (Wu *et al.*, 2006)
Polycyclic aromatic hydrocarbons	Higher (Roshandel *et al.*, 2012)
Gastric dysbiosis	Higher (Nasrollahzadeh *et al.*, 2015)
Human papillomavirus (HPV)	Medium (L. Guo *et al.*, 2016)
Pickled vegetables	Higher (Islami, Ren, *et al.*, 2009)
Tooth loss and poor oral hygiene	Higher (Abnet *et al.*, 2008)
Helicobacter pylori	Lower (Xie *et al.*, 2013)
Low socioeconomic status	Higher (Chen *et al.*, 2022)

Despite all of the risk factors listed in Table 1.1, the aetiology of ESCC remains generally unclear, and risk factors associated with the disease may differ from one geographic region to another.

The ESCC hot spot regions in Africa, described as the ESCC corridor, include Burundi, Ethiopia, Rwanda, Uganda, Kenya, Malawi, Tanzania, and South Africa (Munishi *et al.*, 2015; Schaafsma *et al.*, 2015). In Sub-Saharan Africa, ESCC develops in younger patients compared to other regions of the world (Kayamba *et al.*, 2015). Most ESCC patients in South Africa present with an advanced stage of the disease, and the median survival of these patients is only approximately 120 days *(*Bergquist *et al.,* 2005). Thus, the abysmal outcome of this disease underscores the need for early markers of disease and novel therapeutic options.

A landmark discovery in the 1990s was the discovery of the *BRCA1* and *BRCA2* genes, mutations in which are associated with a high risk of breast and other cancers (Bowcock, 1993). This discovery signalled a new era of our understanding of cancer genetics. Subsequent sequencing of the human genome and genomic profiling of cancers over the past decade has provided us with insight into the impact of germline (heritable) and somatic (acquired) aberrations on cancer (Alldredge and Randall, 2019).

At a molecular level, tumours progress through a range of different mechanisms occurring within and outside the cell. Defective cell signalling pathways affect BPs within the cell, including unregulated proliferation, cell growth, differentiation and cell death (Murakami and Simons, 2008). Tumours proliferate because of oncogene activation or insensitivity to growth-inhibitory signals, resistance to apoptosis or failure to repair damaged DNA caused by a carcinogen. Vascular supply is also important for tumour growth (angiogenesis). Mutations in genes that regulate angiogenesis, such as vascular endothelial growth factor (*VEGF*), basic fibroblast growth factor (*bFGF*) and platelet-derived growth factor (*PDGF*), are seen in most cancers (Murakami and Simons, 2008).

A number of recent technologies that allow genome-wide study of tumour-associated somatic variants have revolutionised the field of cancer. Among these, microarrays reveal loss of heterozygosity and copy number changes (genomic regions of gains and losses), while NGS allows sequencing of the entire exome (coding genome) or whole-genome of

tumours. Sequence comparison between tumours and germ line DNA (peripheral blood or adjacent histologically normal tissue) from the same patient uncovers tumour-specific somatic variants. These two molecular techniques have identified nonrandomly mutated genes and signalling pathways in a number of malignancies, which has allowed the development of personalised therapies implemented in lung, stomach, breast and colorectal cancer, as well as haematological malignancies (Chen *et al.*, 2016).

In 2012, the first whole-exome sequencing study on paired tumour/normal DNA in ESCC was performed on 12 ESCC individuals in the USA. Mutations and copy number alterations were found in the *TP53, NOTCH1, PIK3CA* and *FAT1* genes (Agrawal *et al.*, 2012). Subsequently, several NGS studies have been performed globally on ESCC patients. Zhang *et al.* (2022) reviewed the latest review of somatic genomic data in ESCC patients using the cBIOPortal for cancer genomics (https://www.cbioportal.org). This database integrates genomic data from several sources and is highly credible (Zhang, Wang and Meng, 2022). The top 15 most frequently mutated cancer-related genes in ESCC included *TP53, PIK3CA, KMT2D, LRP1D, PCLO, NFE2L2, NOTCH1, NOTCH3, TRRAP, SMARCA4, KMT2C, ZNF750, TGFBR2, PTCH1* and *HDAC4*. Genes that are frequently involved in copy number variants include *CCND1, FGF3, FGF4, FGF19, CDKN2A, CDKN2B, LPR1B, TBL1XR1, TERC, LPP, PRKCL, TP63, PIK3CA, MECOM* and *MTAP* (Zhang, Wang and Meng, 2022). These alterations affect the cell cycle (*CDKN2A, CDKN2B, CCND1, TP53*), *RTKMAPK/PI3K* pathway (*FGFs, PIK3CA*) and epigenetic modulators (*KMT2C, KMT2D*) (Zhang, Wang and Meng, 2022).

The current review of genetic studies in South Africa informs us that 23 ESCC studies have been performed, of which 16 studies investigated germline susceptibility (Simba *et al.*, 2019). While the ESCC tumour genome has been investigated in patients from Asian countries, the existing literature on genomic sequencing studies in Africa comprises just two investigations.. The first study was conducted in Malawi on 59 ESCC cases subjected to whole-exome and transcriptome analysis (Liu *et al.*, 2016). The tumour genome revealed gene mutations common to previous studies performed in Asia and America as well as a unique set of gene mutations not reported before in this cancer. Apart from deletions on chromosome 1q31, mutated genes were similar to those found in the North

American and Asian ESCC cohorts, namely, *TP53, NOTCH1, FAT1, NFE2L2*, and *FBXW7*, as well as copy number alterations in *CCND1* and *CDKN2A* (Zhang, Wang and Meng, 2022; Liu *et al.*, 2016). In a single African ESCC study, whole-exome sequencing revealed three mutational signatures, of which two are common to the Chinese population (Liu *et al.*, 2016). The first common signature is characterised by C to T transition mutations, an age-associated signature common to many cancer types. The second signature is characterised by both C to G transversions and C to T transitions that occur at TCN trinucleotides, which closely resembles the APOBEC-associated cytidine deaminase signature found in several malignancies. In addition, a third signature was identified in the African cohort, previously unreported in any other ESCC cohort. This signature is associated with gingival buccal oral squamous carcinoma, similar to signature 29. Gene mutations from this signature include: *TP53*, *FAT1*, *CASP8*, *HRAS* and *NOTCH1* and is commonly seen in tobacco-chewing-related gingival buccal squamous cell carcinoma. This signature is characterized by C to A transversions (NCA trinucleotides) and C to T transitions (NCG trinucleotides). The study attributed these mutations to factors such as age-related cytosine deamination, APOBEC-induced cytosine deamination, and an unidentified mutagenic agent. While HPV and EBV were ruled out, further research is needed to pinpoint the exact cause of this signature in SSA ESCC. This is the first time that this signature was reported in an ESCC cohort, and the authors suggested that it could be attributed to an unknown mutagenic process. In the same study, RNA expression analysis showed that ESCC could be divided into three distinct subtypes characterised by their expression of cell cycle and neural transcripts. The authors concluded that the mechanism of mutagenesis of ESCC in Africa differs when compared to the rest of the world, and thus, the epidemiological cause of ESCC in Africa could be attributed to factors other than exposure to a carcinogen such as tobacco and an oncogene virus (Liu *et al.*, 2016). The second more recent study investigated exome sequencing in ESCC patients from Eastern Africa: Kenya (68), Malawi (59), Tanzania (35). Similar to the initial study, the predominant portion of cases (88% and 91%) exhibited mutational signatures related to APOBEC activity (SBS2 and SBS13) (Moody et al, 2021).

1.3 DNA Methylation Studies in ESCC

In addition to loss of heterozygosity, copy number changes and genomic mutations, the tumour genome can be affected by an aberrant methylation pattern. DNA methylation inhibits gene expression by modifying the DNA structure without changing the coding sequence of the genome. DNA methylation is the covalent addition of a methyl group to cytosine (5-methylcytosine (5mC)). Most CpG dinucleotides are methylated; however, those within CpG islands (CGIs) are usually unmethylated (Deaton *et al.*, 2011; Jones, 2012). Carcinogenesis is often characterised by genome-wide methylation changes between tumour and normal specimens. Generally, there is a pattern of global hypomethylation and focal hypermethylation in promoters and first exons across the tumour methylome (Dong *et al.*, 2014). Methylated CpGs prevent transcription factor binding in tumour-suppressor genes, which results in decreased gene expression and loss of tumour-suppressor function. Consequently, this leads to uncontrolled cell growth and tumour development (Dong *et al.*, 2014).

Several Asian studies have investigated the methylation status of different genes in ESCC patients over the past two decades (Table 1- 2). These studies compared methylation patterns between ESCC lesions and adjacent normal esophageal tissue of the same patient or plasma DNA from normal healthy controls.

Table 1- 2: List of all known methylation studies carried out globally in ESCC

Hypermethylated genes*	Number of ESCC cases	Techniques used	Geographical location and Research group
RAR-β (55%) and *FHIT* (45%)	47 ESCC tumour/normal tissue	Methylation specific PCR (MSP)	Asia (Kuroki *et al.*, 2003)
TAC1 (50%)	24 ESCC tumours and circulating DNA in plasma samples	RT-MSP	North America (Jin *et al.*, 2007)
CDH1 (43%) and Integrin alpha 4 (21%)	251 ESCC (only)	MSP	Asia (Lee *et al.*, 2008)
TFF1	106 ESCC and 27 adjacent normal tissue	Illumina GoldenGate Assay (1505 CpG dinucleotides for 807 genes, Illumina)	Asia (Lima *et al.*, 2011)
RUNX3 (51%)	70 cell free DNA samples	MSP	Asia (Zheng *et al.*, 2011)
RASSF10 (44%)	88 ESCC tumours and five adjacent normal tissue	MSP, expression of *RASSF10* confirmed with RT–PCR	Asia (D. Lu *et al.*, 2014)
DDIT3, *FES*, *FLT3*, *NTRK3*, *SEPT5*, *SEPT9*, *SOX1, SOX17*	40 ESCC tumour/normal tissue	Illumina GoldenGate Assay (1505 CpG dinucleotides for 807 genes, Illumina). The results confirmed with pyrosequencing (Qiagen)	Asia (Kuo *et al.*, 2014)
SOX17	169 ESCC and tissue from nine healthy controls	MSP,RT–PCR, Immunohistochemistry. The results confirmed with bisulfite sequencing	Asia (Gao *et al.*, 2014)

Hypermethylated genes*	Number of ESCC cases	Techniques used	Geographical location and Research group
SFRP1 (95%) *SFRP2* (83%)	30 Basaloid ESCC cases	MSP and Immunohistochemistry	Asia (Saito *et al.*, 2014)
DACH1 (62%)	104 ESCC cases	MSP, results confirmed with bisulfite sequencing	Asia (Wu *et al.*, 2014)
RASS2F (35%)	188 ESCC tumour/normal tissue	Bisulfite sequencing and RT–PCR to confirm expression of *RASS2F*	Asia (W. Guo *et al.*, 2016)
CHFR (45%)	109 ESCC and 16 healthy controls	MSP	Asia (Yun *et al.*, 2015)
SFRP1, *SFRP2*, *SFRP4*, *SFRP5*, *SOX17*, *WIF1* (33%). *CDKN2A CHFR* (9%)	57 ESCC tumour/normal tissue	Infinium Methylation 450K array (Illumina)	Asia (Kishino *et al.*, 2016)
BNIP3, *BRCA1*, *CCND1*, *CDKN2A*, *HTATIP2*, *ITGAV*, *NFKB1*, *PIK3R1*, *PRDM16, PTX3*	43 ESCC tumour/normal tissue and 10 tissues from healthy controls.	Infinium Methylation 450K array, results confirmed with Bisulfite NGS (Illumina)	Asia (Peng *et al.*, 2016)
PAX1, ZNF582	14 ESCC tumour/normal	qMSP	Asia (Huang *et al.*, 2017)

Hypermethylated genes*	Number of ESCC cases	Techniques used	Geographical location and Research group
ZNF132	91 ESCC tumour/normal	NGS targeted approach using HiSeq. RNA expression checked with RQ-PCR	Asia (Jiang *et al.*, 2018)
ABCD1, CCDC8, FBXO17	99 ESCC with three normal samples	Bioinformatics approach using methylation and transcriptome data from TCGA database which used 450k array	Asia (Lu *et al.*, 2019)
KMT2D, KMT2C, KDM6A, EP300, CREBBP	20 ESCC Epigenetics modulator genes	Whole-genome bisulfite sequencing	Asia (Cao *et al.*, 2020)
NRN1	1012 tumour and 15 normal samples	Methylation specific PCR	Asia (Du *et al.*, 2021)
PAX9, SIM2, THSD4	108 ESCC and 51 adjacent normal tissue	850k array with transcriptome data from TCGA database	Brazil, Portugal, Iran, Ethiopia, Kenya and India and SA (Talukdar *et al.*, 2021)

The key genes identified from these studies in Table 1- 2 play important roles in key mechanisms of cell cycle signalling, proliferation, apoptosis, DNA damage repair and involved genes of the *WNT*, *TP53* and *TGF-β* pathways. The results from the above methylation studies with regard to the pathways they affect are summarised below.

1.3.1 DNA Repair Genes

Mismatch repair (MMR) proteins recognise mismatches that arise from replication errors in DNA. Two genes that participate in the MMR system, *MLH1* and *MSH2*, are methylated in 32-62% and 29-32% of ESCC tumours, respectively (Tzao *et al.*, 2005; Su *et al.*, 2014; W. Guo *et al.*, 2016). Apart from MMR genes, the *FHIT* gene, which maintains genome stability, is methylated in 45% of ESCC (Kuroki *et al.*, 2003).

1.3.2 Genes Involved in the Cell Cycle

RASS10 and *CHFR* induce G2/M phase cell cycle arrest and are methylated in 44% (D. Lu *et al.*, 2014) and 45% (Yun *et al.*, 2015) of ESCC tumours, respectively.

1.3.3 Canonical WNT Signalling

The activated *WNTβ-catenin* signalling pathway promotes cell proliferation by stimulating *MYC*, *Cyclin D1* and other downstream genes. Secreted frizzled-related proteins (SFRPs) regulate the canonical WNT pathway. *SFRP1* and *SFRP2* are methylated in 95% and 83% of ESCC, respectively (Saito *et al.*, 2014). *SOX17* is another gene that regulates the WNT pathway by degrading the β-catenin/TCF complex. Hypermethylation of the promoter region of *SOX17* reduces *SOX17* expression in ESCC (65%), leading to overexpression of the WNT signalling pathway, causing an increase in cell proliferation and promoting carcinogenesis. *SOX17* methylation is strongly associated with alcohol consumption, and reducing alcohol intake may help to prevent ESCC by inhibiting *SOX17* methylation (Kuo *et al.*, 2014).

1.3.4 TGF-β Pathway

The TGF-β system is a set of multifunctional cytokines that regulate cell growth, differentiation, migration, angiogenesis and apoptosis (Wu *et al.*, 2014). It has been suggested that methylation of *DACH1* in ESCC (62%) activates *TGF-β*, which in turn suppresses *MYC* (Wu *et al.*, 2014).

1.3.5 TP53 and WNT Pathways

One study performed whole-exome sequencing and methylome analysis on the same ESCC samples. The integrated data revealed that genes that negatively regulate the WNT pathway are generally hypermethylated, while *TP53* and other elements in the *TP53* pathway are usually mutated (Kishino *et al.*, 2016). One of the most recent ESCC studies used the 850k array (Illumina®) and found enrichment of hypermethylated genes in theWNT pathway (Talukdar *et al*, 2021).

One can deduce from the methylation studies listed in Table 1- 2 that there is no particular gene or set of genes commonly hypermethylated across all studies. However, evidence from the studies above indicates that aberrant methylation occurs in genes that are involved in key parts of cancer pathways. Some of the studies in Table 1- 2 also suggest that the genes that are hypermethylated play an important role in the early stage of ESCC. *RAR-β*, *FHIT*, *TAC1*, *CDH1*, *SOX17* and DACH1 are methylated in the early stages of ESCC (Kuroki *et al.*, 2003; Jin *et al.*, 2007; Lee *et al.*, 2008; Wu *et al.*, 2014, 2014). These genes could be critical biomarkers for the early diagnosis of ESCC. *RASSF2* (W. Guo *et al.*, 2016) and *CHFR* (Yun *et al.*, 2015) are genes associated with a poor prognosis that are methylated during late stages of ESCC. As mentioned above, whole-exome sequencing (tumour/normal) and RNA transcriptomic analysis from a single study performed in Africa (Liu *et al.*, 2016) suggest that Africa may have unique signatures of ESCC not explained by exposure to carcinogens or oncogenic viruses. A methylome study is likely to inform on environmental factors contributing to the disease, yet it has never been performed in African or South African ESCC patients.

To carry out a comprehensive methylome study, one must take into consideration the cost and effectiveness of techniques used to determine the tumour methylome. Bisulfite sequencing is considered the “gold standard” to determine the CpG methylation status of a DNA sequence (Heyn *et al.*, 2012). However, obtaining the methylation status of the entire genome using bisulfite sequencing is time-consuming and expensive. For this reason, most studies use recently developed DNA methylation microarrays. Since their inception, DNA methylation arrays have been used extensively in cancer studies (Moran, Arribas and Esteller, 2016), autoimmune disease (Sun *et al.*, 2016), mental illness (Nestler *et al.*, 2016) and cardiovascular disease (Zhang and Zeng, 2016). Among commercial platforms, the Illumina Infinium 450k DNA methylation microarray stands out. It has been used in many cancer studies and

has been the selected technique for studies affiliated with The Cancer Genome Atlas (TCGA) project (Liang and Weisenberger, 2017).

Most studies in Table 1- 2 were performed in the past decade and were limited by the number of genes that could be tested simultaneously for aberrant methylation. Only the two most recent ESCC studies (Kishino *et al.*, 2016; Peng *et al.*, 2016) were able to interrogate ~450 000 sites simultaneously for their methylation status owing to recent availability of the Infinium Methylation 450K array (*Infinium Human Methylation 450K BeadChip Support*, no date). However, two minor problems have been identified with the 450k array; both problems are related to the array probe design. The first issue is that the probes used to determine the methylation status have a differing dynamic range, which could potentially create a bias when measuring DNA methylation. Since the advent of the 450k array bioinformatics tools for data analysis, these issues have largely been addressed by ruling out any potential data bias (Moran, Arribas and Esteller, 2016). The second limitation of arrays is the probe design, which does not target enhancers that regulate transcription, as revealed by whole-genome bisulfite sequencing (Pidsley *et al.*, 2016). To counter this problem, Illumina designed the Infinium 850k EPIC BeadChip array, which targets ~850 000 CpG sites, including enhancer regions (Moran, Arribas and Esteller, 2016). Although one study in Table 1-2 used the 850k array across ESCC samples from several world geographical sites (including South Africa), there were a few drawbacks to their study. First, this study used transcriptome data obtained from the TCGA database (instead of matched tumour samples) to correlate with methylation data (Talukdar *et al.*, 2021). While databases such as TCGA are extensive, comparing data from different experiments is challenging due to several factors, such as differences in sample handling and processing, the type of RNA sequencing platform and chemistry used and the analysis pipeline used to normalise batch effects (Wang *et al.*, 2018). Second, the same study also used a short window of only 2 CpGs to infer 67% of their significant differentially methylated regions (DMRs) (Talukdar *et al.*, 2021). In contrast, the literature suggests using a minimum of 3-4CpGs across a 1000 base pair window when infering DMRs (Grolaux et al, 2022; Rezwan et al, 2015). Finally, the top 3 candidate gene promoters (*PAX9*, *SIM2* and *THSD4*) found to be significantly hypermethylated across most of their specimens from different geographical locations were not detected in their South African cohort, nor were any other significant DMRs detected in their South African cohort (Talukdar *et al.*, 2021). This underscores the need to perform genomic studies

of unique geographical locations that could be associated with specific risk factors and aetiologies.

Recently, Illumina has also developed the TruSeq® Methyl Capture EPIC Library Prep Kit. The TruSeq kit builds on the 850k array design and is an NGS-based technique that enables bisulfite sequencing of 3,340,894 CpG sites in the genome simultaneously. It includes additional regions of significance identified by ENCODE, FANTOM5 and the Epigenomics RoadMap Consortium (*TruSeq Methyl Capture EPIC Library Prep Kit Documentation*, no date). In comparison to whole-genome bisulfite sequencing, the TruSeq kit covers five times more methylated sites than the 850k array and is 25% cheaper than the 850k array (*TruSeq Methyl Capture EPIC Library Prep Kit Documentation*, no date)

1.4 Study Aim and Objectives

The main aim of the study was to test the hypothesis that dysregulation of the methylome contributes ESCC carcinogenesis in the South African population.

Specific Objectives

1. To assess whether the quality of samples stored in RNAlater are suitable for histology examination and large-scale genomic studies.
2. To identify novel and potentially African-specific DMRs in ESCC using the TruSeq Methyl Capture EPIC NGS assay and to correlate these findings with transcriptome data generated from the same cohort.
3. To identify a novel and potentially African-specific ESCC expression profile using transcriptome data.

This study achieved these objectives in three manuscripts: Manuscript 1 addressed objective 1, manuscript 2 addressed objective 2, and manuscript 3 addressed objective 3.

Chapter 2: Summary of Methods

This chapter outlines the approach and summarises the methodology used in each manuscript. Details of the methods used are described in the three manuscripts, which encompass chapters 3, 4 and 5. Sample collection, nucleic acid extraction, library preparation for methylome sequencing, validation studies, bioinformatics analysis, writing of chapters and corrections from supervisors were carried out by the PhD candidate.

2.1 Experimental Design

The genomic studies reported in this thesis used prospectively recruited cases. Ethics clearance was obtained from the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Wits) (Certificate number M181188) in line with the Helsinki declaration (2013) and from the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (UKZN) (Certificate number BF270/15). Ethical clearance certificates can be found in Appendix A. Patients enrolled in this study were recruited from the Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and Grey's hospital (GH) in KwaZulu-Natal (KZN).

2.2 Objective 1

The first objective was a collaboration between CMJAH and GH in KwaZulu-Natal (South Africa). Suspected ESCC tumour tissue and matched adjacent normal tissue were collected from 50 CMJAH and 92 tumour tissue from GH ESCC presentation cases, respectively. Specimens were stored in a 2-ml Cryotube and immediately immersed in a vacuum flask containing liquid nitrogen or stored in RNAlater® (Qiagen). Each tissue specimen was divided into two pieces (25%:75%); the smaller portion was processed for histology, and the larger portion was processed for DNA and RNA extraction (Qiagen AllPrep Universal Kit ® (Qiagen)) according to the manufacturer's protocol. Samples with ≥60% tumour content and matched adjacent normal tissue (0% tumour) were the primary criteria for inclusion. The majority of patients in the CMJAH cohort used for this methylome and transcriptome study had 100% tumour content (7/12). Following extraction, both nucleic acids were assessed using Bioanalyzer® (Agilent, US), gel electrophoresis and Qubit quantification. A detailed explanation of sample collection and quality control (QC) is described and published in chapter 3.

After sample collection, histology, nucleic acid extraction and QC of all specimens described in objective 1, the second and third objectives could be carried out, which included methylome and mRNA sequencing on a cohort of 12 matched tumour normal tissue specimens derived from ESCC patients.

2.3 Objectives 2 and 3

Extracted DNA was sonicated using ultrasonication to obtain products of 180–220 base pairs in length according to the company's exact protocol and specifications. First, 500 ng of gDNA was sonicated using a Covaris E220 sonicator (Covaris©, Woburn, MA) to obtain >60% fragments between 100–300 bp. Product sizes were verified on the Labchip© instrument (Perkin Elmer©). We chose the TruSeq Methyl Capture EPIC® assay (Illumina) for methylome sequencing mainly because this assay targets approximately 3.3 million CpG sites and has never been tested in African ESCC patients or any other African cohort. The Illumina EPIC Truseq protocol was followed exactly according to the manufacturer's instructions without any modifications. The kit generates precaptured DNA libraries. The precapture libraries were hybridised to the EPIC oligos, purified by capture with streptavidin beads and washed twice to remove nonspecific binding. After the last elution, hybridised products were subjected to bisulfite conversion with the reagent provided in the TruSeq Methyl Capture EPIC Library Prep Kit. The bisulfite-treated libraries were PCR-amplified for 11 cycles with Kapa HiFi HotStart Uracil + polymerase© (Kapa Biosystems). Libraries were clustered on a V3 paired-end read flow cell and sequenced for 100 cycles (PE100) on an Illumina HiSeq 2500. RNA sequencing was performed on the same tumour normal matched samples using the mRNA stranded prep assay (Illumina®).

A large component of this project was invested in bioinformatics analysis. NGS data were analysed using several command line and R software packages. The commands used for each step of this entire project can be found in Supplementary document 2.1. A summary of some of the tools used for QC will be described below.

2.4 FastQC

After DNA is sequenced on a sequencer, sequencing reads are stored in the FASTQ format, which contains four lines of text: (1) read name identifiers (2) nucleotide sequences, (3) line begins with a '+' character and is optionally followed by the same sequence identifier, and (4) base Phred quality score for each nucleotide in that read (Cock *et al.*, 2010). FastQC (Andrews, 2010) is a tool that is used to calculate some

simple QC checks on raw NGS data. Input files for FastQC include fastq, BAM or SAM files. FastQC provides a quick overview and summary in the form of graphs and tables to assess NGS data. The user will have a quick impression as to whether there are any problems before doing any further downstream analysis by looking at the FasQ file. One of the main features of FastQC is the per base quality graph, which provides the user with a visual image of the distribution of sequencing quality per base position. It is generally accepted that a Phred score of ≥30 is acceptable because this indicates that the base calling is 99.9% accurate (with a 0.1% chance of error) (Cock, 2010). QC of all sequencing reads (methylome and transcriptome) was performed using FastQC version 0.11.841 (Figure 2-1).

2.5 Trimming

Typical NGS data have poor-quality bases (Phred scores ≤ 30) at the ends of reads. Adapter contamination will affect read mapping, resulting in low mapping efficiencies. These bases are removed in a procedure referred to as “trimming”. FASTQ files were first trimmed using Trimmomatic© version 1.2.14. During trimming of low-quality bases with a Phred score ≤ 30, biased read positions, indexes and adapters were removed from each read. FastQC was run again on the trimmed sequences to verify QC (Figure 2- 1). The same sequencing quality distribution was observed for all other samples.

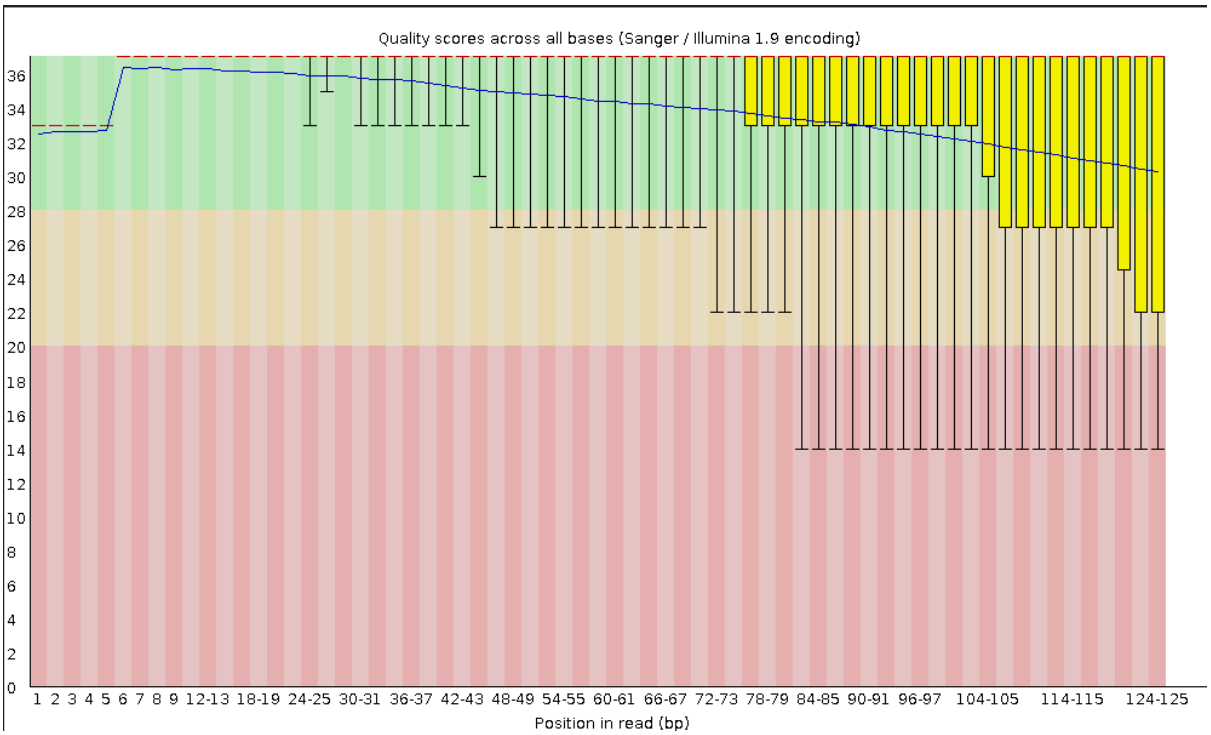


Figure 2- 1: Sequencing quality distribution for sample (2 N). The base position is on the horizontal axis, and the base quality is on the vertical axis. Green represents good quality bases with Phred scores ≥28, orange indicates reasonable quality and red indicates poor quality (Phred scores ≥20).

Primary Analysis and QC for Methylome Sequencing

2.6 Alignment and Sequencing Depth of Coverage

Following trimming, quality paired-end reads were uploaded to Bismark (v0.12.2) for read mapping and alignment to the reference human genome (hg38) using Bowtie2 (v2.2.2). Bismark performs alignment against the reference human genome (hg38) using Bowtie2 (v2.2.2). Bismark runs four alignments simultaneously using the forward and reverse strands. Bisulfite reads (C to T conversions) on the forward and reverse strands are aligned to the converted human reference genome (hg38). Bismark converts all cytosines into thymines in the reference sequence and in the bisulfite-treated samples reads. The reads from bisulfite-treated samples are then aligned to the modified reference sequence to find the positions of the methylated (unconverted) cytosines (Krueger & Andrews, 2011). After alignment, Bismark removes any sequencing duplicates and generates an output file with features that include total read pairs, mapping efficiency, number of CpGs sequenced and average coverage. These sequencing statistics are shown in Table 2- 1.

Table 2- 1: Main sequencing statistics generated from methylome sequencing

Sample name	Total read pairs	Mapping efficiency	CpGs detected	CpGs cov>10	Average coverage
N1	33,444,410	85.1%	9,534,968	3,410,181	13.66
T1	21,536,250	0.00%	1,296	0	1
N2	29,587,814	85.5%	9,087,218	3,207,347	12.6
T2	72,202,632	84.9%	12,931,749	4,542,930	21.52
N3	40,222,539	84.7%	10,310,601	3,655,200	14.74
T3	46,550,417	84.8%	10,969,304	3,851,976	15.97
N4	47,033,141	85.4%	10,851,745	3,868,987	16.33
T4	51,602,720	84.5%	11,376,527	4,036,578	17.25
N5	33,598,500	84.6%	9,595,239	3,393,338	13.36
T5	64,255,092	85.3%	12,382,829	4,319,891	19.51
N6	33,529,095	84.8%	9,440,060	3,350,171	13.45
T6	43,022,311	83.9%	10,835,424	3,799,990	14.94
N7	43,959,976	72.1%	9,505,316	3,585,641	14.55
T7	31,603,965	71%	7,815,367	2,902,765	12.7
N8	44,209,023	72.6%	9,385,976	3,568,877	14.88
T8	62,847,868	71.1%	10,862,501	4,146,899	17.63
N9	22,713,217	77.8%	6,151,854	2,321,961	12.25
T9	63,905,688	80.7%	8,862,278	4,225,536	22.93
N10	34,813,903	81%	9,149,865	2,939,595	11.54
T10	53,504,806	78.6%	10,702,528	3,786,930	14.95
N11	14,219,520	73.05	5,004,653	1,460,226	9.08
T11	21,352,844	73.65	5,931,064	2,228,671	11.9
N12	35,074,592	66.32	9,048,496	3,023,615	11.75
T12	41,348,037	67.05	9,511,438	3,300,804	13.13

• Total read pairs: Number of read pairs sequenced.

• Mapping efficiency: percentage of uniquely aligned read pairs with resp….ect to total sequenced read pairs.

• CpGs detected: number of CpGs called by the methylation extraction script.

• CpGs cov > 10: number of detected CpGs with more than 10 reads mapped.

• Average coverage - average number of reads mapping to CpG positions.

Mapping efficiency refers to the percentage of reads that can be mapped to the reference genome. Samples should have an acceptable mapping efficiency, which usually ranges between 40 and 60% (Doherty and Couldrey, 2014). All of our samples except T1 had a mapping efficiency >66%. Depth of coverage refers to the number of times a nucleotide is read during sequencing. For methylation calling, it is recommended to have a depth of coverage of at least 10x (Ziller, 2015). From our data, 22/24 samples generated a coverage > 10x, one sample generated a depth of

coverage >9, and one sample (T1) failed with a low depth of coverage of 1x. Consequently, sample T1 with low sequencing depth and poor mapping efficiency was excluded from further downstream analysis.

2.7 DNA Methylation Calling

After suitable alignments were generated, the methylation status for individual cytosines was collated into a coverage file (COV) using the Bismark methylation extractor. The COV file contains the methylation status and genomic position (hg38) for each cytosine. The methylation percentage of each cytosine can be calculated by using the number of methylated and unmethylated cytosines at a particular CpG site/s for each sample. For example, if we look at the same read of 125 bases in a normal and tumour sample across one read (Figure 2.2), all ten CpG sites shown in red are methylated, therefore this is an indication that this region is 100% methylated. In contrast, a read from another sample across the same window could have 5/10 of the same CG sites unmethylated (shown in green), indicating that 50% of this region is methylated. These 5 CpG sites that are unmethylated were converted during bisulfite conversion and therefore sequenced as thymine. The CpG data set was also filtered for low coverage, and CpGs with coverage >10x in all windows were discarded.

Sample 1 (Normal Tissue):

TCCTGCCTCAGCCTCCCGGGTAGCTGGGACTACAGGCGCGCGCCGCCACGCCTGGCT

AATTTTTTATATTTTAGTAGAGACGGCGTTTCACCATGTTGGCCAGGATGGTCTCGAT

CTGTTGACGA

Sample 2 (Tumour Tissue):

TCCTGCCTCAGCCTCCCGGGTAGCTGGGACTACAGGCGCGCGCCGCCATGCCTGGCT

AATTTTTTATATTTTAGTAGAGATGGTGTTTCACCATGTTGGCCAGGATGGTCTTGATC

TGTTGATGA

Figure 2- 2: An example of a sequencing reads taken from a matched normal and tumour pairsNormal (sample 1)/tumour (sample 2) pair. Sample 1 has 10/10 CG sites methylated, and sample 2 has 5/10 CG sites methylated at the same positions.
CG sites are shown in red. CT sites are shown in green.

The methylation status of each cytosine can exist in three different sequence contexts: CpG, CHG or CHH (where H is A, C, or T) and is also strand specific: Original Top strand (OT); Complementary To Original Top strand (CTOT); Original Bottom strand (OB) and Complementary To Original Bottom strand (CTOB). Raw data in the form of methylation calls from all four strands are included in the COV file for each sample (Supplementary folder 2.1).

2.8 Methylkit: QC

Additional QC and tertiary analysis of methylome data was generated with MethylKit (Akalin *et al.*, 2012). MethylKit is an R package used to assess QC, analysis and annotation of methylome data. The COV files generated by Bismark for each sample were used as input for MethylKit. It is always recommended to perform some QC analysis to avoid any biases that may affect subsequent downstream analyses. Typical QC after methylation sequencing includes checking for percentage methylation, CpG coverage distribution, adapter contamination and low bisulfite conversion, which can result in an erroneous methylation status of cytosine (Singer, 2019).

2.9 Bisulfite Conversion and PCR Bias QC

The percent methylation distribution plots show the frequency of reads with different methylation percentages. A typical plot should have a bimodal shape showing two

peaks at the ends. This reflects a high number of reads with both low and high methylation and a lower number of reads with intermediate percentages of methylation (Akalin *et al.*, 2012), ie, a CpG site should be fully methylated (100%) or unmethylated (0%). Methylation distribution plots for one of our tumour normal pairs are shown in Figure 2- 3. The same methylation distribution was observed in all our other samples.

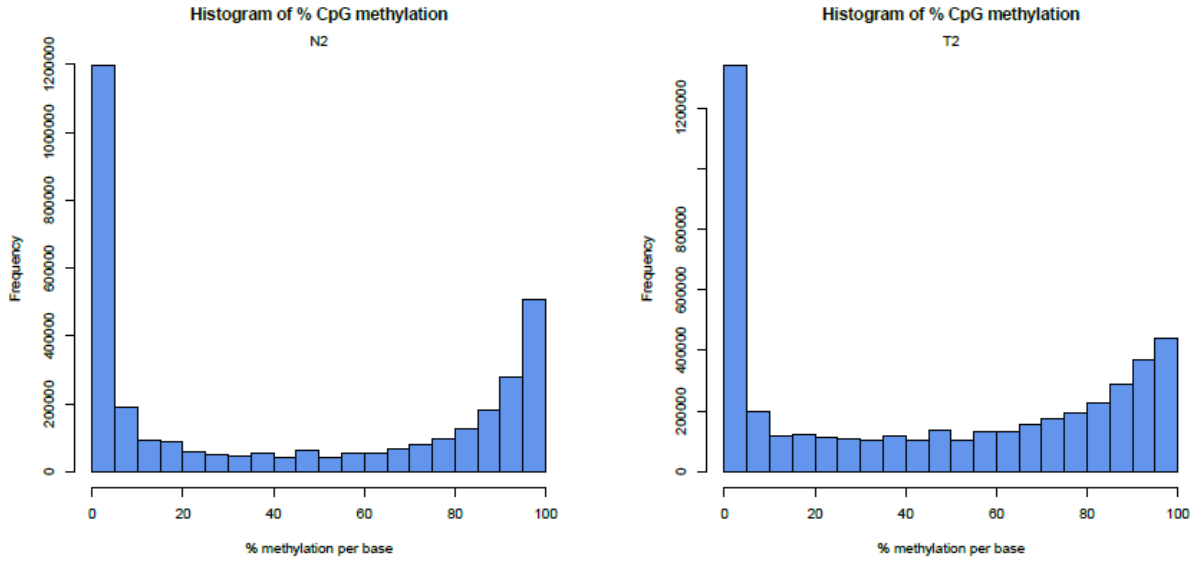


Figure 2- 3: Histogram of percent methylation of cytosine for samples N2 and T2. X-axis: percentage of methylation. Y-axis: frequency

The read coverage distribution is another important metric that would provide evidence if our sequencing suffered from PCR duplication, also known as clonal reads. If there was PCR bias in our samples, some reads will be asymmetrically amplified, which would skew methylation calls for those regions. A high degree of PCR duplication bias would display a read coverage distribution with a secondary peak on the right side (Akalin *et al.*, 2012). Reads from all samples had an equal coverage distribution, with no evidence of a second peak on the right-hand side. Distribution plots for sample 2 (tumour/normal) are shown in Figure 2- 4.

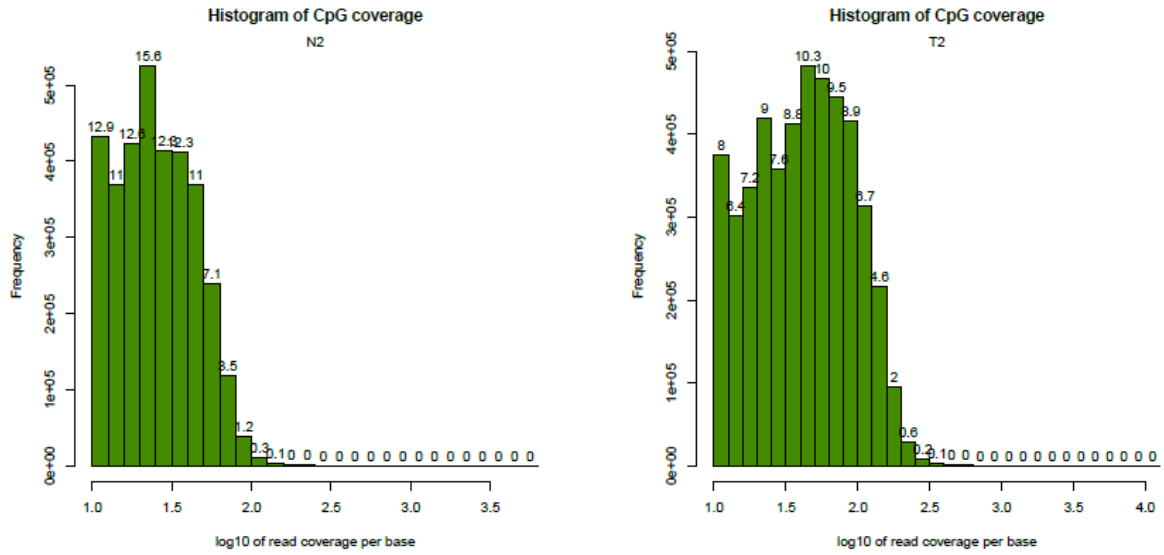


Figure 2- 4: CpG coverage distribution for samples N2 and T2. X-axis: log 10 of read coverage, y-axis: Frequency

The average conversion rate detects the efficiency of the bisulfite treatment when unmethylated cytosines are successfully converted. It is recommended that this number is ≥ 99% (Singer, 2019). The bisulfite conversion rate for all our samples ranged between 99.7% and 99.9%.

2.10 Secondary Analysis of DMRs Between Tumour Normal Samples

Figures and analysis generated with MethylKit included Distribution of differentially methylated cytosines (DMCs), Volcano plot, Principal component analysis (PCA) analysis and distribution of DMRs across gene features. Pathway enrichment analysis across biological processes (BPs), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome was performed in the R environment. Correlation analysis between DNA methylation and transcriptome data was also performed in the R environment. Details of all tertiary analyses for methylome data are described in chapter 4.

2.11 RNA Sequencing

2.11.1 Primary Analysis and Quality Control for Transcriptome Data

Similar to methylation sequencing data, transcriptome data were filtered through a series of filtration methods to obtain high-quality sequencing data for subsequent analysis. The QC of RNA sequencing can be broadly divided into three steps: 1 - raw read data (FASTQ), 2 - alignment and 3 - gene expression. As with DNA methylation sequencing data, the base quality distribution for RNA sequencing data was also visualised with FastQC (v0.11.841), and adapters and reads that had bases with a Phred quality score ≤ 30 were removed with Trim Galore! version 0.4.142. The Q30 score for all our samples ranged between 94 and 97.88%.

2.11.2 Distribution of A, T, G, and C Bases

FastQC was also used to identify the proportion of AT and GC content. According to the latest human genome build (hg38), the GC content is 48.9% for coding RNA (Sheng *et al.*, 2017). Thus, the proportion of AT and GC should be equivalent at each sequencing cycle during the entire sequencing procedure. However, due to primer amplification bias, the first 6 to 7 nucleotides will fluctuate, which is expected in most sequencing experiments. The distribution of GC content is shown in Figure 2- 5. All bases have a similar content (~25%) after the first few cycles of sequencing.

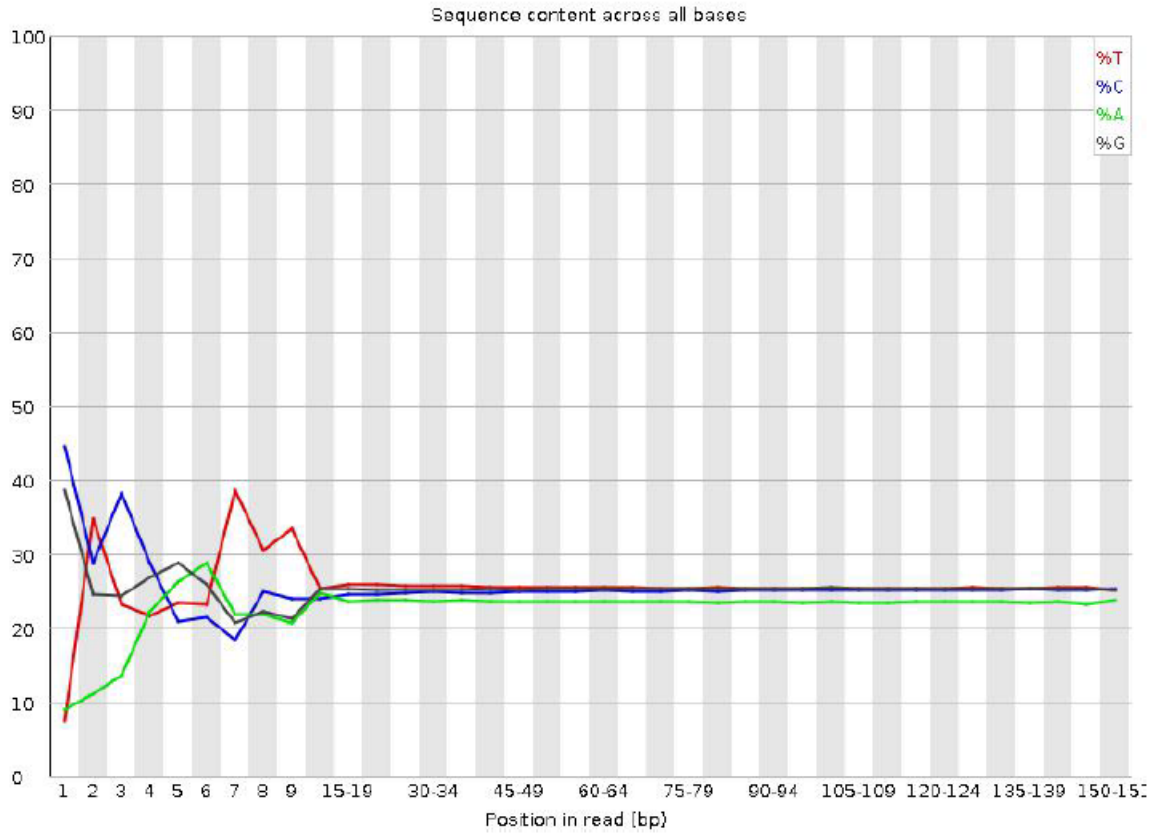


Figure 2-5: Distribution of A, T, G, and C contents for sample 2 N. The horizontal axis represents the position of each base, and the vertical axis represents the percentage of each base.

Raw reads were filtered to remove reads containing adapters or reads of low-quality (Phred quality score ≤ 30) so that downstream analyses only proceeded with clean reads. The filtered clean reads were mapped to the reference genome with HISAT2 (Kim *et al.*, 2019). Mapped reads refer to reads that are mapped to exons, introns or intergenic regions. Regions mapping to the exon should be most abundantly present. The mapping rate of each sample refers to the percentage of clean reads that can be mapped to the reference genome relative to the total number of clean reads (Dobin and Gingeras, 2015). The mapping rate is an important variable because it indicates sequencing accuracy and the presence or absence of contaminating DNA (Conesa *et al.*, 2016). Typically, we expect a mapping rate of 70-90% from human RNA sequencing experiments (Conesa *et al.*, 2016). We observed a high mapping rate of 94 – 98% across all of our samples. The number of sequenced reads per sample is also an important factor for quantification of transcripts when carrying out gene expression studies. Studies have shown that as few as 5 million reads are sufficient to quantify medium to highly expressed genes (Sims *et al.*, 2014), while other studies suggest 20 million reads (Conesa *et al.*, 2016) and sometimes up to 100 million reads

to precisely quantify genes that have low expression levels (Sims *et al.*, 2014). In this study, we obtained an average of 19 million reads per sample. Raw data for RNA sequencing can be downloaded in a link provided in Supplementary folder 2.2. After QC, the next step was to quantitate transcripts for each gene and determine differentially expressed genes (DEGs) between our normal and tumour groups. Pathway analysis for our DEGs was carried out with Gene Set Enrichment Analysis (GSEA). The methodology used for these steps and their results have been described and presented in Chapter 5. A structural outline of chapters 3-5 and all tools used are shown in Figure 2- 6.

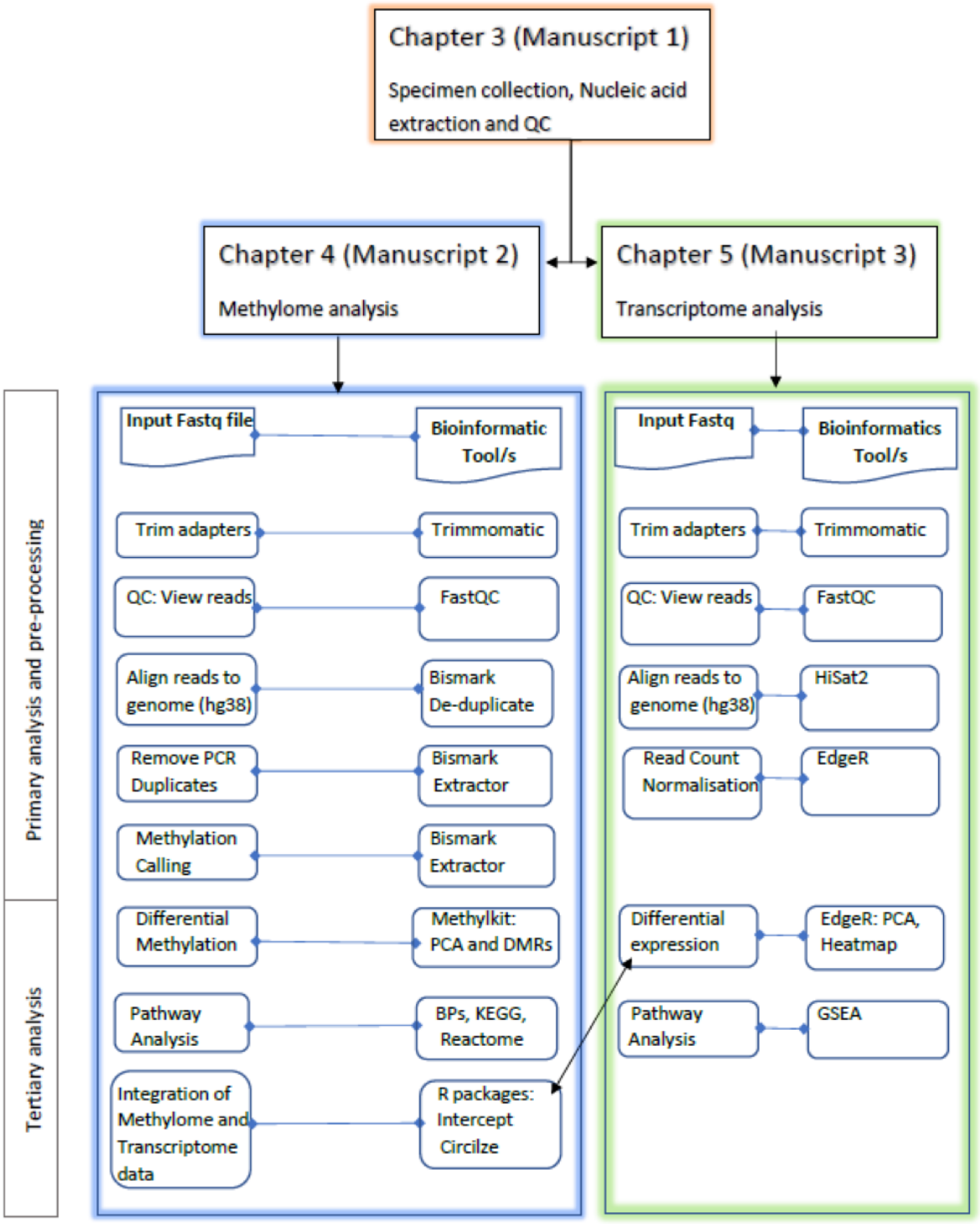


Figure 2-6: Structural outline of chapters 3-5

Chapter 3 (Manuscript 1):

Abstract

Although infectious diseases continue to present a major healthcare problem in Africa, the incidence of cancer is increasing rapidly on the African continent, which merits an increased investment in cancer research in low- to medium-resource settings. Oesophageal squamous cell carcinoma (OSCC) has a high incidence in Eastern and Southern Africa, with late clinical presentation and a very poor prognosis. There is limited research on the molecular pathology of this cancer in Africa, partly as a result of a lack of infrastructure for biobanking and sample processing in many African countries. The aim of this study was to establish a practical and robust workflow to collect, store and process oesophageal cancer samples such that both the tissue architecture and quality of the samples would be preserved and suitable for future genomic research. We developed a workflow that allows storage of fresh biopsy tissue in sterile Eppendorf tubes containing RNAlater, an efficient RNAse inhibitor. We collected 142 OSCC biopsy samples and showed that storage in RNAlater for up to 18 months did not alter tissue morphology, thus allowing histologic assessment by experienced pathologists and determination of tumour content in each biopsied sample. DNA and RNA extracted from tissue samples were assessed for purity, molecular size and yield. The quantity and quality of nucleic acids obtained were suitable for genomic applications, and whole-exome sequencing of DNA from blood and tumour tissues produced sequence data with a high proportion of both usable reads and correct base calling. We conclude that this workflow may be applicable to a wide range of malignancies for future genomic research in low-resource settings.

Introduction

Cancer is a growing health care problem worldwide, particularly in Africa, which is predicted to be the continent with the fastest increase in cancer incidence over the next decade (Sylla and Wild, 2012; Global Burden of Disease Study 2013 Collaborators, 2015). Despite this, the focus of research in Africa has been mainly on communicable diseases such as acquired human immune deficiency syndrome and tuberculosis (Hofman *et al.*, 2009; Adedokun, Olopade and Olopade, 2016). Noncommunicable diseases in Africa have enjoyed less attention, although this appears to be changing (Olesen and Parker, 2012; Kane *et al.*, 2017). Health care systems in Africa are largely geared toward addressing the impact of communicable diseases, leaving them ill-

equipped to deal with the ever-increasing burden of chronic diseases (de-Graft Aikins *et al.*, 2010). Even among publications on noncommunicable diseases, cancer-related research is less prolific than cardiovascular-related research (Hofman *et al.*, 2006).

Oesophageal cancer, particularly OSCC, is considered to be endemic in certain parts of Africa (Parkin *et al.*, 2014; Ferlay *et al.*, 2015). A corridor that extends along the eastern side of Africa from Ethiopia to South Africa is known as a high incidence area of esophageal cancer, with OSCC as the predominant subtype (Schaafsma *et al.*, 2015). Despite the high prevalence, there is limited literature on clinical studies of OSCC and of genomic profiles of OSCC tumours in Africa (Loots *et al.*, 2017; Brown, Stepien and Willem, 2020; Liu *et al.*, 2016). In contrast, significant progress has been made in genomic research on OSCC in non-African countries, particularly in China (Lin, Wang and Koeffler, 2018). The success of these studies is based on (i) large sample sizes, which have the power to identify potential driver genes, i.e., genes that have significant rates of somatic mutations in tumours, and (ii) application of high throughput genomic technology such as whole-exome, whole-genome and RNA sequencing of tumours.

SA and other African countries, on the other hand, face major challenges that stem largely from resource constraints. These challenges include a lack of human resources dedicated to research, inadequate infrastructure for biobanking, and most importantly, very limited funding (Vaught *et al.*, 2014). The establishment of a biobank with nationally accepted, validated protocols and QC measures is still an emerging concept in Africa, and as a result, there is a lack of protocols for biobanking that are validated in this setting (Abayomi *et al.*, 2013; Soo *et al.*, 2017). Following protocols established by facilities in first-world countries, while desirable, may not be feasible in an African setting. There is therefore a need to develop research protocols that are functional but also feasible to follow in a resource-constrained environment. Establishing robust workflows for biobanking of oesophageal cancer tissue samples in South Africa would provide essential tools to facilitate good quality translational research and contribute to improving the quality of care for patients afflicted with this devastating disease. It could also serve as a template for the collection and processing of tissue samples for genomic research in other African cancers.

We set out to assess the feasibility of performing genomic studies in patients with oesophageal cancer from centres with significant resource constraints in two provinces in South Africa by establishing procedures for collection, storage, transport and

processing of tumour samples collected from patients with squamous cell carcinoma of the oesophagus. We also assessed the quality of the genetic material obtained from tissue samples with the aim of developing a workflow for sampling and processing that is feasible in this setting.

Materials and Methods

Patient Recruitment

Patients were recruited from two centres in two provinces of SA, namely, CMJAH in Gauteng, which is a high resource setting, and GH in KwaZulu-Natal, a low-resource setting. CMJAH offers a range of tertiary and specialised services and serves patients from neighbouring provinces. OSCC patients recruited from CMJAH mainly reside in the urban areas of Gauteng. GH is a tertiary hospital providing health care to one of three areas in KwaZulu-Natal. The area covers the Western part of KwaZulu-Natal and has a population of approximately three million, of which approximately two-thirds are from rural areas (Caldwell *et al.*, 2015; Census SA, 2011). Potential OSCC patients arriving for endoscopy at CMJAH and at GH were eligible for recruitment into this study. Only patients with histologically confirmed OSCC were then enrolled. The purpose of the study was explained to each patient in their native language. Informed consent was obtained from each patient who then completed a study questionnaire before their consultation with the surgeons. A research nurse facilitated patient enrolment, consenting and compiling the study questionnaire at CMJAH, while the study clinician (LF) and research assistant facilitated the same processes at GH. Ethical clearance for this study was obtained from the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Wits) (Certificate number M170871) and from the Biomedical Research Ethics Committee of the UKZN (Certificate number BF270/15).

Sample Collection and Storage

The International Agency for Research and Cancer (IARC) protocol was used as a guideline for collecting and processing OSCC biospecimens at CMJAH and GH (Mendy *et al.*, 2017).

Figure 3- 1 illustrates the procedure used to collect, process and store OSCC biopsies from CMJAH and GH. Routine biopsies of oesophageal tissue were taken from patients suspected to have OSCC for diagnostic purposes. For this study, additional tissue specimen biopsies were obtained during the routine procedure. At CMJAH, an additional biopsy specimen in the oesophagus of each suspected tumour was obtained. At GH 2-3, additional biopsy specimens from the suspected tumour were obtained.

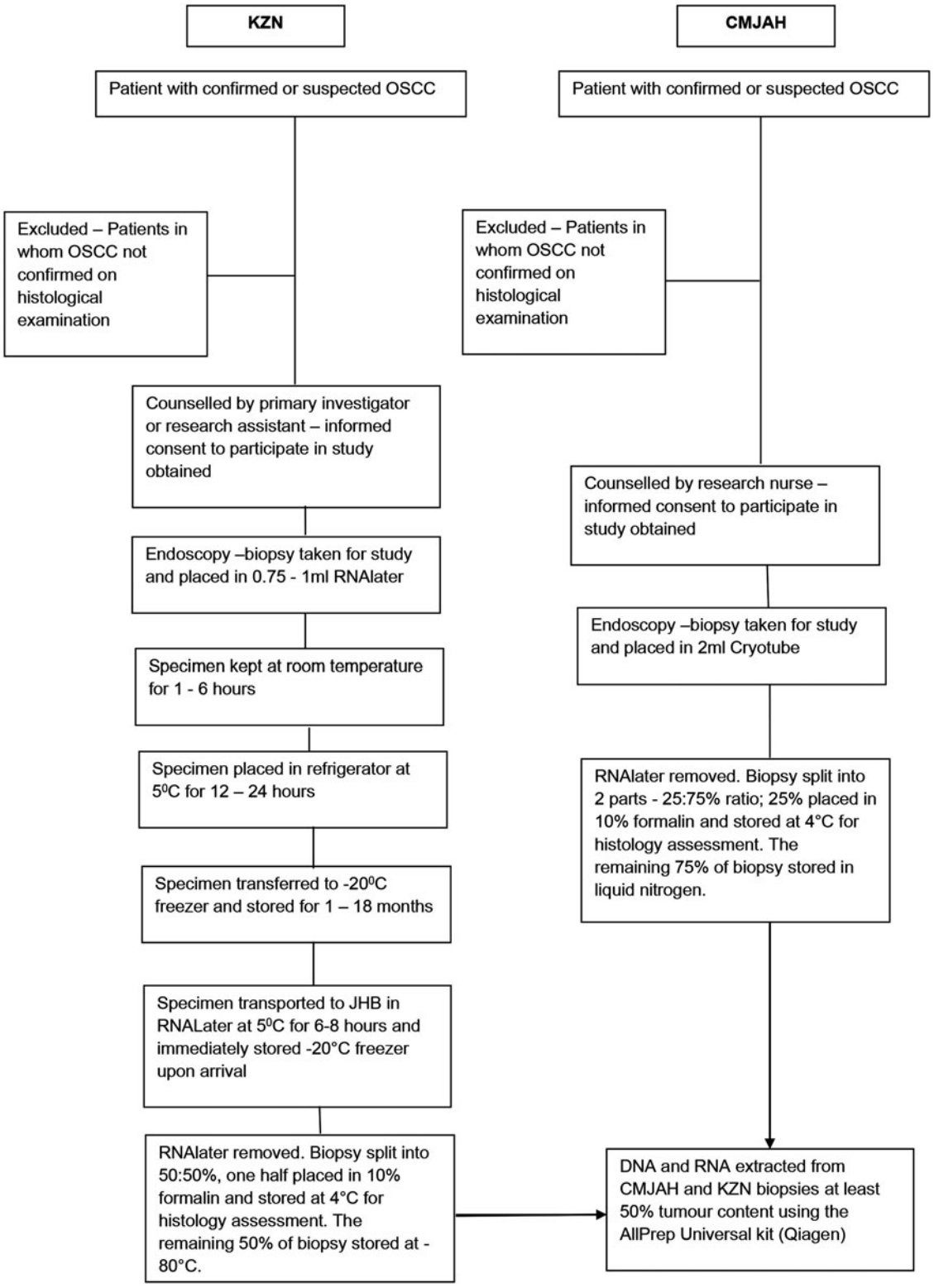


Figure 3- 1: Protocols followed by the two sites for collection, storage, and processing of samples

Histology preparation for FFPE and frozen sections

Biopsy specimens from the CMJAH samples were placed in a 2 ml Cryotube immediately after extraction, and the Cryotube was then immersed in a vacuum flask containing liquid nitrogen. The biopsy was then placed in optimal cutting temperature (OCT) compound (Tissue-Tek®) and stored at -80°C. Biopsies were transported in batches of 10 in liquid nitrogen to a histopathology laboratory off-site for histological processing. Each tissue specimen was divided into two pieces (25%:75%), with the 25% portion being cut from the OCT-embedded frozen block and used for histology. The remaining biopsy (75% of the OCT-embedded frozen block) was stored at -80°C until nucleic acid extraction.

A different protocol was used for the GH samples. Tissue specimens were placed immediately into 0.75-1 ml of RNAlater (Qiagen, USA), kept at room temperature for 30 minutes to six hours and then stored overnight in RNAlater at 5°C before being transferred to a -20°C freezer. The specimens were stored at -20°C for one to eighteen months and then batched and transported at 5°C (temperature controlled) to CMJAH. Upon arrival at CMJAH, the specimens were stored at -20°C. Before processing, the specimens were removed from the -20°C freezer and kept at room temperature for 24 hours, after which the RNAlater was removed. Each biopsy specimen was cut into two equal (50%:50%) portions under sterile conditions. One-half of the specimens were placed in 10% buffered formalin and batched for transportation to an off-site histopathology laboratory for histological processing. The other half was immediately placed in a 1.5 ml Eppendorf tube and stored at -80°C for nucleic acid extraction.

For histological processing, all tissue specimens fixed in formalin (at 6-8°C) were washed repeatedly in different levels of ethanol of increasing concentrations from 70% to 100%. Xylene was then added to each specimen, followed by paraffin wax at 60°C. The tissue specimen was then cooled to allow solidification. The tissue was embedded into a cassette, and serial sections were cut at 5 µm, placed on a single glass slide and stained with hematoxylin and eosin (H&E). The OSCC tissue specimens were sectioned in a Leica microtome cryostat using optimal cutting temperature compound (OCT compound) to embed the tissue samples. Four sections of 10 µm each were mounted onto a single glass slide and stained with H&E.

All tissue sections were then examined microscopically by two experienced histopathologists. The tissue sections were evaluated for the following parameters:

tissue type, tumour percentage, necrosis, inflammation, ulceration and the presence or absence of overlying surface epithelial dysplasia. The percentage of surface epithelial dysplasia was quantified. The tumour was assessed for the degree of differentiation and the presence or absence of lymphovascular invasion. A consensus assessment was rendered for each case.

All biopsy specimens from both sites were stored in an upright position. All refrigeration equipment at the CMJAH site was monitored daily with an alarm-linked notification system. Back-up refrigeration equipment was available if primary refrigeration equipment failure occurred. At GH, there was no alarm system or back-up refrigeration equipment, but the refrigeration equipment was connected to a back-up generator and uninterrupted power supply device.

Tissue DNA and RNA isolation

Tissue DNA (tsDNA) and tissue RNA (tsRNA) were extracted from a subset of tissue specimens with substantial tumour content. All tissue specimens were extracted using the Qiagen AllPrep® Universal kit (Qiagen, USA) according to the manufacturer's protocol. This kit was chosen for its ability to extract high-quality nucleic acids, which are needed for molecular applications such as NGS. All remaining tissue biopsies were used for tsDNA and tsRNA extraction. Tissue from each specimen was added to 350 µl Qiagen buffer RLT (Qiagen, USA) and lysed at 30 mHz for 45 seconds using TissueLyser II (Qiagen®). The lysate was pipetted onto an AllPrep® DNA spin column (Qiagen, USA) for tsDNA binding. The column was centrifuged at 8000 x RCF for 1 minute to collect the flow-through tsRNA in the collection tube. Since RNA is more sensitive to degradation, tsRNA was extracted before tsDNA. RNA flow-through was transferred to an RNeasy spin column. Both nucleic acid specimens were extracted according to the Qiagen Allprep DNA/RNA/miRNA Universal kit (Qiagen, USA). The tsDNA was eluted in 35 µl of elution buffer, and the tsRNA was eluted in 60 µl of buffer.

Assessment of tsDNA and tsRNA quality

Nucleic acid quantity and quality were assessed with a Qubit (Thermo Scientific, USA), 2100 Bioanalyzer (Agilent Technologies, USA) and Nanodrop™ 1000 (Thermo Scientific, USA) spectrophotometer. Qubit was used to determine DNA and RNA concentrations. A Nanodrop 1000 was used to determine the absorbance ratios at 260 nm and 280 nm. A 260/280 ratio of >1.8 for DNA and >2 for RNA was used to indicate good purity. An Agilent 2100 BioAnalyzer was used to determine RNA quality. RNA was extracted from a total of 49 biopsies that were prioritised as having substantial proportions of tumour or dysplasia content. RNA quality was measured using the LabChip GX RNA Assay (Perkin Elmer, USA) according to the manufacturer's instructions using the RNA integrity number equivalent (RINe) score, which has a scale of 1-10, with 10 being the highest quality (Matsubara *et al.*, 2020). tsRNA and tsDNA samples were stored at -80°C.

Whole-Exome Sequencing

Exome sequencing of tumour DNA with matched germline DNA was performed on samples from six OSCC patients. A total amount of 500 ng of DNA from each sample was used as the starting material. DNA was fragmented using the Covaris shearing system (Covaris, USA). Fragments that were generated ranged between 180-280 bp. Exonucleases removed the 5' and 3' ends to ensure that each fragment had a blunt end. DNA was end-repaired after adenylation of the 3' ends. Adapter molecules were ligated to both ends to ensure that DNA fragments were enriched in a PCR. The exonic region of each sample was captured, enriched for PCR and indexed to prepare for hybridisation using SureSelect Human All Exon V6 (Agilent Technologies, CA, USA). NGS of the library was performed on an Illumina HiSeq 2500 sequencer (Novogene, UK). The quality of the sequence data was assessed by calculation of the % of usable reads and the Phred Q score. A sequenced fragment is considered "usable" if it maps uniquely to the genome and remains after removing PCR duplicates. The Phred Q score is the quality score of a base, which is an integer value representing the estimated probability of an error. If P is the error probability, then $P = 10^{-Q/10}$ and $Q = -10 \log_{10}(P)$.

Personnel resources

Gray's Hospital, KwaZulu-Natal: Patients consented, and biopsies were collected and stored by one of the authors (LF) or one of two additional endoscopists (see Acknowledgements). Biopsies for research were collected at the same time as diagnostic biopsies and placed in RNAlater in the endoscopy suite. Samples were then placed in a refrigerator overnight and in a -20°C freezer the next morning by LF. Thereafter, the specimens were transported in batches to the laboratory in Gauteng.

CMJAH, Gauteng: Patients consented to one research nurse, and biopsies were collected by one of three endoscopists (see Acknowledgements). Samples were processed for histology and DNA/RNA extraction and subjected to QC by a medical laboratory scientist (MM). Slides were assessed for tissue morphology and tumour content by two consultant histopathologists (RW and CAW).

Statistical analysis

The Kruskal–Wallis nonparametric test was used to assess differences in tsDNA and tsRNA quality (A260/280) between study sites and between tissue storage media (RNAlater vs. liquid nitrogen). Spearman's rank-sum test was used to assess whether there was any correlation between the number of days of sample storage and the quality of the tsDNA and tsRNA. The Mann–Whitney U test was used to assess differences in RNA quality scores by the sample storage method.

Results

A total of 142 patients were recruited into the study from the two study sites. Histological examination was performed on tissue samples from all patients, and nucleic acids (DNA and RNA) were extracted from a total of 49 samples with substantial tumour or dysplastic tissue content as part of our ongoing study on the genomic analysis of African oesophageal cancer. The total number of biopsies collected at each site with the storage method and histology profile is shown in Table 3- 1. Important differences between protocols followed at the two different sites included storage time, temperature and storage medium. Samples at the GH were stored for long periods (several months) in RNAlater before downstream processing,

while biopsies collected at the CMJAH were stored for shorter periods (minutes to hours) in liquid nitrogen before downstream processing.

Table 3- 1: Number of Biopsies Collected at Each Site with Histology Profile, Nucleic Acid Yield, and Quality for Tissue DNA and Tissue RNA Isolated from Tissue Specimens

Description	CMJAH-JHB	GH-KZN
Number of biopsies collected at each site	50	92
Stored as frozen section	26	0
Stored in RNAlater	24	92
Squamous Cell Carcinoma	16 (32%)	60 (63%)
Dysplasia	25 (50%)	18 (19%)
Dysplasia and Squamous Cell Carcinoma	1 (2%)	5 (5%)
Other	8 (16%)	9 (13%)
Number of biopsies that were used for nucleic acid extraction	37	12
Tissue size (cm)	0.3 – 0.5	0.5 – 1
Mean tsDNA yield (ug)	4.5 µg	9.2 µg
Mean A260/280 tsDNA (std.dev)	1.91 (0.13)	1.96 (0.11)
Mean tsRNA yield (ug)	5.7 µg	11.5 µg
Mean A260/280 tsRNA (std.dev)	2.02 (0.09)	2.03 (0.04)
Mean tissue storage time (days) (std.dev)	92.8 (64.2)	650 (84.4)

Histological analysis of the tumour content of all biopsy samples was obtained and evaluated as described in the Methods. The quality of cellular morphology from all 142 OSCC specimens from CMJAH and GH and both the frozen section and RNAlater protocols was sufficient to allow assessment of all the required parameters, including tissue type, tumour percentage and the percentage of surface epithelial dysplasia. As shown in Table 3- 1, the GH protocol (RNAlater) produced a higher percentage of samples with histology positive for squamous cell carcinoma compared to the protocol followed at CMJAH (Frozen section/Liquid nitrogen). A substantial proportion of the research biopsy samples from CMJAH (50%) were assessed as dysplasia rather than invasive carcinoma, although all patients had a clinical diagnosis of OSCC. Representative photomicrographs of biopsy tissues from both centres are shown in

Figure 3- 2. The panels on the left half of the image (A, C, E, G) show a squamous cell carcinoma from a frozen section tissue biopsy, while the panels on the right half (B, D, F, H) show, at the same set of magnifications, a squamous cell carcinoma that was placed in RNAlater. Extensive freeze artefact is apparent in the images on the left panel, which restrict areas suitable for histopathological assessment. Freeze artefacts with distortion of cellular detail are best shown in image E (3 thin arrows). Freeze artefacts cause blurring of the epithelial/stromal interface, making assessment of invasion difficult. Additionally, there is no clarity of nuclear detail, which is critical for a definitive diagnosis of malignancy, and identification of mitoses is problematic. The panels on the right show islands of tumour cells with clear cellular morphology and well-formed keratin pearls (2 thin arrows in H).

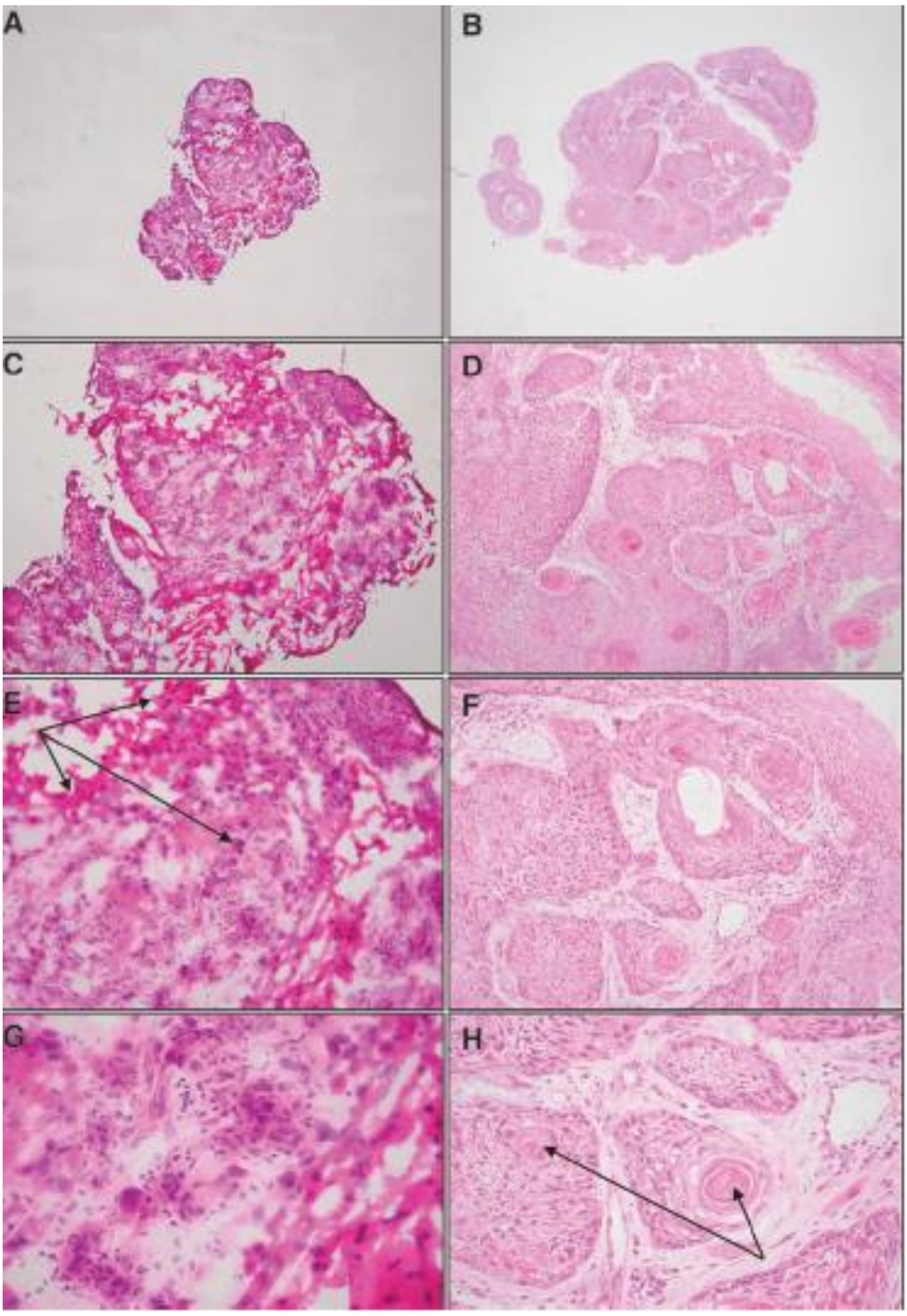


Figure 3- 2: Composite photomicrograph of esophageal invasive squamous cell carcinomas (part of figure legend at increasing magnifications. Panels (A, C, E, G) squamous cell carcinoma from a frozen section tissue biopsy; Panels (B, D, F, H) squamous cell carcinoma placed in RNAlater. Extensive freeze artefacts noted in (A, C, E, G) with freeze artefacts and distortion of cellular detail best shown in image E (three thin arrows). Panels (B, D, F, H) show islands of tumour cells with clear cellular morphology and well-formed keratin pearls (two thin arrows in H). Original magnification: A and B: 40×, C and D: 100×, E and F: 200×, G and H: 400.)

Table 3- 1 shows the yield (amount) and quality of the genetic material obtained from samples from the two centres. The yield of nucleic acids obtained from the GH samples

was more than twice that of the CMJAH samples. This is likely due to the size of the biopsies obtained from GH, approximately twice the size of biopsies obtained from the CMJAH. There were no statistically significant differences between the quality (A260/280) of tsDNA ($p=0.072$) and tsRNA ($p=0.625$) from the two study sites. The tsDNA from both sites was of good quality, with mean A260/280 ratios of 1.91 and 1.96 for CMJAH and GH, respectively, and mean A260/280 ratios of 2.02 and 2.03 for tsRNA. The mean RINe scores for RNA samples for both protocols were over eight and were not significantly different, with 43 of 49 (88%) samples meeting the generally accepted metric (>7) required to produce high-quality RNA sequencing libraries.[22] An example of the gel electrophoresis of the tsDNA samples is shown in Figure 3- 3. High molecular weight tsDNA was obtained from all samples from both study sites.

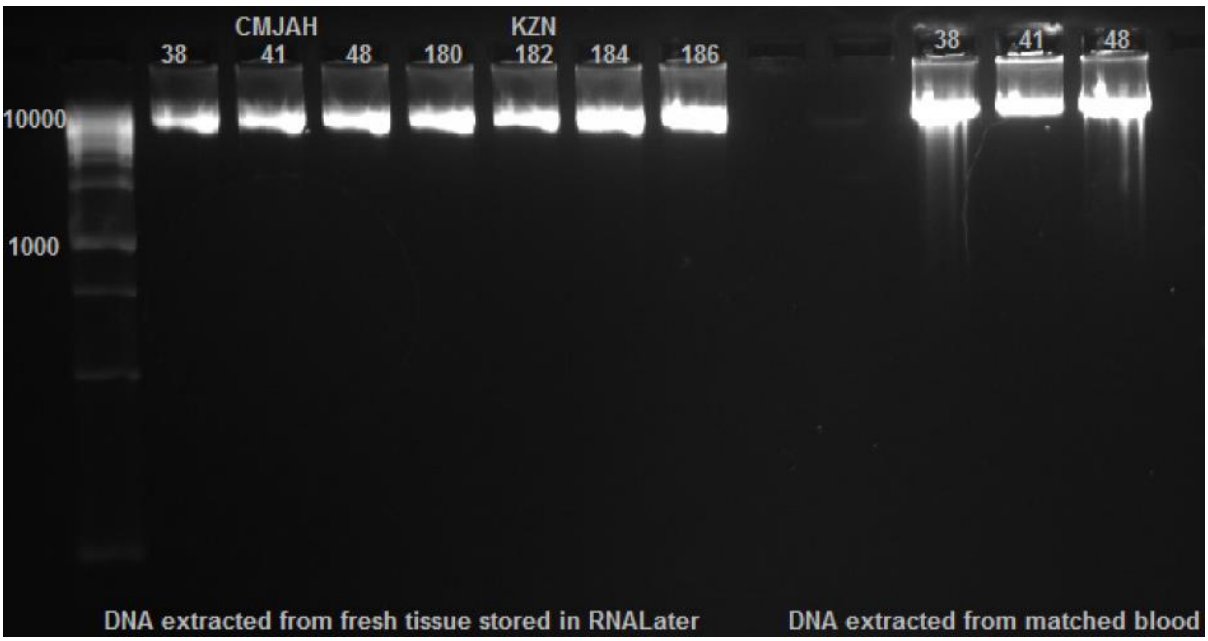


Figure 3- 3: DNA samples extracted from tumour tissues on a 1% agarose gel

DNA extracted from a subset of six tissue samples and their paired germline DNA was submitted for whole-exome sequencing to determine whether the samples were of sufficient quality for genomic studies. The QC analysis of the sequence data is shown in Table 3- 2. The average number of usable sequence reads across the six tissue samples was 97.4%. The average proportion of bases called with a Phred score of ≥ 30, which indicates a 99.9% probability that the base was called correctly, was 92.6%. These metrics indicate good quality next-generation sequence data (Ewing and Green, 1998; Ewing *et al.*, 1998), which were similar between samples sequenced from both sites. Furthermore, the quality of the sequence data from the tissue samples was very similar to that of DNA extracted from paired blood samples (data not shown).

Table 3- 2: Quality Control Statistics for DNA Exome Sequencing Data from the Tissue DNA Samples

Sample	Centre	Raw reads	Usable reads (%)	Q30 score (%)
180 Blood	GH - KZN	25 203 721	97.12	93.12
180 tsDNA		28 178 349	98.29	92.54
182 Blood		36 860 022	96.65	92.73
182 tsDNA		33 810 584	97.15	92.62
186 Blood		29 055 883	97.50	91.84
186 tsDNA		32 985 596	97.79	92.63
41 Blood	CMJAH - JHB	13 965 249	97.74	93.00
41 tsDNA		37 575 723	96.89	92.34
48 Blood		20 282 587	95.33	91.66
48 tsDNA		31 801 242	96.91	92.44
38 Blood		40 151 807	93.92	93.12
38 tsDNA		37 587 499	97.26	92.89

The personnel required to execute the two protocols are described in the Methods. The additional effort required over and above that needed for a standard diagnostic workflow is taking additional biopsies at the same time as the diagnostic biopsies (endoscopist), then either freezing the research biopsy in liquid nitrogen and storing it at -80°C or placing it in RNAlater. The freezing protocol requires a trained laboratory assistant. Preparation and analysis of slides is required for both protocols and is done by histopathologists. DNA and RNA were extracted from the tissue samples by a laboratory scientist. An estimate of the costs of reagents, personnel and equipment for the two protocols in our two institutions is shown in Table 3- 3. This shows that the running costs per sample are very similar for the two protocols. However, the frozen section protocol requires a substantial initial investment for equipment. The methods for extraction of DNA/RNA and subsequent QC of the extracted nucleic acids are identical for the two protocols.

Table 3- 3: Cost per sample breakdown for the two protocols

	CMAHJ - JHB	GH - KZN
Running cost	Frozen section	RNALater
Consumables per sample: (RNALater, cryovials for GH; Liquid nitrogen, cryovials for CMJAH)	USD 1.69	USD 3.51
Histology costs per sample	USD 27.50	USD 27.50
Reagents for DNA/RNA extraction	USD 16.50	USD 16.50
Staff time for sample processing and extraction	USD 18.00	USD 18.00
Total running cost per sample	**USD 63.69**	**USD 65.51**
Fixed costs for equipment		
Liquid nitrogen 21 L storage tank	USD 1 515.47	
Liquid nitrogen flask (Dewar)	USD 283.17	
-80C Ultra low temperature freezer	USD 5789.12	
Cryo gloves (1 pair for long term usage)	USD 806.53	
-20C Chest Freezer		USD 174.66
Total fixed costs	**USD 6 245.50**	**USD 174.66**
Sample shipment		
Dry ice shipment (100 samples)		USD 365.07

Exchange rate of 1 USD = 13.7400 ZAR was used. Notes: Sample shipment when relevant is estimated at USD 365 per batch of 100 samples.

Discussion

This study was undertaken to ascertain whether genomic research could be carried out successfully in centres and study sites with significant resource constraints. In our study, one of the centres had no on-site genetic laboratory or staff and no on-site histopathology service and relied on a refrigerator/freezer, a supply of RNAlater, specimen collection tubes and clinical staff to collect and store specimens. Despite these limitations, substantial proportions of tumour tissue could be identified in the majority of specimens on histological assessment, and genetic material (both DNA and RNA) could be extracted with acceptable quality and quantity, with the DNA being used successfully for NGS.

The first important consideration when performing genomic studies on tumour tissue is ensuring that tissue samples taken are representative of the tumour. Histopathological confirmation of tumour content in biopsy specimens is an essential step in successful biobanking (Basik *et al.*, 2013). Since wide disparity in tumour content may occur between biopsies, each specimen should be evaluated by a pathologist to confirm the presence of a substantial fraction of tumour tissue prior to using samples for costly genomic research (Basik *et al.*, 2013). This is further illustrated by our study, where tumour content varied from 15% - 100% between different samples. One of our initial concerns was that biopsies stored in RNAlater for long periods (>12 months) would lose the integrity of the tissue and thus their morphology and would not be suitable for histological analysis. However, histology results were obtained successfully from all 142 OSCC specimens from CMJAH (high resource) and GH (low-resource), indicating that oesophageal biopsies stored for up to 18 months did not affect the integrity of the biopsies, so they could still be used for histological assessment. A previous study showed that storing tissue biopsies in RNAlater or OCT compound for 24 hours did not compromise histological, immunochemical and genomic qualities (Florell *et al.*, 2001). However, another study reported that 45% of RNAlater samples had poor nuclear morphology preservation, rendering pathological evaluation challenging (Aguilar-Mahecha *et al.*, 2017). The protocol followed for samples collected at GH, where specimens were immersed in RNAlater and kept at room temperature for several hours, then at 5°C overnight before being stored at -20°C, still in RNAlater, may offer improved protection of tissue morphology (Ellis *et al.*, 2002; Basik *et al.*, 2013). Another important consideration when using RNAlater for storage is the size of the biopsy specimen since RNAlater has been shown to interfere with histological assessment in smaller samples (Basik *et al.*, 2013; Aguilar-Mahecha *et al.*, 2017).

A major difference between the two centres was the method of specimen storage, with liquid nitrogen being preferred at CMJAH, while RNAlater was used exclusively at GH. In our experience, the logistics used for storing and extracting nucleic acid from liquid nitrogen was more challenging compared to RNAlater. Liquid nitrogen must be available at the biopsy site, careful handling is needed to prevent injury to research staff, and a long term liquid nitrogen storage facility must be available to maintain frozen sections. Somewhat surprisingly, we found that the quality of the nuclear morphology was superior in the RNAlater samples compared to the samples prepared in liquid nitrogen. The use of liquid nitrogen can make histological assessment more

difficult since specimens are fragile and more difficult to section (Diaz *et al.*, 2013). In our study, we found that certain samples had freezing artefacts, which can lead to an inconclusive interpretation (Chatterjee, 2014). RNAlater is also easier to use within a busy clinical environment and a low-resource setting where clinicians would not always have immediate access to refrigerators or freezers. There was a difference in the percentage tumour content between the two centres. As shown in Table 3- 1, specimens from GH had a substantially higher rate of squamous cell carcinoma-positive cases (63%) compared to patients from CMJAH (32%). The reason for the higher tumour content from GH samples may be related to the biopsy size, since samples from GH were larger than those from CMJAH despite the same size forceps being used at both centres (Table 3- 1). Another possible reason may be the experience of the endoscopist performing the biopsy. At GH, the biopsies were collected either by the primary investigator or by one of two other endoscopists, all of whom had several years of endoscopy experience. In contrast, at CMJAH, biopsies were collected by several different clinicians and, in most instances, by trainee surgeons with less experience, who might have been less effective in deep sampling of the core of the tumour.

QC of DNA and RNA samples extracted from tumour tissue showed similar A260/280 ratios between the two centres, while the mean yield of DNA and RNA extracted from GH cases was approximately twice as high as the yields obtained from CMJAH cases. This could also be explained by the fact that biopsies from GH were larger (0.5-1 cm for KZN compared to 0.3-0.5 cm for CMJAH). As mentioned previously, this is most likely because the majority of biopsies were taken by a single very experienced endoscopist at GH (LF) in contrast to less experienced endoscopists at CMJAH. However, the quality and quantity of DNA and RNA from both sites were adequate for genomic studies such as NGS. There was a marked difference in the storage time of the tissue from the two different sites (Table 3- 1), which was necessary since samples at GH were batched for shipment to CMJAH. Our study demonstrates that despite a storage time of several months, the quality of genetic material is not compromised. This is an important consideration in centres without on-site laboratories such as GH. The absence of these facilities does not, therefore, preclude genomic research. The quality scores from the exome sequencing data are an indication that DNA extracted

from tissue samples at both sites generated high-quality sequences whether biopsies were stored in RNAlater or liquid nitrogen.

The personnel resources required for the two protocols are very similar except that a trained laboratory assistant is required for freezing samples in liquid nitrogen, whereas the biopsies can be placed in tubes containing RNAlater by the endoscopist. Additionally, although the use of liquid nitrogen has a slightly lower consumable cost than RNAlater, it requires a substantial initial investment for equipment. The only equipment required for storage in RNAlater by GH was a -20°C freezer; thus, the RNAlater protocol has a significant advantage in settings where there is a lack of any existing storage facilities. This is particularly common in resource-constrained settings. However, these estimates come with the *caveat* that they are based on our institutions and may vary considerably in other settings.

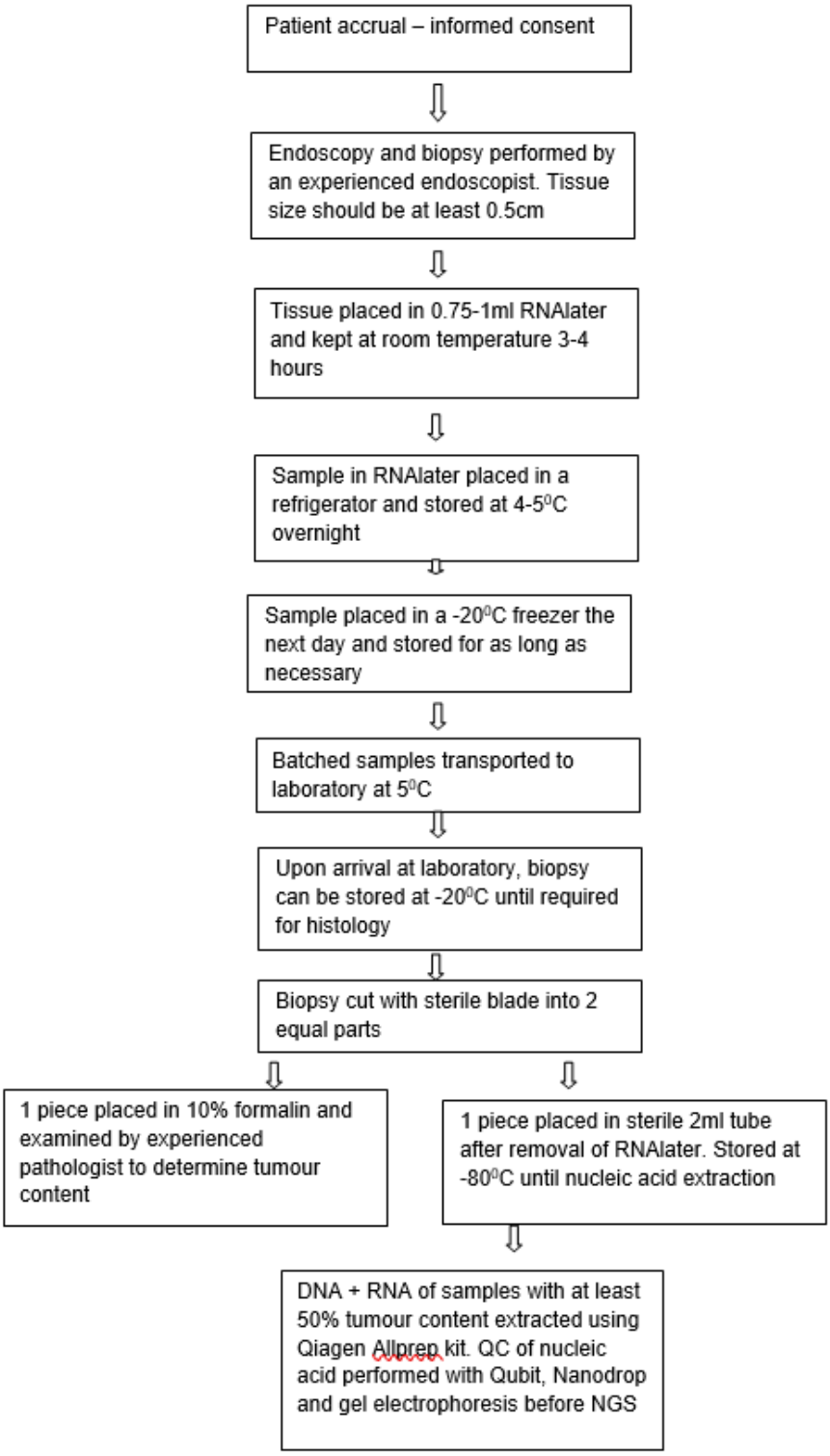


Figure 3- 4: Recommended operational workflow for tissue collection in a low-resource setting

In conclusion, our results indicate that it is possible to produce good quality samples for genomic research in an environment lacking on-site laboratory facilities geared towards this type of research without compromising important preanalytical variables that may influence results. The use of RNAlater, the availability of a standard

refrigerator and -20°C freezer as well as collaboration with a centre with the appropriate laboratory facilities and staff are essential. A recommended workflow for tissue biobanking in centres with limited resources is shown in Figure 3- 4. We suggest that this workflow is both practical and cost effective and thus feasible for use in resource-constrained settings. It may also be applicable to other types of tissues and tumours, subject to the outcome of the kind of QC assessments outlined in this study.

Chapter 4: The Methylome of South African Esophageal Squamous Cell Carcinoma (Manuscript 2)

4.1 Abstract

Esophageal squamous cell carcinoma (ESCC) is common in Southern and Eastern Africa with a high mortality, yet little is known about the contribution of genetic factors to this disease. Epigenetic mechanisms such as DNA methylation have been shown to contribute to ESCC carcinogenesis in Asian populations, but only one study has carried out a methylome study in African populations. The most recently developed methylation targeted assay, the Illumina TruSeq Methyl Capture EPIC library prep (TruSeq EPIC), leverages the power of targeted bisulfite sequencing, builds on the foundation of the Infinium MethylationEPIC BeadChip Microarray (850k EPIC Array) and targets 3.3 million CpG sites. We performed a methylome study using the TruSeq EPIC assay and matched transcriptome analysis on 11 tumour normal pairs derived from Black South African ESCC patients. We identified three critical gene promoters (*CDC42EP3, TNC* and *KRT13*) and 1 enhancer (*COL6A3*) that exhibited significant differential methylation and a significant inverse relationship with gene expression. All of these transcriptional regulatory elements function in the extracellular matrix (ECM), suggesting that the ECM through influencing the tumour microenvironment may promotes ESCC tumourigenesis.

4.2 Introduction

Esophageal squamous cell carcinoma (ESCC) is the seventh-prevailing tumour globally (Huang et al, 2021), particularly within sub-Saharan African countries together with South American endemic pockets and the 'Asian Esophageal Cancer Belt' (Northern Iran, China, Japan and Central Asia) (Reed and Johnston, 1993). Recent reviews concerning ESCC prevalence within South Africa (SA) highlighted its accountability for the second-peaking highest cancer-related mortality rate (46.7/100,000 mortalities and 19.2/100,000 mortalities for males and females, respectively) (Loots et al, 2017).

ESCC etiology remains unclear, and risk factors are thought to differ geographically. Most ESCC patients in SA present at a late-stage , with a median survival of 120 days (Bergquist et al, 2005). Such poor prognoses underscore the need to identify early stage diagnostic biomarkers to improve timely intervention and therapeutic strategies. Several studies have investigated ESCC methylation status across selected genomic regions, whereby key methylation-sensitive genes were found to be linked with cell cycle signalling, cell proliferation, apoptosis, DNA damage repair, and the *WNT*, *TP53, TGF*-β pathways and cell part morphogenesis (Cao et al, 2020; Gao et al, 2014; Guo et al, 2016; Huang et al, 2017, p. 1; Jiang et al, 2018, p. 132; Jin et al, 2007; Kishino et al, 2016; Kuroki et al, 2003; Lee et al, 2008; Lima et al, 2011; Lu et al, 2019; Peng et al, 2016; Saito et al, 2014; Talukdar et al, 2021; Wu et al, 2014; Xi et al 2022; Yun et al, 2015; Y Zheng et al, 2011). The 450k array and 850k array (Illumina™), which target 450 000 and 850 000 CpG sites, respectively, expanded the scope of methylation investigational studies . Two recent ESCC studies employing Infinium Human Methylation 450K and 850K BeadChip® (Illumina™) integrated methylome data with transcriptomic data through TCGA. Talukdar *et al.* identified three genes, *PAX9*, *SIM2*, and *THSD4*, demonstrating a significant inverse correlation between gene expression and DNA methylation. The authors concluded that such oncogenomic dysregulations are common events regulated by aberrant DNA methylation (Talukdar et al, 2021). Zheng *et al.* identified two molecular subtypes associated with immune-related pathways, the tumour microenvironment and prognosis (Zheng et al, 2022). Recently, the TruSeq Methyl Capture EPIC sequencing assay (Illumina™) was launched, probing > 3.3 million CpGs, with increased sensitivity (Lin et al, 2021) and a wider dynamic range when quantifying hypo/hypermethylated regions (Lin et al, 2021). Furthermore, an expanded probe range targets additional regulatory regions on this novel platform, allowing increased coverage of enhancers, DNAse hypersensitive sites and transcription factor binding sites, increasing the discovery potential for this assay.

This study employed the latest TruSeq Methyl Capture EPIC® (Illumina™) assay for probing DMRs within ESCC tumour/matched normal tissue (Figure 4- 1).

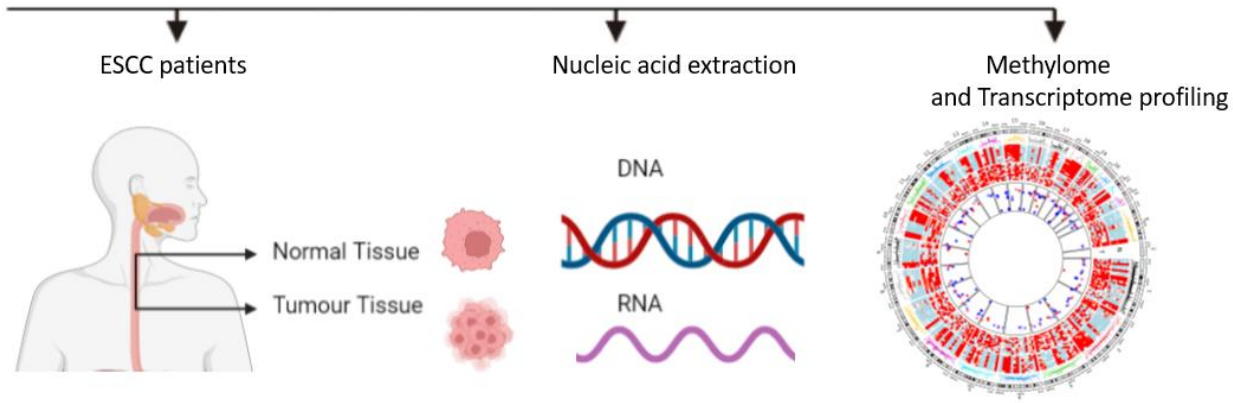


Figure 4- 1: First genome-wide methylation study on 3.3 million CpG targets with matched transcriptome data in South African ESCC patients

4.3 Methods

4.3.1 Specimen Collection, Methods and Study Design

The International Agency for Research on Cancer (IARC) protocol was used as a guideline to collect and process ESCCs' patient biospecimens (Mendy et al, 2017). Our detailed protocol for sample collection, storage, histology and nucleic acid extraction is described in Ferndale-Moodley (Ferndale et al, 2021b). In brief, patients with a provisional diagnosis of ESCC were referred for endoscopy at CMJAH. We recruited a total of 50 suspected ESCC cases of which 17/50 patients (34%) were positive for ESCC, 25 cases (50%) had dysplasia and the remaining 13 (16%) of cases were negative. Ethics clearance was obtained from the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Wits) (M170871) in line with the Helsinki declaration (2013). Samples with ≥60% tumour content and matched adjacent normal tissue (0% tumour) were the primary criteria for inclusion. The majority (58%) of patients in our cohort had 100% tumour content. For this study, twelve samples were chosen in accordance with the above criteria and also accounting for budget limitations. One sample was later excluded due to poor quality sequencing data. Therefore, a total of 11 paired tumour and normal samples were sequenced for the analysis of the methylome and transcriptome. Patient characteristics are shown in Table 4- 1. All patients were of African ethnicity, and all samples had confirmed squamous cell lesions. Photomicrographs of histology results can be found in Ferndale-Moodley *et al.* (Ferndale et al, 2021b). The mean age of the patients was 60.81 years (ranging from 38 to 78), and the cases included seven females and four

males. Three patients consumed both tobacco and alcohol, one patient had a smoking history, and the other seven had no exposure to known ESCC risk factors.

Table 4- 1: Patient characteristics of specimens processed in this study

Sample number	Age	Gender	Tumour grade	Tumour %	Current smoker	Current alcohol consumption	Current fuel used for cooking
2	69	Female	Invasive Carcinoma	100	Never	Never	Wood
3	65	Female	Invasive Carcinoma	100	Never	Never	Wood
4	78	Male	Invasive Carcinoma	100	Never	Never	Wood
5	56	Male	Invasive Carcinoma	100	Never	Never	Electricity
6	67	Male	Invasive Carcinoma	75	Smoked for 3 years	Never	Paraffin
7	53	Female	Keratinising Carcinoma	80	Never	Never	Electricity. Wood in the past
8	55	Female	Keratinising Carcinoma	60	Yes. 7 per day	750 ml twice per week	Electricity
9	66	Female	Keratinising Carcinoma	100	Never	Never	Wood
10	49	Male	Free lying Squamous	100	Yes. 5 per day	1 litre twice per week	Wood
11	38	Female	Keratinising Carcinoma	90	Yes. 60 per day	18 litres per week	Wood
12	73	Female	Keratinising Carcinoma	90	Never	Never	Electricity

4.3.2 Sample Processing Using the Truseq Methyl EPIC Assay

Library enrichment for methylome profiling was performed using the Illumina TruSeq Methyl Capture EPIC library kit (Illumina©). The TruSeq assay builds on the 850k bead array (Illumina©) and includes additional regions of known methylation sites of interest. This assay uses the power of NGS to perform targeted bisulfite sequencing simultaneously at multiple points in the genome. The 3.34 million CpGs targeted by this assay include 3'UTR regions, Transcription Start Site 200, Transcription Start Site 1500, promoters, 5'UTR regions, gene exons, introns and intergenic regions ("TruSeq Methyl Capture EPIC Library Prep Kit Documentation," n.d.). The Illumina EPIC Truseq protocol was followed according to the manufacturer's instructions without any modifications. In brief, 500 ng DNA was sonicated with the Covaris E220 sonicator (Covaris, Woburn, MA) to obtain >60% fragments between 100–300 bp in length as per the manufacturer's protocol specifications. Product sizes were checked on the Labchip© instrument (Perkin Elmer®). DNA fragments were hybridised to EPIC oligos, captured with streptavidin beads and washed twice to remove nonspecific binding. Hybridised products were subjected to bisulfite conversion and PCR amplification for 11 cycles using Kapa HiFi HotStart Uracil and DNA polymerase (Kapa Biosystems). Libraries were clustered on a dual flow cell using TruSeq Paired-End (PE) cluster (v3) for 100 cycles (2x150 PE) on an Illumina HiSeq 2500 sequencer (Illumina®).

4.3.3 Quality Control

QC of all sequencing reads (methylome and transcriptome) was performed using FastQC version 0.11.8 (Andrews, 2010). Adapters and bases with a Phred quality score below 30 were removed using Trim Galore! version 0.4.1 (Krueger, 2019). Reads were then aligned to the human reference genome (hg38). The methylation status of all cytosines (for every uniquely mappable read) was called with the cytosine2coverage and _methylation_extractor packages (Bismark v0.20.0). Spike-in control sequences were used to check the bisulfite conversion rates and validate the efficiency of the bisulfite treatment. The percentage of methylation at each CpG site was computed for each sample. Only CpGs with a coverage > 10X in all paired samples were included, and reads

with extremely high coverage (>99.9th percentile) were discarded due to possible PCR bias(Akalin et al, 2012). All downstream analyses were performed in the R/Bioconductor environment.

4.3.4 Differential Methylation Analysis

Methylkit is an R-based software package that is the preferred tool among other methylome software packages for DMR detection under several parameters, such as different methylation percentages, sequencing coverage, depth of sequencing, length (bp) of DMRs and sample sizes, (Wreczycka et al, 2017) and thus, was chosen for this study (MethylKit v1.7.05).

Methylkit

Methylkit uses logistic regression to compare methylation percentages between matched tumour and normal tissues at given CpG regions. The results are tested against the null hypothesis (no differences in methylation between our two groups). For multiple comparisons, testing p values were adjusted to q-values using the sliding window model (SLIM) (Akalin et al, 2012). Paired analyses were performed on tumour/normal matched samples to identify differentially methylated CpGs (DMCs) and DMRs. DMRs were assigned a window of 1000 bp in length with at least two CpGs or more. Statistically significant DMCs and DMRs were identified by a q-value cut-off ≤ 0.01 and a methylation difference higher than 20% for promoters and 10% for enhancers that needed more stringent criterias. All DMCs and DMRs were annotated with the R/Bioconductor package annotatr (Yu et al, 2015) using the RefGene and CpG island annotations from the UCSC (hg38) website (https://genome.ucsc.edu/cgi-bin/hgGateway). The annotation comprised two categories: (1) distance from a gene to a CpG island and (2) regional annotation. Regions flanking CpG islands located at a distance of up to 2000 bps are assigned as shores; regions beyond 2000 bps are referred to as shelves, and open-sea are other regions further away. For gene region annotation, DMCs and DMRs were classified into three groups, namely, overlapping, exons, introns and intergenic regions. Using the above criteria, genomic plots were generated using the 'ChIPseeker' R package (Yu et al, 2015), which allowed visualisation of the above genomic annotations. To identify

significant differential methylation in promoters, MethylKit was used to segment the methylome into distinct categories of high and low methylation regions based on change-point analysis. In brief, change points are recorded, and the genome is partitioned into regions between consecutive change points using ≥2 CpGs with significant changes (q-value ≤ 0.01) in low or high methylation(Akalin et al, 2012). These so-called “segment DMRs” were then annotated to specific promoter regions for further analysis. We also investigated segments that overlapped with enhancers according to predefined annotations (“enhancer DMRs”). Enhancer coordinates were downloaded from the Functional AnnoTation Of the Mammalian genome (FANTOM) database (hg38 assembly). To plot DMRs in enhancer regions, we used the plot DMRs function of the ‘dmrseq’ R package (Korthauer et al, 2019), which transforms the output file obtained with MethylKit software into a BSseq object that provides methylation and coverage information for all the samples and their corresponding sample IDs.

4.3.5 RNA Sequencing

RNA sequencing and transcript normalisation

Samples (2-12) were sequenced on the Illumina HiSeq 2500. RNA was extracted from patient tissue with the AllPrep assay (Qiagen®). RNA integrity number (RIN) was calculated using the Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). Samples with a RIN value ≥8 were included for mRNA sequencing. Library preparation was performed with the TruSeq Stranded mRNA assay according to the manufacturer’s guidelines. Briefly, 1 µg RNA was purified with poly-DT magnetic beads to capture mRNA and fragmented. RNA was reverse transcribed into cDNA using Superscript VI® (Thermo Fisher). Second strand cDNA was generated to ensure strand specificity by substituting dTTP with dUTP. Double-stranded cDNA was processed for end repair, A-tailing, adaptor ligation and library enrichment. The quality and quantity of libraries were analysed with a Labchip GX® analyser (Perkin Elmer) and Qubit® assay (Thermo Fisher). RNA libraries were sequenced on the 125 bp paired-end runs on the Illumina HiSeq2500 platform. Reads from transcriptome data were demultiplexed before analysis.

After initial QC with FastQC, transcripts were mapped to the reference genome (hg38) and quantified using the software package Hisat2 (Kim et al, 2019). Count data for all samples were combined and normalised with EdgeR (Robinson et al, 2010) (version 4.1). The strategy used for differential analysis between tumour/normal samples was to first obtain differential expression on transcripts and then aggregate transcript-level p values to obtain gene-level p values (Yi et al, 2018). Recently, this approach has been shown to increase sensitivity and accuracy in identifying differentially expressed genes (DEGs) (Mehmood et al, 2021; Yi et al, 2018). To account for individual effects, we used a linear regression with the tissue type (normal or tumour) as the main variable and the patient identification as the covariate. Genes with a fold change ≥2 (up and down) and FDR-adjusted p value ≤ 0.05 were considered significant.

4.3.6 Pathway Analysis

Pathway analysis was carried out on hypermethylated and hypomethylated promoters and enhancers using Gene Set Enrichment Analysis (GSEA) (Subramanian et al, 2005) software through the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and Gene Ontology (Biological Processes – BP) databases. GSEA was also used for pathway analysis of the DEG data. Reactome pathway analysis ('ReactomePA)' (Yu and He, 2016), an R package, was used for the visualisation of enriched pathways with gene set concept network plots.

DMRs in promoters and enhancers with a 20% difference in methylation and q-value ≤ 0.01 were correlated with RNA expression if the same gene had an inverse correlation with an expression level that exhibited a 2-fold log change with an FDR ≤ 0.05 using the Intersect function, which is built in R studio (Chan, 2018). Circos plots comprising transcriptome and methylome data were rendered using the 'circlise' package.

An overview of the bioinformatics approach used in this study is shown in Figure 4- 6.

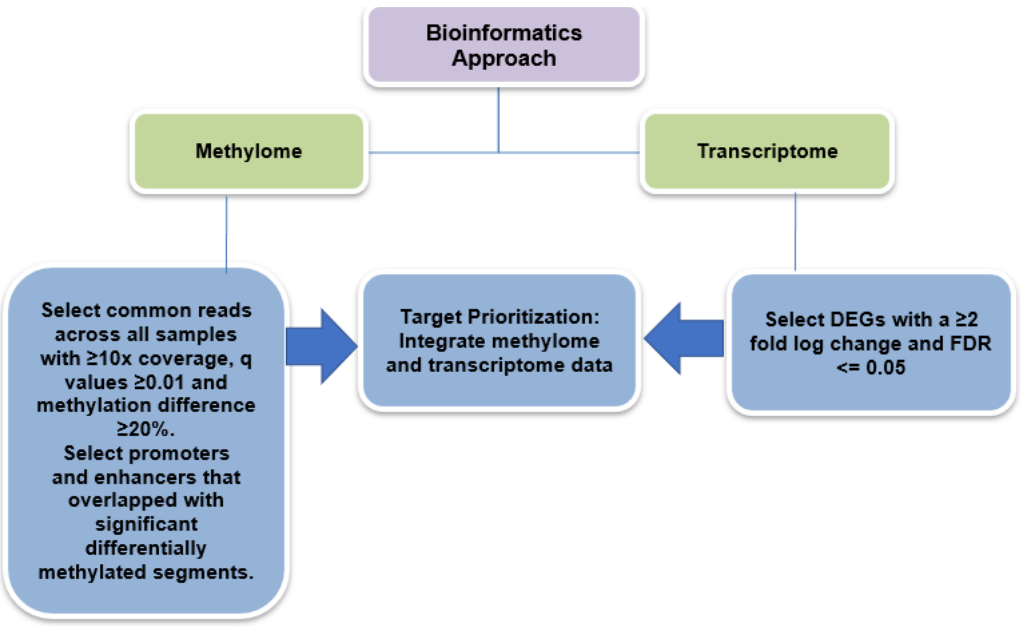


Figure 4- 2: Overview of the bioinformatics approach used for this study

4.4 Results

4.4.1 CpG Differential Methylation across the tumour genome

Figure 4- 2 shows the percentage of hypo and hypermethylated cytosines, highlighting a predominance of hypomethylated CpGs (67%) in tumour samples compared to hypermethylated CpGs. Paired analysis identified 953,132 DMCs, from which 106,257 were differentially methylated, with a q-value ≤0.01 and ≥25% difference in methylation (Figure 4- 3). Despite the heterogeneity of tumour samples, CpG methylation-based PCA could discriminate most tumour samples from their nontumour counterparts (Figure 4- 4), with the exception of tumour sample T7, which clustered with most normal samples and normal adjacent tissue. Normal sample N11 clustered closer to tumour samples.

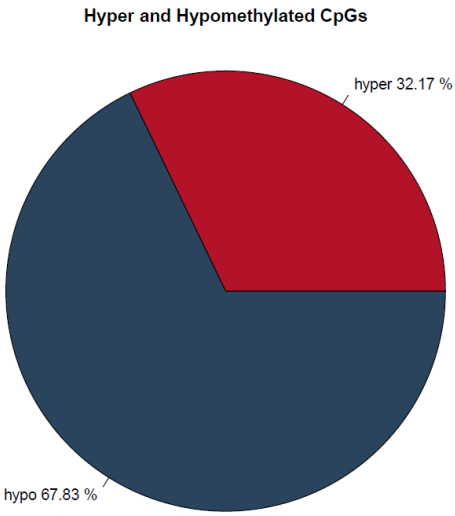


Figure 4- 3: Pie chart illustrating hypo (blue) and hypermethylated (red) DMCs in tumour vs normal samples

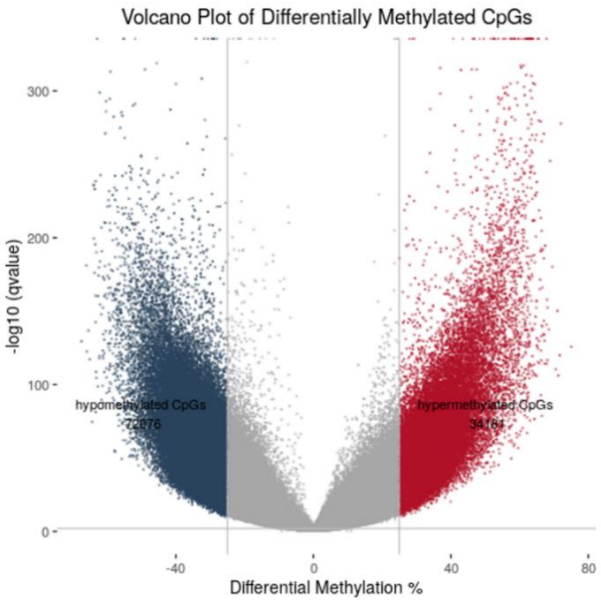


Figure 4- 4; Volcano plot of DMCs. Blue represents hypomethylated Cytosines (≥20%) and red represents hypermethylated Cytosines (≥20%), with q-values ≤0.01. Among 106,257 CpGs having ≥20% difference (hypomethylated or hypermethylated) and a q-value ≤0.01, 67% were hypomethylated while 33% were hypermethylated. Differentially methylated CpGs with higher significant q-values are located to the upper-left/-right sides for the plot with highly significant changes appearing higher on the plot. Values on the x- and y-axes are percent methylation differences and negative log10 for corrected q-values, respectively.

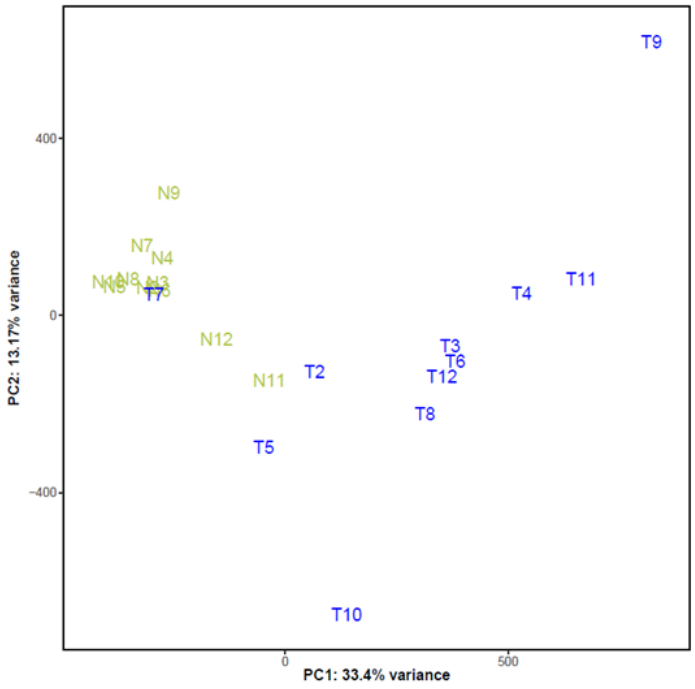
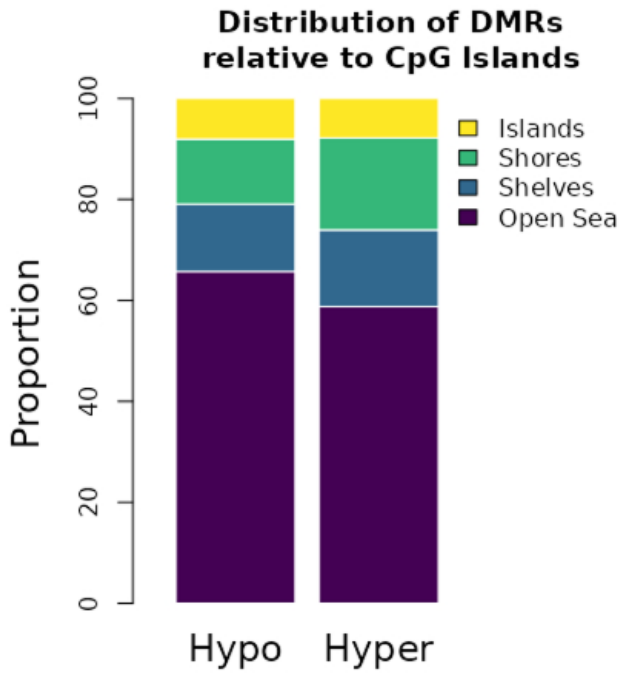


Figure 4- 5: Illustrations of differential methylation across the genome and PCA analysis between our 2 groups. PCA using all DMCs (Normal - green, Tumour – blue). PCA scores on the x-axis (PC1) and the y-axis (PC2).

4.4.2 Genome-wide DMR distributions

Most hypo/hypermethylated CpGs (>50%) were located in open-sea regions (Figure 4-6A). Paired segmentation analysis identified 36,264 differentially hypomethylated and hypermethylated regions (q-value ≤ 0.01), of which 13,486 spanned annotated genes. Most DMRs were located in promoters and intergenic regions (Figure 4- 5B). The proportions of hypomethylated DMRs were as follows: promoters (28%), 5'UTR (1.4%), 3'UTR (3.3%), exons (18%), introns (22.7%) and distal intergenic regions (26.5%). The proportions of hypermethylated DMRs were as follows: promoters (40%), 5'UTR (1.2%), 3'UTR (4.7%), exons (8.5%), introns (30.3%) and distal intergenic regions (14.7%). Since most DMRs overlapped with promoters and distal intergenic regions, the study focused on differential methylation (tumour/normal) in these regions (Thurman et al, 2012).

A


Distribution of DMRs
relative to CpG Islands
Proportion
0
20
40
60
80
100
Hypo
Hyper
Islands
Shores
Shelves
Open Sea

B

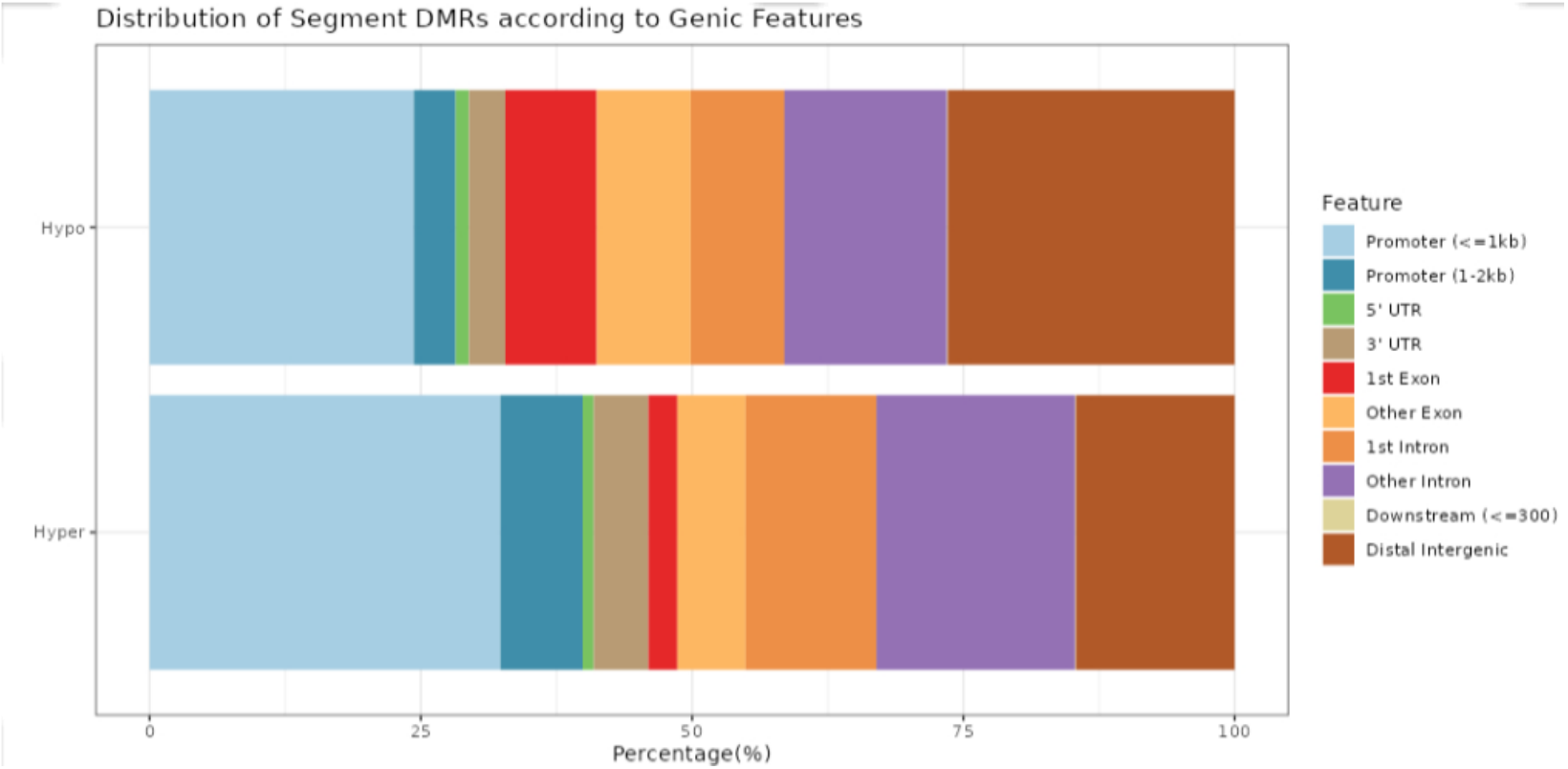


Figure 4- 6: Distribution of DMRs relative to CpG islands (A) and gene features (B). A, Differential methylation pattern: proportion of DMRs (Tumour/Normal) relative to CpG islands across all samples. B, Distribution of Segment DMRs relative to gene features, their distance to Transcription Start Sites (TSS) and gene feature context (promoters, 5' UTR, 3' UTR, exons and introns. The y-axis shows hypomethylated (top-panel) and hypermethylated regions (bottom-panel). The proportion of methylation (percentage) is shown on the x – axis.

4.4.3 Hypomethylated and hypermethylated pathways across gene promoters

Interrogation of pathways analysis softwares: BP, KEGG and Reactome, revealed that hypomethylated promoters were enriched in genes involved in cell–cell adhesion mechanisms, adaptive immune response, cyclic AMP (cAMP) signaling pathway, hemostasis, embryonic morphogenesis and adaptive defence processes. Hypermethylated genes analyzed across identical above repositories and algorithms, belonged to myeloid cell activation, innate immune response, chemical synaptic signalling, cytokine–cytokine receptor interaction, G protein-coupled receptors (GPRC) pathway, central nervous system (CNS) processes and nervous system development pathway (supplementary Figure S 4- 1A-C). Some of these pathways, including cell–cell adhesion (Li et al, 2017), cytokine–cytokine receptor interaction (Davies, 2011), innate immune system mechanisms and nervous system processes, were detected in previous ESCC studies (Huang and Fu, 2019; Griffin et al., 2020). Within these pathways, specific genes previously reported as hypomethylated in in ovarian and bladder cancer included *ATG7, HRAS* and *CALML3* (Zhu et al., 2019). Hypermethylated CNS-linked gene promoter included the *LHX6* gene promoter. Furthermore, multiple interleukin (IL) gene family members (*IL6R, IL6ST* and *IL10RB*) were hypermethylated and belong to the cytokine–cytokine receptor interaction pathway. Within the cytokine-cytokine interaction pathway the *RAP1* and the signalling pathway *CDH1* act as a tumour suppressor genes. The protein product of *CDH1* acts as an adhesion molecule and makes it difficult for tumour cells to detatch from the original tumour. In ESCC hypermethylation of *CDH1* results in decreased cell adhesion which stimulates tumour metastasis (Ling et al, 2011).

4.4.4 Enhancer Methylation Analysis

FANTOM-based analyses identified 1008 genes overlapping with differentially methylated enhancers, of which 51 enhancers were significantly (≥20%) hypermethylated and 132 enhancers were significantly (≥20%) hypomethylated within the tumour group compared to normal tissue. Since only a few enhancers were significantly differentially methylated (n=183), these were insufficient to interrogate the KEGG, GSEA and Reactome algorithms. Once the parameters were relaxed to include additional

hypo/hypermethylated enhancers, using a 10% methylation shift (instead of 20%) while keeping a q-value ≤ 0.01, we identified 239 hypomethylated DMRs in enhancers (neighbouring 219 annotated genes) and 195 hypermethylated enhancer DMRs (in the vicinity of 182 annotated genes). These could be used for pathway analysis (Supplementary Table S 4- 1) as detailed below.

Using the Reactome algorithm, genes found to be associated with hypomethylated enhancers regions were mostly related to the degradation of the extracellular matrix (ECM) (Supplementary Figure S4- 2A). Six hypomethylated enhancers governing the expression of *COL6A3, COL6A2, CACNB2, CTSD, TPSAB1 and TIMP2* genes were affected. Other hypomethylated, Reactome-enriched enhancers included *RHOB* and *RHOA GTPases*, the protein product of which interact with a wide range of targets affecting changes in the cytoskeleton (Gutierrez et al, 2019) and cell adhesion (Ridley, 2015) processes. ECM linked cell migration and adhesion through activation of small RhoGTPases, was previously described in OSCC (Ramos et al, 2016). Overall, pathway analysis from enhancer data suggests changes to the tumour microenvironment through ECM disruption.

Profiling of hypermethylated enhancers using the Reactome analytical algorithm revealed changes in ESCC involving the transcription factor AP-2 alpha (*TFAP2*) and the TGF-beta (β) complex pathways (Supplementary Figure S4- 2B). The gene promoter of the *ESR1* gene involved in the TFAP2 pathway was hypermethylated. *ESR1*-enhancer hypermethylation was described in breast cancer studies and was proposed as a mechanism of endocrine resistance in these malignancies (Bos et al, 2022; Stone et al, 2015). Although TGF-beta (β) promoter hypermethylation has been previously detected in several cancers, including lung (Shah et al, 2008), prostate (Shah et al, 2008), ovarian (N Wang et al, 2012) and ESCC (Ma et al, 2020), there were no published reports of *TGF-B* enhancer hyper-methylation.

4.4.5 Matched Transcriptome Analysis

Overall, 241 DEGs were identified, with 71 downregulated and 170 upregulated in tumours relative to their non-tumour adjacent tissues and consistent across all samples. The top-ranking 20 DEGs are depicted in Table 4- 2.

Table 4- 2: Top-ranking 20 DEGs, according to most significant p value (False Discovery Rate – FDR)

Gene	logFC*	FDR
FN1	9.7	7.88E-19
POSTN	9.1	4.23E-16
COL6A3	8.5	9.64E-09
LPIN1	-2.7	9.82E-09
PTHLH	7.7	9.82E-09
TGFBI	5.9	2.22E-08
KRT78	-6.9	1.10E-07
COL11A1	7.7	2.43E-07
SPINK5	-3.8	1.41E-06
KRT13	-4.8	3.84E-06
TMPRSS11B	-8	8.71E-06
PHLDB2	5.2	1.02E-05
IL32	4.3	1.41E-05
EMP1	-3.5	2.40E-05
C1S	4.3	5.71E-05
MGLL	-2.6	6.32E-05
COL16A1	5.26	6.79E-05
ECM1	-4.5	1.03E-04
SGIP1	5.1	1.07E-4
SPP1	6.6	1.48E-4

Significant changes in gene expression identified eight top genes involved in the ECM (Demir et al, 2002; Hamada et al, 2002; Ho et al, 2022; Singh and Schwarzbauer, 2012;

Zhang et al, 2012): *FN1, POSTN, COL6A3, KRT78, COL11A1, KRT13, COL16A1* and *ECM1*. All genes listed in Table 4- 2 also had a logFC>2, with 12 genes that were overexpressed and 8 underexpressed.

Unsurprisingly, ECM was also enriched through GSEA pathway analysis. Several other pathways were KEGG-/Reactome-enriched for proteoglycans and actin cytoskeleton. Aberrant expression of actin-related proteins and ECM was previously reported in ESCC patients (Palumbo et al, 2020; Zhan et al, 2017). In the present study the ECM pathway was enriched in both tumour enhancers and transcriptome data.

4.4.6 Genomic Integration

Methylome and transcriptome data integration was first studied at the genomic level, with Circos representation of DNA methylation (≥20%) and differential expression (≥2-fold) locations across differing chromosomes shown in Figure 4- 5.

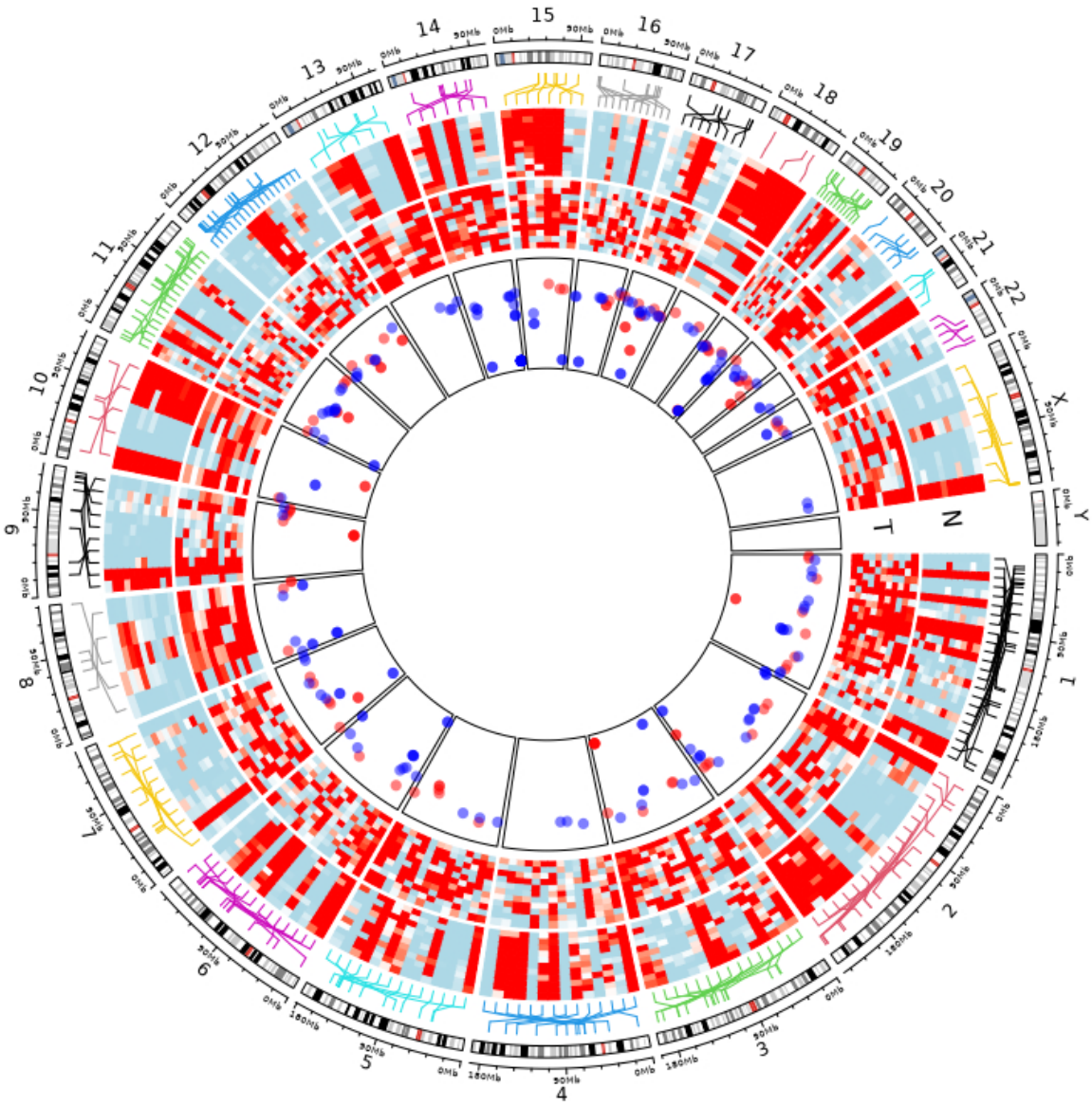


Figure 4- 7: Circos plot illustrating differentially methylated promoters with a methylation difference ≥20% and gene expression with an expression difference ≥2-fold between tumour and normal samples. From outer to inner circles: the ideogram with hg38 chromosomal locations having coloured links between heatmaps and corresponding locations. The first heatmap track corresponds to gene expression across all normal samples (N), while the second heatmap track refers to the gene expression across tumour samples (T) with red representing overexpression and blue underexpression relative to normal samples The innermost track corresponds to DMRs (q ≤ 0.01) overlapping promoters and enhancers. Red dots are hypermethylated (≥20%), and blue dots are hypomethylated (≤20%) in tumours.

It is well established that cancer cells overexpress more genes than normal cells (Lotem et al, 2005) and that there is also a general pattern of global hypomethylation in tumour cells compared to normal cells (Feinberg and Vogelstein, 1983). The Circos plot also reflects such a pattern for mRNA expression and hypomethylation (≥20%) in tumour versus normal samples. Specific chromosomes that display contrast between the tumour and normal tissue pattern include visible regions across chromosomes 1, 3, 7, 13, 16, 20 and 22.

To prioritise candidate genes that may be affected by differential methylation (tumour versus matched normal tissue), we selected promoters and enhancers that showed significant methylation differences (tumour/normal) and showed inverse RNA expression of their corresponding genes. We referred to these as “Hypo-up” (undermethylated/ increased gene expression) and Hyper-down (hypermethylated/ decreased gene expression). Analysing the fraction of transcriptional changes specific to ESCC associated with DNA methylation dysregulation, 322 Hypo-up promoters and 163 Hyper-down promoters were identified, together with 192 Hypo-up enhancers and 71 Hyper-down enhancers (Supplementary Table S 4- 3). These data also demonstrated that methylome and transcriptome changes in ESCC are spread throughout most chromosomal locations.

4.4.7 Pathway analysis after integrating differentially methylated promoters and enhancers with matched transcriptome data

GSEA-derived gene lists uploaded within the Reactome database showed that receptor tyrosine kinases, nervous system development and hemostasis pathways were enriched across both transcriptome and promoter methylation data. After integrating methylation with RNA expression, four genes were hypo-upregulated (*PIK3R5, TUBB6, PLXNA4, PRKCZ*), and two were hyper-down regulated (*RGMA*, *ESRP2*). The phosphatidylinositol 3-kinases (PI3Ks) family of proteins play roles in cellular signalling and cancer. The PI3K gene family integrates signals from growth factors and cytokines and regulates several signalling pathways through intracellular signalling (Thorpe et al, 2015). Although there are no reports of Hypo-up uregulation associated with *PIK3R5*, somatic mutations were previously described in the *PIK3R5* gene that promote cell survival and oncogenesis in colon cancer (Jaiswal et al, 2009) and melanoma (Shull et al, 2012). There has only been

one report of hypomethylation of the *TUBB6* gene promoter in colon cancer (Beggs et al, 2019). Regarding enhancers notably, six hypo-up enhancers affected genes invoved in the ECM degradation: *TIMP2, COL6A2, CACNB2, CTSD, TPSAB1* and *COL6A3*.

Analyses of enriched pathways from promoter and enhancer candidates revealed that most pathways belonged to the tumour microenvironment (TME). The main components of the TME in ESCC include immune cells, cytokines, the ECM (Wei et al, 2020), and CNS related genes (Griffin et al, 2020). Interactions of cell–cell adhesion proteins with the TME also drive tumour development (Harjunpää et al, 2019).

4.4.8 Candidate Gene Prioritisation

Using these Hypo-up/Hyper-down criteria, we further filtered our results to include promoters and enhancers that showed significant methylation variations (≥20%) with q-values ≤ 0.01 and that also had an inverse relationship to RNA expression (FDR ≤ 0.05 and a ≥2-fold change in gene expression). We detected 15 significant gene promoters (12 Hypo-up and three hyper-down), with four significant Hypo-up (Table 4- 3). When examining the function of individual candidates, three gene promoters (*CDC42EP3*, *TNC* and *KRT13*) and one enhancer (*COL6A3*) were found to have ECM function(s).

Table 4- 3: Promoters and enhancers demonstrating significant methylation variations (≤20%) with q-value ≥0.01 and RNA fold-changes in opposite direction (≤2 logFC)

Gene	DMR	Class	*Methylation difference (%)	Number of CpGs	Number of samples (n=11)	Differential expression (logFC change)	FDR
*KRT13**	Promoter	Hyper-down	21	9	9	-4.8	3.84E-06
TNC	Promoter	Hypo-up	-28	11	9	4.3	0.0002
INHBA	Promoter	Hypo-up	-27	9	11	5.7	0.0006
FAP	Promoter	Hypo-up	-32	6	11	4.6	0.0020
UPK1A	Promoter	Hyper-down	25	6	7	-5.3	0.0027
DUXAP10	Promoter	Hypo-up	-34	7	6	5.0	0.0029
ABLIM1	Promoter	Hyper-down	36	7	11	-2.2	0.0080
LRRC32	Promoter	Hypo-up	-35	5	11	4.6	0.0100
PARP12	Promoter	Hypo-up	-29	6	8	4.9	0.0122
EGFLAM	Promoter	Hypo-up	-33	5	8	4.2	0.0166
SYNE1	Promoter	Hypo-up	-35	16	11	4.5	0.0185
THY1	Promoter	Hypo-up	-30	8	10	3.4	0.0290
PCCB	Promoter	Hypo-up	-29	19	9	4.4	0.0245
CDC42EP3	Promoter	Hypo-up	-40	20	10	6.8	0.0488
CDKN2A	Promoter	Hypo-up	-46	14	10	3.2	0.0493
*COL6A3**	Enhancer	Hypo-up	-41	13	11	8.4	9.65E-09

Gene	DMR	Class	*Methylation difference (%)	Number of CpGs	Number of samples (n=11)	Differential expression (logFC change)	FDR
SMYD3	Enhancer	Hypo-up	-29	7	9	6.0	0.001
NEK6	Enhancer	Hypo-up	-22	23	10	3.7	0.010
TTI1	Enhancer	Hypo-up	-28	11	9	5.5	0.041

*Genes are among the most differentially expressed and feature in Table 4.2

4.5 Discussion

DNA methylation provides a promising ESCC biomarker since methylation shifts occur in the early phase of the disease (Nishiyama and Nakanishi, 2021a). To our knowledge, we are the first group to probe the TruSeq Methyl Capture EPIC® assay in ESCC. Combining this technique with full transcriptome profiling, this dataset revealed deep-genomic dysregulation in ESCC at the gene regulation and pathway levels. It has been established that 70% of CpG islands are found in promoter regions (Bitzer and Malek, 2016), which could explain promoters having DMR abundance in this study. Aberrant methylation of gene promoters is a hallmark of several tumour models, and widespread differential methylation is an early event that leads to gene amplification in ESCC (Alvarez et al, 2011). Consequently, several studies have focused on the inverse relationship between DNA methylation in promoter regions and the transcriptome (Aran and Hellman, 2013; Bell et al, 2016). Enhancers inhabit a much larger genome segment than promoters and have a significant influence on cancer progression and prognosis (Aran and Hellman, 2013; Bell et al, 2016). Moreover, it has been suggested that predicting gene expression levels using enhancer sites is better than using promoters (Aran and Hellman, 2013). This study sought to identify promoters and enhancers with significant methylation variations, which also demonstrated an inverse relationship with their associated genes expression in matched tumour/normal ESCC samples. Using the FANTOM database for enhancers, 1390 enhancers were identified, which had sufficient sequencing coverage (>10 counts), of which 1163 were differentially methylated, with a q-value ≤ 0.01 for the tumour vs normal paired cohort (Supplementary Table S4- 1). Only 257 CpG sites from the 850K array fell within the 1163 enhancer DMRs identified through

the TruSeq EPIC assay. Consequently, ~80% of enhancer DMRs, potentially linked to gene regulation, would have been missed if the 850K bead array was utilised. Compared to another study that employed 850K arrays, 67% of DMRs were defined mostly as two CpGs in the assigned window(Talukdar et al, 2021). Here, enhancers and promoters were identified through several consecutive significant CpGs. Talukdar et al (2021) did a comprehensive study on methylation changes in ESCC using 850K arrays. They included several distinct geographical populations with a high incidence of ESCC and they identified three genes promoters with significant hypermethylation: *PAX9*, *SIM2*, and *THSD4 genes*. However, none of these three gene promoters were shown to be significant within the 21 South African cases included in their cohorts.This underscores the need to perform genomic studies for unique geographical locations, possibly associated with specific risk factors and etiologies. In the same study RNA sequencing data were obtained from the TCGA data base to correlate expression with their methylation results (Talukdar et al, 2021). While databases such as TCGA are extensive, comparing data from differing patients and experiments is challenging due to variations in sample handling/processing, type of RNA sequencing platform/chemistry used and the analysis pipeline employed to normalise batch effects (Q Wang et al, 2018). Also, gene expression from TCGA ESCC patients were mostly Caucasian descent (Talukdar et al, 2021)).In contrast to this , in the present study methylome and transcriptome data were all from matched patients, thus reducing inter-individual variability. We identified robust candidates from methylome and transcriptome data that may be of benefit for understanding ESCC initiation/progression in SA and identify potential clinical biomarkers. To prioritise candidate genes, we probed promoters and enhancers that showed significant methylation variations (≥20%) with q-value ≤0.01 and that also had an inverse relationship to RNA expression (FDR ≤ 0.05 and a ≥ 2-fold change in gene expression), referring to these states as Hypo-up and Hyper-down promoters or enhancers. Using these criteria, 15 promoters and four enhancers were identified. We noted significant differential methylation in three of these gene promoters (*CDC42EP3* (Hypo-up), *TNC* (Hypo-up) and *KRT13* (hyper-down)) and one enhancer (*COL6A3* (Hypo-up)), which affect the ECM. GSEA from transcriptome data and hypomethylated enhancers also highlighted enrichment for the ECM pathway. Our results reflect dysregulation of ECM that may contribute to the landscape of ESCC carcinogenesis in South African patients.

In particular, one hyper-down promoter*, KRT13*, topped the list of prioritised genes. *KRT13* had almost a 5-fold decrease in expression in tumours compared to normal tissues and had the most significant p value from transcriptome analysis. Keratins (KRTs) are structural proteins expressed in epithelial cells and interact with several proteins related to cell survival, death, migration, proliferation, metastasis and invasion (Hendrix et al, 1996; Kim et al, 2006). In lung cancer, KRT13 interacts with ECM (Busch et al, 2021). Promoter hypermethylation with *KRT13* silencing was reported in oral cell squamous carcinoma (OSCC) cases. OSCC and ESCC are thought to be related(Lee et al, 2017). They are both associated with identical major risk factors: alcohol and/or tobacco consumption (Ishiguro et al, 2009; Johnson, 2001). It has been suggested that *KRT13* could be used as a therapeutic target in OSCC patients and that treatment with 3-deazaneplanocin hydrochloride (DZNep) restores transcription of *KRT13* (Naganuma et al, 2014). This could be an interesting scenario if applicable to ESCC.

It is well established that the ECM plays a critical role in the TME, and its disruption has been frequently reported in ESCC (Palumbo et al, 2020; Saldin et al, 2019; Zhang et al, 2021). In cancer, tumours make use of ECM remodelling and create a microenvironment that promotes tumourigenesis and metastasis (Winkler et al, 2020). The ECM composition becomes dysregulated, altered, stiff and more abundant in most cancer (Lu et al, 2011). The *CDC42EP3* and *TNC* genes that form components of the EMC, were identified as Hypo-up promoters in this cohort. *CDC42EP3* gene was previously found to be upregulated and associated with advanced tumour progression in gastric cancer (Chen et al, 2021, p. 42). In addition, *CDC42EP3* has been demonstrated to be required for matrix remodelling, invasion, angiogenesis, and tumour-growth-promoting abilities of cancer-associated fibroblasts (CAFs) (Calvo et al, 2015). CAFs increase ECM deposition and remodelling, leading to elevated cellular/tissue tension and enhanced stiffness in solid tumours (DuFort et al, 2011; Levental et al, 2009). These physical changes exacerbate tumour cell growth, survival, and migration (DuFort et al, 2011). The other Hypo-up promoter in the present study, was the Tenascin-C (*TNC*), gene promoter an upregulated ECM glycoprotein within the TME in most epithelial malignancies (Orend and Chiquet-Ehrismann, 2006). Cancerous epithelial/stromal cells interact to facilitate tumour remodelling by upregulating *TNC* (Hynes, 2002). Both Hypo-up promoters identified in this study suggest that such dysregulation are likely to result in ECM remodelling.

Among the four identified hypo-up enhancers, *COL6A1* is known to play a role in ECM, and may be considered as a potential drug target in ESCC. The *COL6A3* enhancer region acted as a peak hypomethylated DMR (41%), demonstrating eight-fold upregulation of expression in tumour tissue. Collagen type VI α1 chain (*COL6A3*) encodes the α1 (VI) chain of type VI collagen and is a primary protein for ECM (Tanaka et al, 2003), proving vital for multiple tumour-model TMEs (Chen et al, 2013). Collagen VI consists of three subunits: alpha-1 (VI), alpha-2 (VI), and alpha-3 (VI) chains, providing ECM structural support. *COL6A3* encodes the alpha-1 (VI) repeat subunit (Chen et al, 2013). Early studies demonstrated that *COL6A3* overexpression leads to ECM disruption in ovarian cancer (Sherman-Baust et al, 2003). Since then, *COL6A3* has been shown to be overexpressed in almost 20 types of different cancers (Wang and Pan, 2020). Furthermore, the *COL6A3* related gene was also hypomethylated and transcriptionally upregulated in tumour development within uterine leiomyomas (Maekawa et al, 2013) and glioblastoma (Hyman et al, 2015). Functional studies have shown that overexpression of *COL6A3* in bladder (Huang et al, 2018), colorectal (Liu et al, 2018) and gastric (Ao et al, 2018) cancer cell lines. Collectively, all Hypo-up enhancers and promoters identified in this study were linked to ECM disruption. Only one other recent Chinese study investigated ESCC with matched methylome data and transcriptome data derived from the same patient samples. Although they used the earlier 450k array, which targets > 450,000 methylation sites in contrast to the 3.3 million CpG sites used in this study. Like our study, they selected promoters that showed significant methylation variations (≥20%) with q-value ≤0.01 and that also had an inverse relationship to RNA expression (FDR ≤ 0.05 and a ≥ 2-fold change in gene expression) and proposed a 12-marker diagnostic panel to distinguish high-risk ESCC patients (Xi et al, 2022). From their candidate genes, 2 genes also associated with ECM function: *COL14A1* and *MMP13*. This further reflects dysregulation in the ECM in ESCC.

In addition, we identified 12 other differentially methylated promoters of which 10 were hypomethylated (*CDKN2A, LRRC32, SYNE1, DUXAP10, EGFLAM, FAP, THY1, PARP12, PCCB* and *INHBA*) and two were hypermethylated (*UPK1A, ABLIM1*). Among these genes *UPK1A* (Kong et al, 2010; Li et al, 2013), *DUXAP10* (Z Wang et al, 2018) and *INHBA* (Lyu et al, 2018) were previously found to be upregulated in ESCC. UPK1A is a highly conserved protein in several organisms and act as a tumour-suppressor gene in ESCC (Kong *et al.*, 2010). This gene affects cell cycle arrest at the G1/S checkpoint and is associated with downregulation of cyclin D1, cyclin-dependent

kinase 4 and MMP7. Promoter hypermethylation of *UPK1A* with decreased RNA expression has been previously described with a poor prognosis in ESCC (Kong *et al.*, 2010). The *DUXAP10* pseudogene derived from long noncoding RNA (lncRNA), was found to be upregulated in in ESCC and colorectal cancer (CRC) and is associated with a short survival time. Knockdown of *DUXAP10* negatively influences ESCC cell proliferation. It has also been suggested that *DUXAP10* be used as a marker for early diagnosis and treatment of ESCC (Wang *et al.*, 2018). Inhibin subunit beta A (*INHBA*) encodes a member of the TGF-beta superfamily of proteins. Overexpression of *INHBA* has been reported in several cancers, including bladder, CRC, breast, gastric, lung, esophageal adenocarcinoma and in ESCC where it is associated with a poor prognosis (Lee et al, 2015; Okano et al, 2013; Bashir et al, 2015; Q Wang et al, 2012; Lyu et al, 2018; Seder et al, 2009). Although the molecular mechanism of *INHBA* has not been explained in ESCC (Lyu et al, 2018), in esophageal adenocarcinoma *INHBA* was found to be overexpressed as a result of promoter demethylation and histone acetylation (Seder et al, 2009) The same molecular mechanism could explain the overexpression of *INHBA* in ESCC, although further studies are warranted.

Last, this study investigated top-hypomethylated/top-hypermethylated gene promoters without integrating transcriptome data. One particular pathway enriched in KEGG analysis was cytokine–cytokine receptor interaction. During the early stage of ESCC, the immune system plays a critical role in preventing tumourigenesis (Zhang et al, 2021). Conversely, ESCC cells expand tumours by secreting cytokines, chemokines and growth factors (Chen et al, 2020). Several hypomethylated genes in this study were linked with immune system function or belonged to the chemokine ligand (CCL) and interleukins (IL) family of genes (involved in cytokine interactions) (Zhang et al, 2021) (Supplementary Table S4- 3). Among these, *IL6* is upregulated in ESCC (Soares-Lima et al, 2021) and was significantly hypomethylated within most of this study's samples. Several cell myeloid activation pathways were also hypomethylated, corroborating previous investigations whereby myeloid-derived suppressor cells (especially cluster of differentiation (CD)) were enriched in ESCC and linked to poor prognoses (Chen et al, 2014; Huang et al, 2015). This study found *CD53, CD86* and *CD200* to be hypomethylated within the tumour cohort. Concerning hypermethylated gene promoters, this study detected several gene promoters linked to the CNS/neural development pathways. Cells involved in the CNS are established as TME components by regulating tissue function and playing essential roles in cancer progression. The interaction between nerves and tumour cells is established in breast,

pancreatic, prostate, head and neck cancers and cholangiocarcinoma (Egashira et al, 2011; Kamiya et al, 2019; Magnon et al, 2013; Saloman et al, 2016).

Another important contribution of this study is the identification and use of enhancers as regulatory regions that could serve as clinical biomarkers. Several studies have shown that there is a stronger correlation between methylation of enhancers and gene expression compared to promoters (Andersson, 2015; Bell et al, 2016; Cho et al, 2022; Song et al, 2019). Despite enhancers inhabiting a much larger portion of genomic DNA than promoters and having a significant effect on cancer progression and prognosis (Aran and Hellman, 2013; Bell et al, 2016), the study of DNA methylation in enhancers has been underrepresented in most cancers, and no study to our knowledge have investigated the differential methylation of enhancers in ESCC. This lack of research could be due to DNA methylation assays with limited coverage up until recently with the introduction of the Infinium Methylation EPIC Array (850 K). The TruSeq Methyl Capture EPIC assay used in this study with 3,3 millions targets rectifies such issues (Moran et al, 2016).

In this study, we focused on DNA methylation combined with whole transcriptome profiling. Our findings highlight the prevalence of differentially methylated regions (DMRs) in promoter and enhancer regions that influence gene regulation and signalling pathways. The TruSeq EPIC assay outperformed the 850K bead array in detecting enhancer DMRs. This work demonstrates the importance of site-specific studies given the diverse geographic and etiologic factors that contribute to ESCC. Hypo-up and hyper-down candidate promoters and enhancers were mainly associated with ECM and contributed to ESCC carcinogenesis. In particular, the *KRT13* gene emerged as a prominent candidate that may warrant potential implications for ESCC therapy. A noteworthy aspect of this study is the focus on enhancers as potential clinical biomarkers. Enhancers are usually underrepresented in cancer studies, although it has been suggested that they show a strong or stronger inverse correlation between methylation and gene expression compared to promoters. This research leverages the comprehensive coverage of the TruSeq Methyl Capture EPIC assay to address this knowledge gap. Finally, this multi omics study serves as a foundation for understanding the development and progression of ESCC in the context of South African ESCC patients.

Chapter 5: Investigating the Transcriptome in African Esophageal Squamous Cell Carcinoma (Manuscript 3)

5.1 Abstract

Esophageal squamous cell carcinoma (ESCC) is a leading and aggressive cancer that is endemic in certain parts of Sub-Saharan Africa. To improve outcomes, there is an urgent need for the application of genomics to early detection. In this study, we used RNA sequencing on eleven matched ESCC tumour and adjacent normal tissue samples to investigate gene expression in South African ESCC patients. Transcriptome data mapped to 13 452 annotated genes between normal and tumour groups that clustered independently on principal component analysis (PCA). We investigated significant differentially expressed genes (DEGs) using criteria that included a fold change difference of ≥2 and an FDR-adjusted p value below ≤ 0.05. We identified 241 DEGs, of which 170 were upregulated and 71 were downregulated. Our top three overexpressed genes, *FN1, POSTN* and *COL6A3* (8-9-fold increase) had the most significant FDR ($<10^{-9}$ to 10^{-16}) values and all belonged to the extracellular matrix (ECM) pathway:. Our data suggest that the ECM plays an important role in the progression of ESCC in our cohort, facilitated by serine proteases.

5.2 Introduction

Cancer is the first or second leading cause of death in 112 of 183 countries for people aged ≥70 years. The latest global statistics show that esophageal cancer accounts for more than 540,000 deaths each year (Sung et al, 2021). There are two types of esophageal cancer: esophageal squamous cell carcinoma (ESCC), which begins in squamous cells in the upper and middle parts of the esophagus, and esophageal adenocarcinoma (EAC), which develops in the glandular tissue in the lower part of the esophagus (Mariette et al, 2005). ESCC is associated with geographically endemic regions, including China, Central Asia, Japan, Northern Iran, South America, African Americans and Sub-Sharan Africa (Alaouna et al, 2019). In Sub-Saharan Africa and central Asia, ESCC incidence is 20-fold higher than in Western countries (Somdyala et al, 2010). ESCC accounts for the second highest number of cancer-related deaths in South Africa, with an incidence of 46.7/100,000 and 19.2/100,000 for males and females respectively (Alaouna et al, 2019). The high mortality rate is mainly due to patients presenting late with the disease. Treatment for ESCC is largely palliative and includes insertion of a stent to enable swallowing (Loots et al, 2019). Although there are also radio- and chemotherapy options for these patients in South Africa, the median survival of ESCC patients is only 112 days (Ferndale et al, 2021a). The high incidence and poor prognosis of ESCC in South Africa underscores the need to better understand the aetiology, to identify early markers of disease and to implement novel therapeutic options. A number of risk factors have been associated with ESCC, including tobacco smoking (Segal et al, 1988), alcohol consumption (Sumeruk et al, 1992), smoke inhalation from cooking and heating fires (Sammon, 2007), poor nutrition (Kamangar et al, 2009), and unhygienic water sources (Yang, 1980). Homemade beer, which is regularly contaminated with fungi such as Fusarium verticillioides, has also been suggested as a contributing factor of ESCC in South Africa (Alaouna et al, 2019).

The recent progress in high throughput sequencing technology has led to genomic analyses (e.g., mutational profiling, transcriptome, methylome, etc.) becoming a fundamental source of experimental and clinical information. Although extensive genomic profiling of ESCC has been used for the classification of ESCC subtypes (Gao et al, 2014; Kim et al, 2017; Lin et al, 2014; Nagata et al, 2017; Sawada et al, 2016), an important and unanswered question is whether genomic data can provide information to improve our understanding of the aetiology of this deadly disease. A search of the literature for gene expression studies in ESCC is summarised in Table

5.1. We did not include any studies that investigated microRNA and lncRNA sequencing.

Table 5- 1: Gene expression studies on ESCC over the past decade

Cohort and reference	Number of samples	Technique	Upregulated gene/s	Downregulated Gene/s	Pathways
China (Fu et al, 2015)	3 paired tumour/normal	RNA Sequencing (Single ended)	*COL1A1, COL1A2, COL1A10, MMP*	*KRT4, CRISP3, CRISP4*	ECM, Focal adhesion, small cell lung cancer
China (L-Y Zhang et al, 2016)	3 paired tumour/normal, 218 validation cohort	RNA sequencing	-	*PSCA*	Cell cycle arrest and Differentiation
China (Xing et al, 2017)	6 paired tumour/normal	RNA Sequencing (Paired ended)	*ADAM12, CHI3L1, MMP13, SPP1*	Not taken into consideration	Degradation of ECM
China (Yu et al, 2019)	119 paired tumour/normal	RNA expression microarray and exome sequencing	*ANO1, GAL,* and *MMP3*	-	Degradation of ECM, TGF-beta pathway
India (Kashyap et al, 2009)	20 paired tumour/normal	DNA microarray candidates were confirmed with mRNA expression	*SPARC, MMP1, MMP3, MMP10, MMP9, MMP12, MMP13, MMP11, SPP1, OPN3, IL6, IL1B, THBS2, TACSTD1*	*SYNPO2, NMES1, TGM3, ZNF12, ZNF750, ZNF185, TMPRSS11E, SPINK5, SERPINB4, SCEL, RHCG, PPP1R3C, IVL, HOP, GPX3, PPL*	Degradation of the ECM, Epithelial formation, Arachidonic acid metabolic pathway, Tryptophan metabolic pathway
China (Z Liu et al, 2020, p. 17)	90 paired tumour/normal	RNA sequencing	*KRT17*	-	AKT pathway
China (Xu et al, 2020)	3 paired tumour/normal. The results validated against 53 samples from GEO database	RNA sequencing	*MMP1, MMP9, MMP12, MMP10, MMP13, MMP17, MMP23B*	-	Degradation of ECM
China (Zhang et al, 2020, p. 3)	155	RNA and exome sequencing	*TSTA3*	-	Cell migration

Cohort and reference	Number of samples	Technique	Upregulated gene/s	Downregulated Gene/s	Pathways
China (G Liu et al, 2020)	10 paired tumour/normal	RNA sequencing	*CLIC4*	*CLIC3*	Cell cycle, P53
TCGA Database (Zhou et al, 2022)	162 tumour and 11 normal	RNA Sequencing	*METTL3*	-	MAPK and immune response
Africa (Liu et al, n.d.)	59 paired tumour/normal	RNA Sequencing	*TP63, KRT5, KRT15, BNC1, DSC3, DSG3*	-	Neural differentiation, morphogenesis, DNA repair, metabolism
Japan (Takemoto et al, 2021)	88 paired Tumour/normal	RNA sequencing, exome, methylation (450k array)	-	*FAT1* and *FAT2*	transmembrane proteins and members of the cadherin superfamily.

One of the most notable observations from Table 5- 1 is that most expression studies on ESCC were carried out in Asia (10/12), with Africa severely under-represented. There is only a single African study that carried out exome and RNA sequencing in a cohort of Malawian ESCC patients (Liu *et al.*, 2016) despite the high prevalence of ESCC in sub-Sharan and Eastern Africa.

Another important observation from Table 5- 1 is that in almost half of the studies (5/12), DEGs are enriched in biological terms related to degradation of the ECM. The ECM is a component of the tumour microenvironment (TME) and is a dynamic network of proteins, glycoproteins, proteoglycans, and polysaccharides that provides signalling, stability and structural support for organs. Other components of the TME include the tumour's connective tissue, vasculature and infiltrating immune cells. There are three main constituents of the ECM: structural proteins, adhesive proteins and proteoglycans. The most common type of structural protein is collagen, which provides the matrix with strength and support. Adhesive proteins mediate interactions between cells or between cells and the ECM. There are four main families of adhesive proteins: immunoglobulin-like adhesion molecules, integrins, cadherins and selectins (Ren et al, 2011). Proteoglycans, the third major constituent of the ECM, are proteins that are usually attached to glycosaminoglycan polysaccharides and are responsible for cell adhesion, migration and proliferation (Wight et al, 1992). Excess ECM production or reduced ECM turnover is a prominent feature in several cancers (Frantz et al, 2010). Degradation of the ECM in cancer is facilitated through proteinases such as matrix metalloproteinases (MMP) and serine proteinases (Cawston and Young, 2010), which are released by tumour and stromal cells. In ESCC, degradation of the ECM is thought to occur as a result of a reduction in collagen deposition due to overexpression of the matrix metalloproteinase (MMP) family, which is made up of 23 isoforms (Bode and Maskos, 2003). Several studies have investigated the role of *MMPs* in ESCC. *MMP9, MMP12, MMP10, MMP13, MMP17*, and *MMP23B* enhance cell viability and cell migration in ESCC (M Liu et al, 2016). The MMP family is also linked with lymph node metastasis and advanced TNM stage of ESCC (M Liu et al, 2016). Studies have shown that *AJUBA* upregulates *MMP10*, which promotes cell migration and cell invasion (Shi et al, 2016) and is associated with poor survival in

early ESCC (Liu et al, 2012). *PTK7* activates *AP-1* and *NF-κB*, which promotes the expression of MMP9, increasing cell invasiveness in ESCC (Shin et al, 2016).

Despite these findings on the importance of ECM in ESCC, there are no genome-wide expression studies specific to the South African population. In the present report, we addressed this issue using RNAseq profiling of a cohort of South African ESCC patients with matched tumour/adjacent tissues.

5.3 Methods

Specimen Collection, Methods and Study Design

Specimens from patients with a histologically confirmed diagnosis of ESCC were collected and processed according to the IARC protocol (Mendy et al, 2017). Sample collection, storage, histology and nucleic acid extraction have already been described and reported (Ferndale, Moodley et al, 2021b). In brief, tumour and adjacent normal tissue were collected from 50 presenting patients suspected to have ESCC from CMJAH. Specimens were stored in RNAlater® (Qiagen) or placed in a 2 ml Cryotube and immediately immersed in liquid nitrogen. Each specimen was dissected into two pieces (25%:75%); the smaller portion was processed for histology to determine tumour content, and the larger portion was used for DNA and RNA extraction [Qiagen AllPrep Universal Kit ® (Qiagen)] according to the manufacturer's protocol. Paired samples with ≥60% tumour content and matched adjacent normal tissue (0% tumour) were the primary criteria for inclusion. The majority of patients in our cohort had 100% tumour content (7/11).

Patient characteristics are shown in Table 5- 2. All patient specimens were of self-declared African ancestry . The mean age of the patients was 60.81 years (ranging from 38 to 78). Our cohort included seven females and four males. Four patients consumed tobacco, and three consumed alcohol.

Table 5- 2: Patient characteristics of specimens processed in this study

*Sample number	Age	Gender	Tumour grade	Tumour %	Current smoker	Current alcohol consumption	Current fuel used for cooking
2	69	Female	Invasive Carcinoma	100	Never	Never	Wood
3	65	Female	Invasive Carcinoma	100	Never	Never	Wood
4	78	Male	Invasive Carcinoma	100	Never	Never	Wood
5	56	Male	Invasive Carcinoma	100	Never	Never	Electricity
6	67	Male	Invasive Carcinoma	75	No. Past smoker	Never	Paraffin
7	53	Female	Keratinising Carcinoma	80	Never	Never	Electricity. Wood in the past
8	55	Female	Keratinising Carcinoma	60	Yes. 7 per day	Yes. 750 ml twice per week	Electricity
9	66	Female	Keratinising Carcinoma	100	Never	Never	Wood
10	49	Male	Free lying Squamous	100	Yes. 5 per day	Yes. 1 litre twice per week	Wood
11	38	Female	Keratinising Carcinoma	90	Yes. 60 per day	Yes. 9 litres twice per week	Wood
12	73	Female	Keratinising Carcinoma	90	Never	Never	Electricity

*Please note: Sample number 1 failed QC for sequencing and therefore was not included for the rest of this study

RNA Sequencing

RNA sequencing and transcript normalisation

RNA was extracted from patient tissue with the AllPrep assay (Qiagen®). RIN was calculated using the Agilent 2100 BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA). Library preparation was performed with the TruSeq Stranded mRNA assay according to the manufacturer's guidelines. Briefly, 1 µg RNA was purified with poly-DT magnetic beads to capture mRNA. RNA was reverse transcribed into cDNA using Superscript VI® (Thermo Fisher). Second strand cDNA was generated to ensure strand specificity by substituting dTTP with dUTP. Double-stranded cDNA was processed for end repair, A-tailing, adaptor ligation and library enrichment. The quality and quantity of libraries were analysed with a Labchip GX® analyser (Perkin Elmer) and Qubit® assay (Thermo Fisher). RNA libraries were sequenced as 125 bp paired-end reads on the Illumina HiSeq2500 platform. Reads from transcriptome data were demultiplexed before analysis.

After initial QC with FastQC, transcripts were mapped to the reference genome (hg38) and quantified using the software package Hisat2 (Kim et al, 2019). Since all matching pairs were in the same batch, we performed a paired analysis. Count data for all samples were combined and normalised (based on library size) with EdgeR (Robinson et al, 2010) (version 4.1). The strategy used for differential analysis between tumour/normal samples was to aggregate p values derived from individual transcripts to obtain gene-level results (Yi et al, 2018). Traditionally, the detection of DEGs used gene-level read counts as starting material to test for differential gene expression between tumour and normal samples. However, recently, it has been shown that statistical testing is more powerful with the 'analysis first, aggregation second' approach in which p values derived from transcript analysis are first aggregated to obtain gene-level results. This has been shown to increase sensitivity and accuracy for DEG analysis (Yi et al, 2018). Thus, a p value was obtained for each transcript and aggregated to obtain a single p value for each corresponding gene. To account for individual effects, we used a linear regression with the tissue type (normal or tumour) as the main variable and the patient identification as the covariate. Genes with a fold change (up and down) ≥2 and FDR-adjusted p value ≤0.05 were considered significant.

5.4 Results

5.4.1 Transcriptome Profiling of South African ESCC

The number of reads that are mapped to a gene can be affected by different factors, such as gene length (Oshlack and Wakefield, 2009), sequencing depth (Robinson and Oshlack, 2010) and GC content (Risso et al, 2011). Thus, these factors need to be taken into consideration when converting raw reads into meaningful differential expression results. The number of reads after normalisation and filtering out reads with low coverage (>10) are shown in Table 5.3.

Table 5- 3: The number of reads after normalisation and filtering for all samples in our cohort

Group	Sample number	Read counts after normalization and filtering
Normal	2N	21976469
Tumour	2T	21808442
Normal	3N	20369913
Tumour	3T	13117084
Normal	4N	26040758
Tumour	4T	17409313
Normal	5N	21176982
Tumour	5T	20859610
Normal	6N	19141767
Tumour	6T	15421812
Normal	7N	2422961
Tumour	7T	1664606
Normal	8N	834498
Tumour	8T	1319364
Normal	9N	1123674
Tumour	9T	3341811
Normal	10N	1134526
Tumour	10T	5511597
Normal	11N	4306920
Tumour	11T	1432194
Normal	12N	2556530
Tumour	12T	1324877

From Table 5- 3, the mean read count after normalisation across all samples was 10195259, sufficient for further downstream analysis. A visual display of the read counts obtained after normalisation is shown in Figure 5-1.

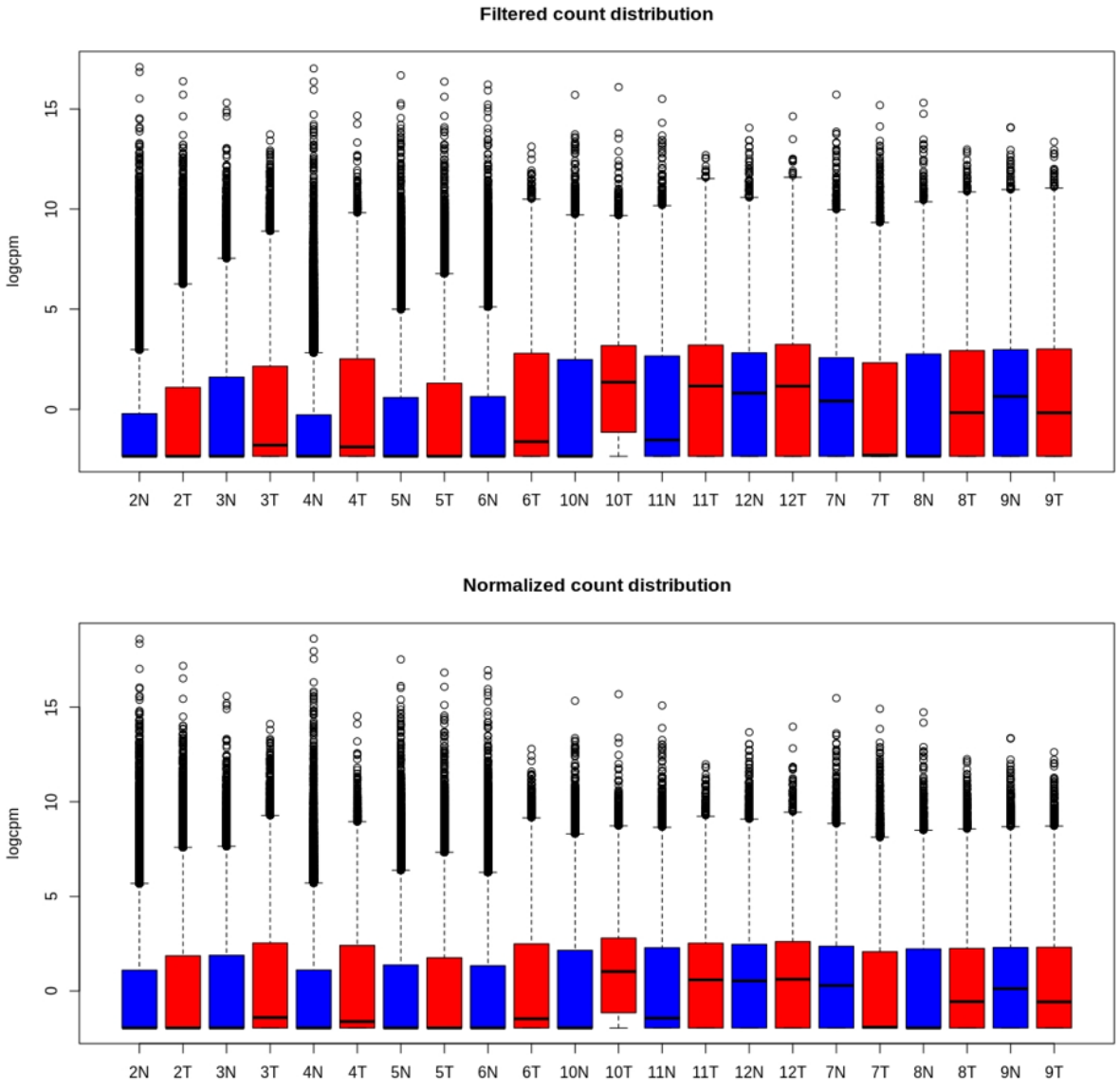


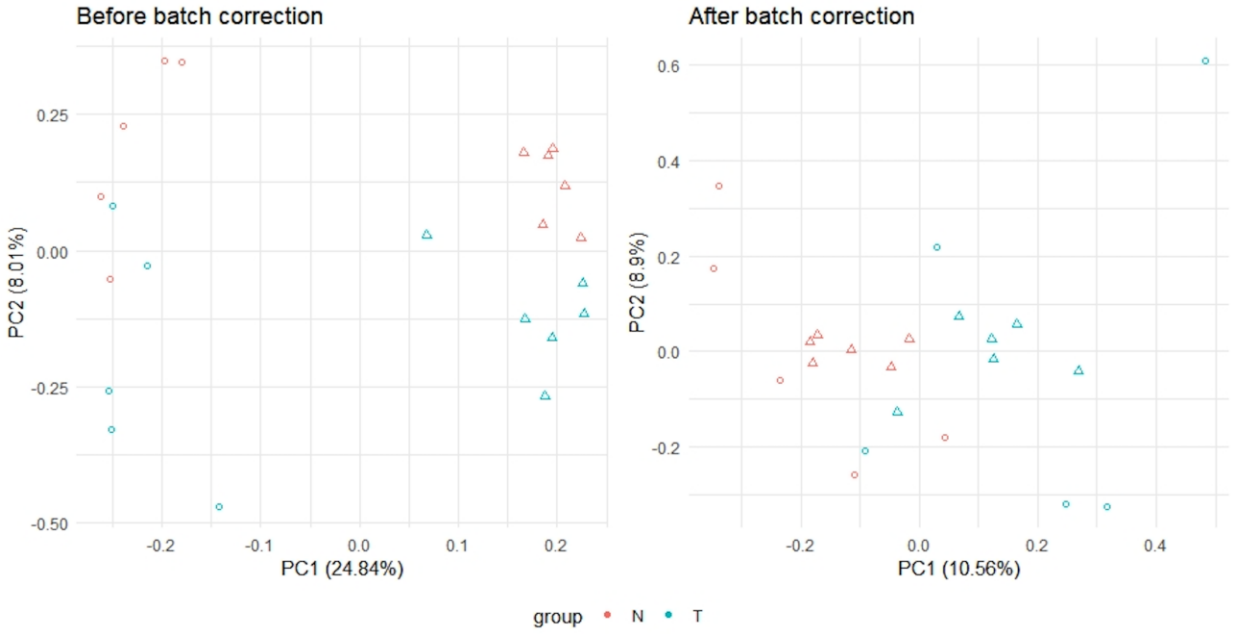
Figure 5- 1: Normalisation of RNA transcripts

Figure 5.1 displays read counts after normalisation measured in log2 counts per million. Each boxplot represents the median with standard deviation (top and bottom). The extremes of the boxes are the first and third quartiles. The dots are outliers (genes with several counts). Low coverage was defined using the "filterByExpr" function of the "edgeR" package, with a minimum number of reads ≥10 (using the "min.count" function).

The difference between both batches (tumour and normal) was significantly reduced when filtering out transcripts with low coverage (<10x). PCA showed that most of the ESCC samples clustered together (Figure 5- 2 A) and were sufficient to distinguish

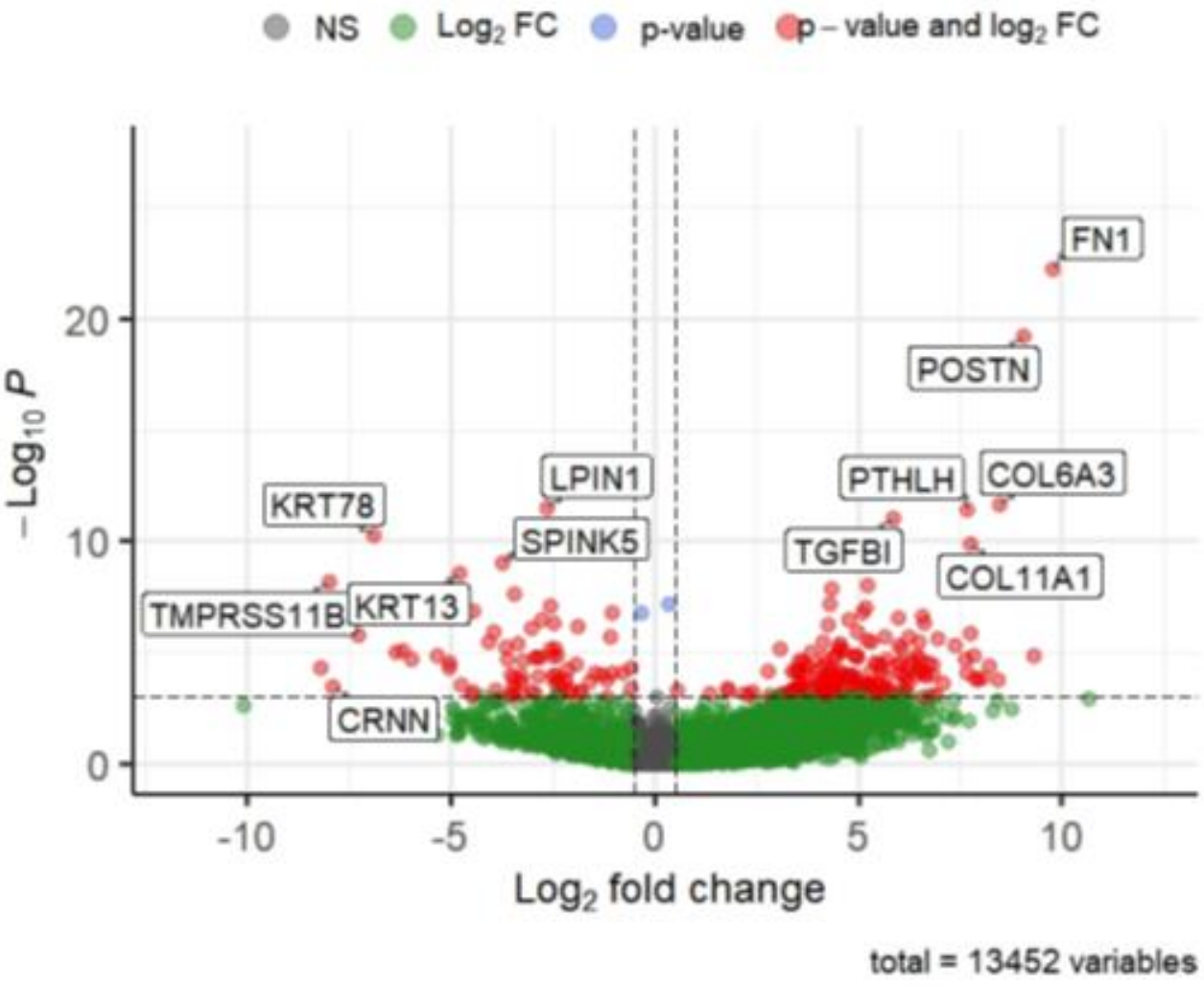
tumour from normal samples. Differential analysis was performed by aggregating p values derived from the transcript-level to obtain gene-level results, as previously described (Yi et al, 2018). The volcano plot highlights all significant transcripts, using a FC => 2 and adjusted p value (FDR) ≤ 0.05 (Figure 5.2B).

A



B **Volcano plot : DEGs**

Tumor vs. Normal



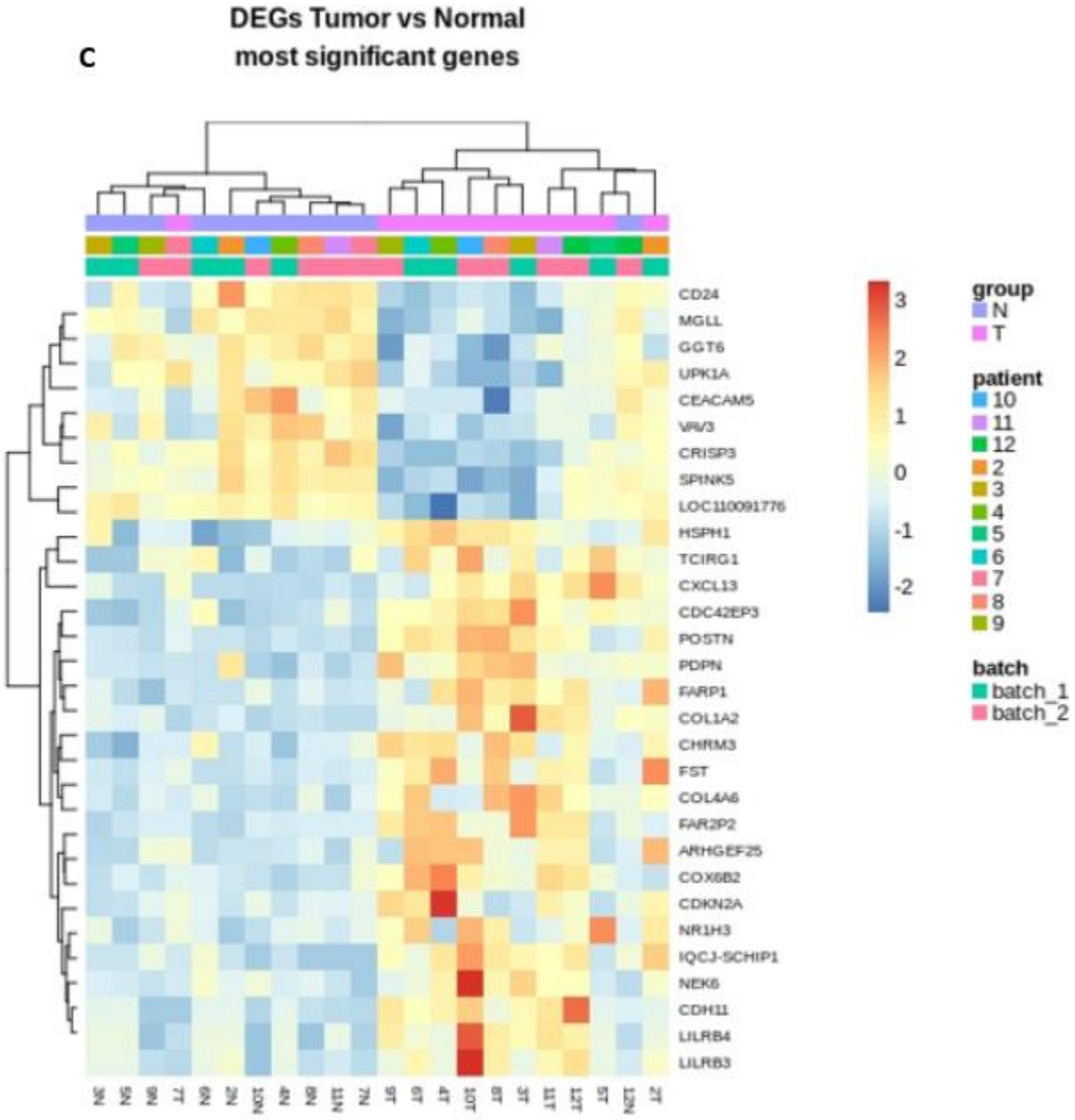


Figure 5-2: PCA, volcano and heatmap between tumour and normal samples A, PCA plot of tumour (blue) and normal (red) tissues from 11 South African samples. Before batch correction accounts for all transcripts regardless of their coverage after RNA sequencing, after batch correction filters out transcripts with low coverage (>10x). Circle: samples sequenced from batch 1, triangle: samples that were sequenced from batch 2. B, Volcano plot of DEGs. Statistical significance P values (Y-axis) versus fold change on x-axis. The most upregulated genes in tumours are towards the right, and most downregulated genes in tumours are towards the left. The horizontal dashed line around 2.5 represents log 0.003 (p value) corresponding to the corrected FDR of 0.05, which was the level of significance

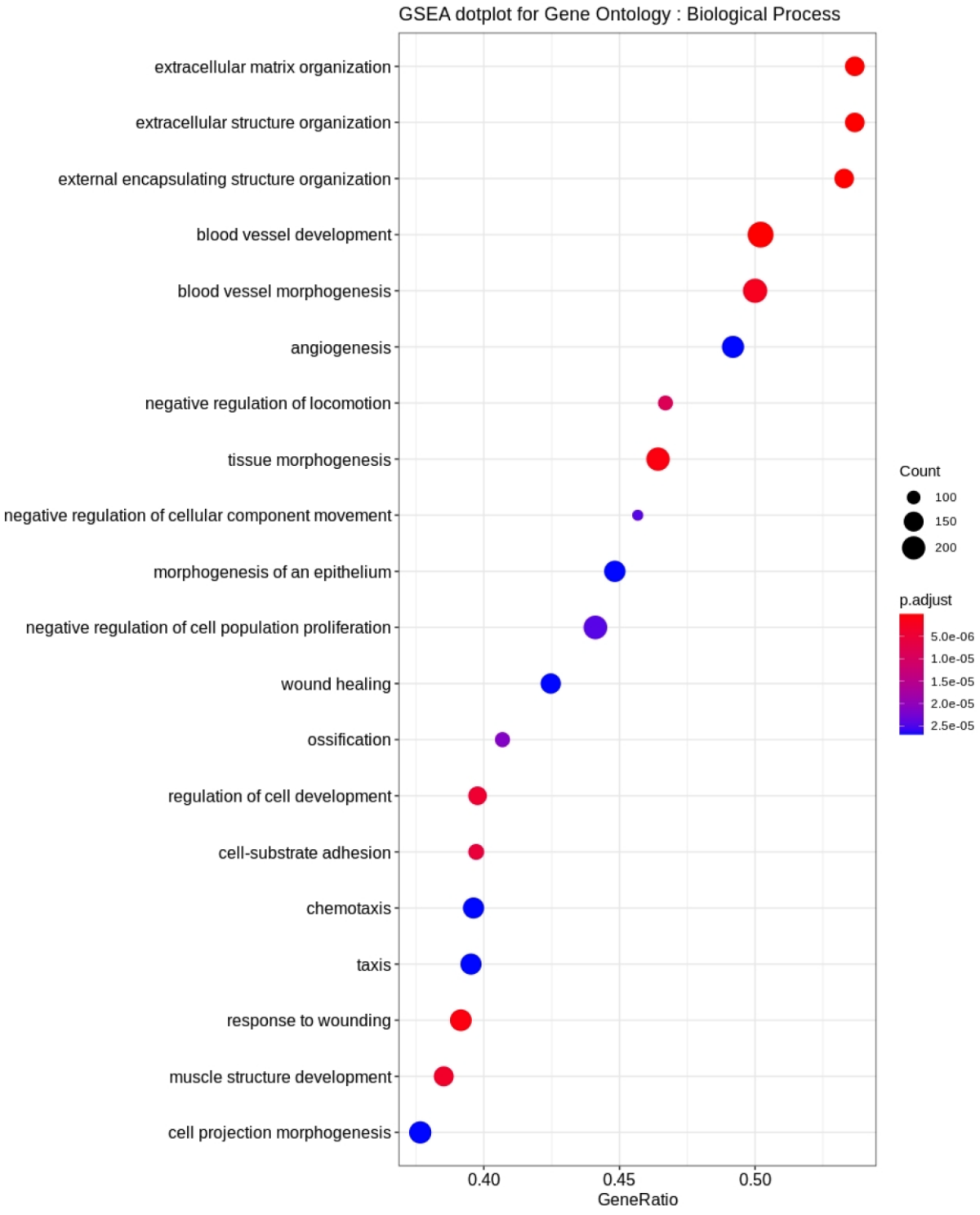
used for differential gene expression.C, Heatmap of the top 30 DEGs between normal and tumour samples. The dendrograms are based on unsupervised clustering of genes (rows) and samples (columns).

From the volcano plot (Figure 5-2 B), a total of 13 452 genes were compared between our normal and tumour groups. Some of the significant genes that had both a FC ≥ 2 in expression and an FDR ≤ 0.05 are highlighted in the volcano plot. Genes towards the upper-left were downregulated, and genes towards the upper right were upregulated with significant p values. We observed that most of the genes that were upregulated with significant p values belonged to the ECM pathway. In total, there were 241 DEGs, with 71 downregulated and 170 upregulated in tumours relative to their nontumour adjacent tissues, which was consistent across all samples (supplementary table 5.1). The top 30 DEGs were able to correctly discriminate between tumour and nontumour tissues, as shown in a heatmap of the normalised expression data (Figure 5.2 C).

5.4.2 Gene Set Enrichment Analysis in South African ESCC

To further investigate the differences between ESCC and normal tissue, GSEA was used to investigate gene expression patterns and pathways between ESCC and normal tissue. GSEA has been shown to be a powerful method in omics studies that involve cancer (Zito et al, 2021). GSEA is a computational process that is used to determine if a set of genes can provide an understanding of the BPs or pathways that underlie a phenotype. GSEA showed that the ECM pathway was consistently enriched across the different databases, and ECM was the most enriched pathway across BPs and Reactome (Figure 5-3). ECM has been shown to be enriched in several ESCC studies, as described above. Some of the most enriched pathways from KEGG included Epstein–Barr virus (EBV), proteoglycans and MAPK signalling pathways, which have also been previously described and associated with ESCC (Mabi, 2017; S Zheng et al, 2011; Zhu and Cheung, 2021) (Supplementary Figure S5- 1).

A


GSEA dotplot for Gene Ontology : Biological Process
extracellular matrix organization
extracellular structure organization
external encapsulating structure organization
blood vessel development
blood vessel morphogenesis
angiogenesis
negative regulation of locomotion
tissue morphogenesis
negative regulation of cellular component movement
morphogenesis of an epithelium
negative regulation of cell population proliferation
wound healing
ossification
regulation of cell development
cell-substrate adhesion
chemotaxis
taxis
response to wounding
muscle structure development
cell projection morphogenesis
0.40
0.45
0.50
GeneRatio
Count
100
150
200
p.adjust
5.0e-06
1.0e-05
1.5e-05
2.0e-05
2.5e-05

B

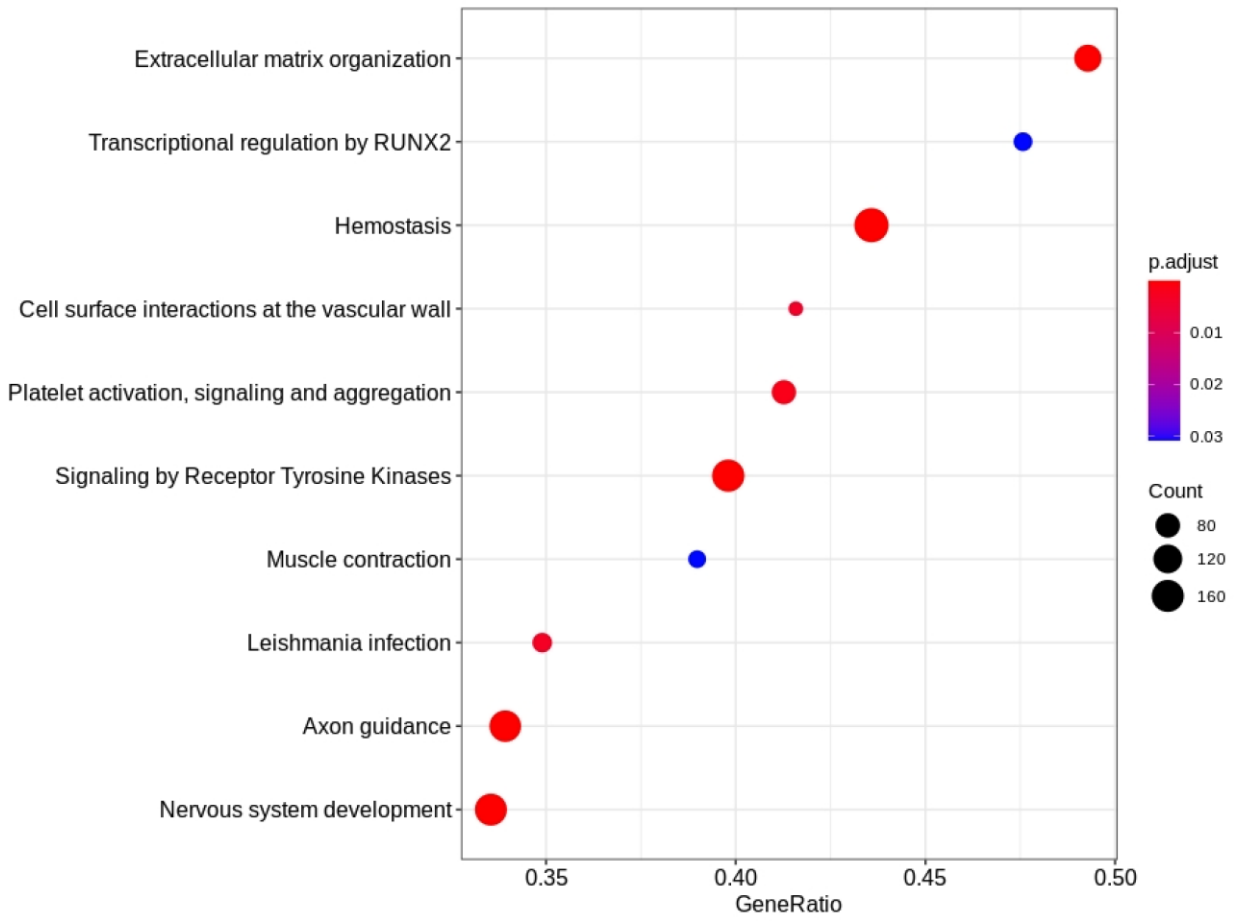
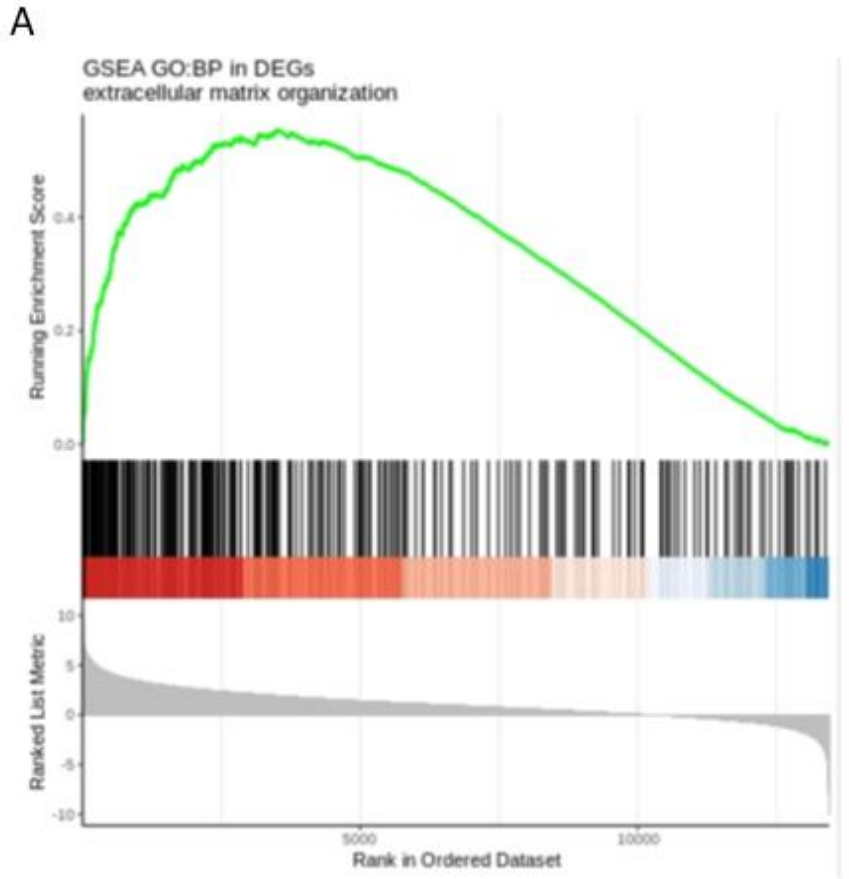


Figure 5-3: Pathways enriched after transcriptome analysis with BPs (A) and Reactome (B). "Count" = number of genes

Most gene ontology tools only investigate genes that have a significant change, for example, p values ≤ 0.05. In addition to significant genes, GSEA also looks at changes in all genes across a particular pathway. GSEA focuses on a ranking system and determines if a given gene set has statistical significance between both groups (Zito et al, 2021). The enrichment plot from GSEA showed a significant enrichment score for ECM (Figure 5-4A).



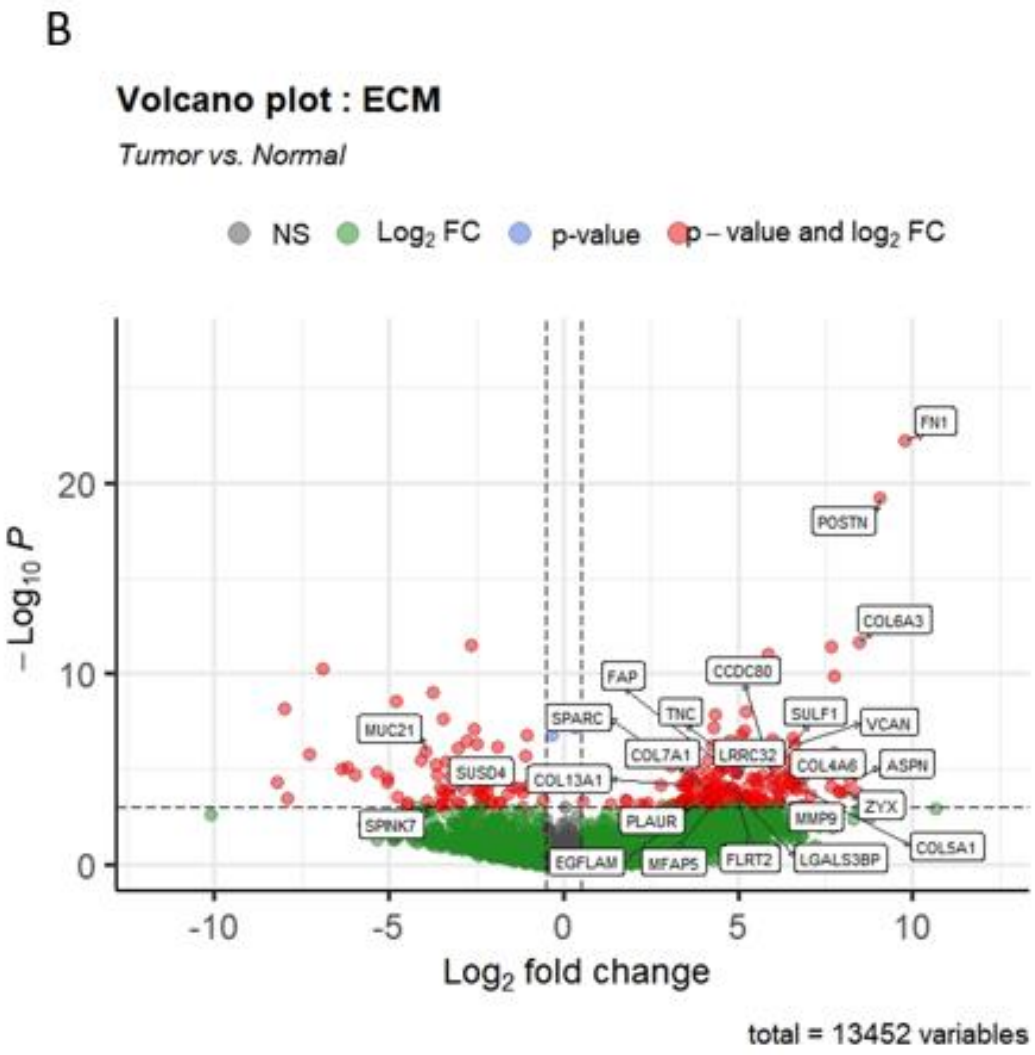


Figure 5-4: GSEA and Volcano plots illustrating over representation of the ECM in our cohort A, Along the X-axis is the ranked gene set (ECM) with genes ranked from left to right. On the left are genes with the highest expression and on the right are genes with the lowest expression when comparing our tumour to normal group. Red: high expression, Blue: low expression. Each black bar represents a gene in the ECM pathway. B, Volcano plot of all significant genes that belong to the ECM pathway among all 13452 genes expressed. All labelled dots (genes) belong to the ECM pathway. The position of each gene is based on both the fold change (X-axis) and Log P-value (Y-axis). NS: no significance in expression or p-value. The top 3 genes that were overexpressed and had the most significant p-values were FN1, POSTN and COL6A3 (upper right).

From the GSEA plot in Figure 5.4 A, most of the bars are located on the left-hand side and thus were overexpressed in the ECM pathway. ECM was the top most enriched GO pathway (NES= 1.7, adjusted p value = 3.35E-08). Since ECM was highlighted in most of

our pathway analyses, we sought to determine how many genes in the ECM pathway displayed both a FC > 2 in expression (up or down) and a significant FDR (≤ 0.05) (Figure 5.4 B). We identified 25 genes that belonged to this category. Incidentally, the top three overexpressed genes that also had the most significant p values across all genes belonged to the ECM pathway: *FN1, POSTN* and *COL6A3* (also shown in Figure 5.4 B). All three genes had an 8-9-Log_2 FC increase in expression in tumour tissue compared to normal tissue and very significant FDRs that ranged between $>10^{-9}$ and 10^{-16} (Supplementary table 5.1).

5.5 Discussion

Over the past decade, transcriptome studies have been used extensively to understand the underlying causes of cancer and to identify novel biomarkers, gene signatures and aberrant expression. NGS has facilitated these discoveries and consequently has paved the way for anticancer therapies (Supplitt et al, 2021). In our study, we compared gene expression between matched tumour/normal tissues derived from ESCC patients. We identified previously identified biomarkers and novel candidates. The data generated from this study were focused on providing more insight into RNA sequencing when comparing normal and tumour tissue in ESCC, a deadly and high incidence disease in Sub-Saharan Africa, a region that is also severely lacking in genomic studies.

The PCA plot showed that although the intracluster variation was broad, most of the ESCC and normal tissues independently clustered together. The list of significantly expressed genes (2-fold change in expression and FDR ≤ 0.05) included some that were previously reported: *SPINK5*, *SPP1*, *SCEL*, *RHCG*, *ZNF185*, *COL1A2*, *KRT17*, *SPARC*, *CRISP3*, *KRT4*, *MMP9*, *SYNPO2L*, *FN1*, *COL6A3* and *POSTN* (Kashyap et al, 2009; Xing et al, 2017; Fu et al, 2015; Z Liu et al, 2020; Fu et al, 2015; Samantaray *et al.*, 2004; Xu *et al.*, 2020; Zhang et al, 2005; Nanda et al, 2004; Wang et al, 2014). Most of the significantly overexpressed genes detected in our study (8/15) that were also previously reported in one specific study that was carried out in a cohort of Indian (Kashyap et al, 2009) ESCC patients. Importantly, from our list of 241 genes that had a ≥2-fold change in expression and an FDR ≤0.05, a significant number of genes (15%) were associated with the ECM (25) and serine proteases (11).

5.5.1 ECM Genes are Differentially Expressed in South African ESCC

ECM was the most consistent biological term enriched in our pathway analysis. Cancer progression is strongly correlated with the interaction between tumour cells and the tumour microenvironment (Walker et al, 2018). Of the significant genes that exhibited a 2-fold change in gene expression and had FDR values ≤ 0.05, 3 were underexpressed, and the remaining 22 were overexpressed. The majority of the genes (9) were involved in cell adhesion, six were involved in structural support, four were involved in signalling, two were involved in degradation of the ECM, and one gene each was involved in cell migration, binding of collagen and serine proteinase inhibitor.

Disruption of the ECM and proteinases has been implicated in cancer progression. Extracellular proteases promote cell invasion through ECM remodelling (Rowe and Weiss, 2009). Serine proteases are thought to disrupt critical signalling pathways such as growth factors (Yang et al, 2008), chemokines (Naito et al, 2019), cytokines (Leuning et al, 2018) and kinases (Duffy, 2004). A substantial amount of ESCC studies listed in table 5-1 have attributed ECM dysregulation to MMP proteases. In this study, although *MMP9* was a significant overexpressed gene that exhibited a 2-fold change in gene expression and had FDR values ≤ 0.05, several aberrantly expressed genes enriched for serine proteases, not MMPs, contribute to disruption of the ECM. Although serine proteases, including urokinase-type plasminogen activator, fibroblast activation protein (FAP), granzymes, and kallikrein-related peptidases, have all been reported to have increased expression levels and abnormal activity in the tumour microenvironment in several other cancers (Duffy, 2004; Mj et al, 1994; Y Liu et al, 2020; Zhou et al, 2020; Seiz et al, 2012). One of the genes that could contribute significantly to ECM degradation in our study was the serine protease inhibitor Kazal type (*SPINK7*), which was found to be significantly underexpressed (~4-fold) with an FDR ≤ 0.05. *SPINK7* belongs to the SPINK family, which includes *SPINK1-14* and is the largest branch in the serine protease inhibitor family (Jalkanen et al, 2006; Ma et al, 2013, p. 13; Rawlings et al, 2004; Wapenaar et al, 2007). *SPINK7* is also known as esophageal cancer-related gene 2 (*ECRG2*) and has been well described to be significantly downregulated in primary ESCC (Cui et al, 2010). The same observation was made in our study.

In addition to detecting several serine proteases that were aberrantly expressed, we also looked at the most highly expressed genes and genes with the most significant p values (≤ 0.05). The top 3 genes with the most significant p values were fibronectin-1 (*FN1*), periostin (*POSTN*) and collagen type VI alpha 3 chain (*COL6A3*). Incidentally, the same three genes also displayed the highest fold change in expression with a 9-fold difference between ESCC tumour and adjacent normal tissue (supplementary table 5.1). High expression of *FN1* has been reported in several cancers. In colon cancer, *FN1* promotes cell invasion (Ding et al, 2008). In breast cancer, *FN1* promotes progression by activating *STAT3* (Balanis et al, 2013). High levels of *FN1* were reported in renal cancer with advanced disease (Waalkes et al, 2010). In advanced ESCC, *FN1* was shown to activate the Erk pathway (Zhang et al, 2005). High levels of POSTN protein have also been previously reported in ESCC and were also shown to be associated with tumour progression, angiogenesis and poor survival (Wang et al, 2014). High levels of *COL6A3* have been reported in several cancers, including breast (Iyengar et al, 2005), liver (Lee et al, 2019), lung (Nanda et al, 2004), pancreatic (Arafat et al, 2011) and bladder (Huang et al, 2018) carcinomas. In ESCC, it has been suggested that high levels of *COL6A3* interact with *TEM8* during angiogenesis (Nanda et al, 2004).

The top six underexpressed genes that were downregulated with ≥ 6-fold change were *CAPN14, TMPRSS11B, CRNN, DYNAP, KRT78, PADI1*, and *CLCA4*. Down regulation of most of these genes have been shown to play a critical role in cancer progression. *CAPN14*, a protease, has been shown to be underexpressed in patients with eosinophilic esophagitis (Litosh et al, 2017). Downregulation of *CRNN* gene expression has been previously described in ESCC, in association with advanced tumour stage and poor survival in ESCC patients (Hsu et al, 2014). *TMPRSS11B, CLCA4 and KRT78* have also previously been shown to be downregulated in African-American ESCC patients. These genes affect the NRF2 pathway, which leads to dysregulation of the stress response and detox networks (Erkizan et al, 2017).

In summary, we present here for the first time a comprehensive transcriptome data set derived from tumour/normal paired in South African ESCC patients. We generated a substantial list of dysregulated genes that were enriched in the ECM pathway, notably,

serine proteases. These data provide insight into the molecular mechanisms and pathogenesis of ESCC in a cohort that is underrepresented in genomic studies.

Raw data for RNA sequencing in the form of count files can be downloaded using the link in the supplementary section (Supplementary folder 2.2).

Chapter 6: Discussion and Conclusion

Our study aimed to evaluate the contribution of methylome dysregulation and altered gene expression in African ESCC patients and to identify the genes and molecular pathways affected in our cohort.

6.1 Objective 1

In contrast to higher resource settings, South African and other African countries encounter several challenges regarding infrastructure, procuring tissue and human resources necessary for biobanking. Consequently, obtaining ESCC samples is challenging. First, we established a protocol to obtain good quality samples for genomic research in a resource limiting setup that only requires RNAlater, a standard refrigerator and a -20°C freezer. The workflow we provide is practical and cost effective and could also be applied to other types of tumour tissues.

6.2 Objective 2

Carcinogenesis is usually accompanied with genome-wide methylation changes (Pfeifer, 2018). After obtaining good quality nucleic acid for methylome and transcriptome sequencing, we attempted to identify whether DNA methylation changes across the genome were associated with transcriptome wide gene expression changes. The genetic contribution of methylome dysregulation in African ESCC is limited. Only one other cancer study investigating cervical carcinoma utilized methylome and transcriptome data in an African cohort (Gagliardi et al, 2020).

It is possible to sequence all 28 million CpG sites in the human genome using whole-genome bisulfite sequencing (WGBS). However, this method is costly, laborious and not practical (Suzuki et al, 2018). At the time this study was initiated, several previous and current studies employed the 450k or 850k microarray assay from Illumina for ESCC methylation profiling (Kishino et al, 2016; Lu et al, 2019; Peng et al, 2016; Talukdar et al, 2021). Both kits target ~ 450 000 or 850 000 CpG sites respectively. Since microarrays only target a set of predefined sequences and also taking onto consideration that African

populations have the highest levels of genetic diversity and phenotypic variations across all populations, we chose a sequencing approach to integrate methylation status across DMRs between our normal and tumour groups. NGS is more flexible than microarrays and is independent of previously selected targets, thus, allowing for higher discovery potential. Indeed, in our cohort we compared the 850k array to the TruSeq EPIC assay and found that only 257 CpG sites from the 850K array fell within the 1163 enhancer DMRs identified with the TruSeq EPIC assay. After sequencing, we had to choose a bioinformatics pipeline to determine methylation status and assign DMRs. Several methods exist and each has advantages and disadvantages. We opted for a practical method to analyse methylation calling based on the number of CpGs across a DMR and the depth of coverage of ≥10x. This is the recommended depth of coverage for methylation sequencing based on analysis of several high coverage reference data sets (Ziller et al, 2015) and recommendations provided from the supplier of the assay ("TruSeq Methyl Capture EPIC Library Prep Kit Documentation," 2022.). Furthermore, it was also shown that at this depth there is a strong correlation between sequencing and arrays regardless of the mapping tool used (Silva et al, 2021).

For our analysis, we first generated methylation calling of each individual CpG across our two groups. Differential methylation across the tumour methylome revealed predominantly hypomethylated CpGs (67%) in tumour samples compared to hypermethylated CpGs, a ubiquitous feature of carcinogenesis (Ehrlich, 2009). PCA analysis showed us that although the intra cluster variation was broad, most of the ESCC and normal tissues clustered separately After generating methylation calls, we inferred DMRs between our two groups using MethylKit®. Methylkit segments the methylome based on change-point analysis. We looked at DMRs distributed across gene features and most DMRs were located in promoter and intergenic regions. The proportion of hypermethylated promoters was highest in promoters (40%), in accordance with site specific methylation being a common feature across the tumour methylome (Nishiyama and Nakanishi, 2021b). We then identified DMRs such as promoters and enhancers that inversely correlated with gene expression. A key finding in our study was the identification of one hyper-down promoter (*KRT13*), two Hypo-up promoters (*CDC42EP3* and *TNC*) and one Hypo-up enhancer (*COL6A1*) all of which belong to the ECM.

The ECM is a 3D structure that provides structural support and participates in range of activities from cell adhesion, migration, proliferation to cell differentiation and cell death (Lu et al, 2012). During cancer or other pathological processes ECM dynamics are dysregulated (Lu et al, 2012) through degradation, hyperproduction or alteration (Naba et al, 2016). Components in ECM remodelling and degradation include: proteoglycans, glycoproteins, fibrillar collagens, elastin, fibronectin, laminin and nidogen (Daley et al, 2008; Vakonakis and Campbell, 2007). Remodelling of the ECM can involve removal of one of its components. Fibrillar collagens and elastin contribute to tensile strength and viscoelasticity of tissue, while the other proteins are involved in building the matrix network (Daley et al, 2008; Vakonakis and Campbell, 2007). The major enzyme in ECM remodelling belongs to the metalloproteinases which is made up of two main families: matrix metalloproteinase (MMP) and a disintegrin and metalloproteinase with thrombospondin motifs (ADAMTS) (Cawston and Young, 2010).

KRT13 is an intermediate filament that belongs to a group of tough fibrous proteins found on the moist lining (mucosae) of the esophagus, mouth, genitals, nose and anus (Moll et al, 1982). Intermediate filaments protect the mucosae from frictional damage or any other daily physical stress. Several studies in bladder cancer have shown downregulation of KRT13 in advanced stage and high-grade bladder cancer (Celis et al, 1996; Moll et al, 1982; Schaafsma et al, 1990; Southgate et al, 1999; Worst et al, 2014). A later study showed that hypermethylation of *KRT13* leads to downregulation of the gene in aggressive bladder cancer (Marsit et al, 2010). It has been suggested that this correlation falls in line with previous studies in which there is a dedifferentiation of epithelial tumour cells into an amesenchymal (EMT) phenotype. During EMT transition, cells are more likely to invade and metastasise surrounding tissue as a result of a higher grade of plasticity (Craene and Berx, 2013; Dave et al, 2012; Worst et al, 2014). KRT13 works together with KRT4 to form intermediate filaments essential for structural support for the mucosae layer (Shibuya et al, 2003). In our study both *KRT13* and *KRT4* were also downregulated (FDR ≤0.05 and a ≥2-fold change in gene expression). White Sponge Nevus (WSN) disease is an autosomal dominant disease characterised by mutations in *KRT4* and *KRT13* (Qiao et al, 2022). As a result, this leads to the production of damaged intermediate filaments, inflammation and abnormal proliferation and division of epithelial

cells (Qiao et al, 2022). Histologically WSN is characterised by the keratinisation of surface cells and it has been suggested that lesions derived from other congenital diseases including squamous cell carcinomas exhibit the same appearance as WSN. Since half of our cohort displayed keratinising carcinoma (Table 4- 1), which could be attributed to a lack of intermediate filaments as a result of hypermethylation and downregulation of *KRT13* and downregulation of *KRT4*.

Cancer-associated fibroblasts (CAFs) are spindle-shaped cells that build up and remodel the ECM structure and is a dominant component of the tumour stroma (Kalluri, 2016). The functions of CAFs include matrix remodelling, invasion, angiogenesis and promoting tumour growth (DuFort et al, 2011; Gaggioli et al, 2007). In our cohort *Cdc42Ep3* showed a Hypo-up status. *Cdc42EP3*, also known as *BORG2*, has been shown to induce CAFs by coordinating actin and septin networks (Calvo et al, 2015). Without the expression of *Cdc42EP3*, CAFs are unable to remodel the ECM resulting in reduced matrix stiffness and cancer cell invasion (Calvo et al, 2015). It is also important to note that the *FAP* gene, a type-II transmembrane serine protease, is almost exclusively expressed in pathological conditions including cancer, arthritis and fibrosis (Fitzgerald and Weiner, 2020). *FAP* was one of our candidate genes that was highly overexpressed with a 5-fold increase and a significant p value (<0.05) in our tumour ESCC samples compared to matched normal tissue. *FAP* expression results in poor clinical outcomes through its effects on remodelling of the ECM. *FAP* also affects angiogenesis, epithelial-to-mesenchymal transition, immunosuppression and regulation of intracellular signalling. Since over expression of both *FAP* and *Cdc42EP3* affect matrix remodelling, these two genes overexpression may severely impact the ECM.

TNC was another Hypo-up gene identified in our study. Tenascin-C (*TNC*) is a ECM glycoprotein that is highly expressed in the tumour microenvironment (TME) and promotes proliferation, invasion and tumour cell survival (Sun et al, 2018). *TNC* expression is limited in normal adult tissue, however, *TNC* is re-expressed in several diseases including lesions that undergo remodelling, tissue repair and cancers. *TNC* expression is associated with a poor patient survival prognosis in several cancer types (Midwood et al, 2016; Orend and Chiquet-Ehrismann, 2006). Under normal

circumstances, expression of *TNC* expression is tightly controlled. However, in cancer, *TNC* is highly expressed around blood vessels, in matrix fibres and orchestrating an immunosuppressive tumour microenvironment and promoting metastasis (Yilmaz et al, 2022). In glioblastoma, high TNC levels were found to be associated with tissue stiffness (Miroshnikova et al, 2016). In ESCC Yang et al showed that *TNC* expression was highly expressed in their cohort of 136 ESCC samples. Furthermore, *TNC* expression also correlated with expression of four CAF markers: *PDGFRα, PDGFRβ, SMA, FAP* (Yang et al, 2016). We also found over expression of *FAP* in our cohort.

Taking advantage of our high resolution methylome assay, we were able to interrogatemethylation levels of enhancers between our two groups. The only enhancer that fit our criteria and showed significant methylation variations (≥20%) with q-value ≤0.01 and that also had an inverse relationship to RNA expression (FDR ≤0.05 and a ≥2-fold change in gene expression) was *COL6A3*. Collagen is made up of 3 alpha chains: alpha-1 (VI) and alpha-2 (VI) chains, the alpha-3 (VI). *COL6A3* encodes for the largest of these alpha chains, the alpha-3 (VI) chain which has more domains compared to the other alpha chains (Zanussi et al, 1992). These domains organise matrix components (Zanussi et al, 1992). *COL6A3* is also highly expressed in several malignancies (Wang and Pan, 2020) including ESCC (Nanda et al, 2004). There have been two proposed mechanisms of overexpression of *COL6A3* in cancers: alternate splicing (Arafat et al, 2011) and aberrant methylation (Hyman et al, 2015; Maekawa et al, 2013). In uterine leiomyoma and glioblastoma, *COL6A3* was shown to be hypomethylated and over expressed (Hyman et al, 2015; Maekawa et al, 2013). It has been suggested that these findings may have implications for tumour progression and may also have diagnostic and prognostic potential in cancer (Wang and Pan, 2020). Collagen provides strength, regulates cell adhesion and is the most abundant fibrous protein within the interstitial ECM constituting up to 30% of the total protein mass (Rozario and DeSimone, 2010). An extensive study that used expression data from 275 ESCC cases from the GEO and TCGA databases identified 22 collagen family genes that were highly over or under expressed. The authors suggested a 7-gene panel to predict the prognosis in ESCC patients as this was shown to better predict ESCC patient survival compared to the TNM staging model (Li et al, 2019). Two of these genes overlapped with our cohort: COL10A1, COL11A1. We also

found additional collagen family member genes that were significantly highly expressed (FDR ≤0.05 and a ≥2-fold change in gene expression). These collagen family members include: *COL12A1, COL13A1, COL16A1, COL1A2, COL4A6, COL5A1, COL6A3, COL7A1* and *COX6B2*.

Taken together our findings suggest that the inverse relationship between hypomethylation and gene expression of two promoters (*Cdc42Ep3* and *TNC*) and one enhancer (*COL6A3*) all point to remodelling of the ECM, that may promote tumour invasion and angiogenesis. This coupled with Hyper-down *KRT13* and concomitant loss of protection of mucosae lining in the esophagus, all result in dynamic deregulated and disorganised changes to the ECM which could be the main molecular driving force in our African ESCC cohort. Unlike other signalling pathways previously described to be enriched or disrupted in ESCC (described in Chapter 1), our transcriptome data heavily points point dysregulation of the ECM with most of the genes belonging to the families of serine proteases and collagen members responsible for degradation of the ECM. Mutations in the tumour suppressor gene TP53 are seen in a variety of gastrointestinal cancers, including ESCC (Sankalecha et al, 2017). TP53 mutations are one of the most common aberrations in ESCC, occur in the early phases, throughout all stages and progression of ESCC and are thought to be one of the main drivers of ESCC (Zhong et al, 2023). There has been some link between TP53 mutations and the ECM. Cancers with TP53 mutations have shown to have an enhanced production of ECM proteins, such as collagens and fibronectin, prompting a tumour-supportive microenvironment (Procopio et al, 2015). Additionally, a recent study in pancreatic ductal adenocarcinoma revealed a high correlation between cancer aggressiveness and interaction among driver mutations like TP53, the ECM, and tumour acidosis. Consequently, it is suggested that treatments that target acidosis should be taken into consideration for upcoming treatment (Czaplinska et al, 2023). It will be interesting to determine the mutational status of TP53 in our cohort. This observation may imply that TP53 mutations are driving ESCC in African patients in conjunction with the ECM.

6.3 Proposed Model for ECM Remodeling in ESCC Through Aberrant Methylation and Therapeutic Implications

Taking our findings into consideration, we propose that aberrant methylation of key ECM-related genes (e.g., *KRT13*, *Cdc42EP3*, *TNC*, and *COL6A3*) significantly contributes to ECM remodeling in ESCC, ultimately promoting tumor invasion and progression. Specifically, the hypermethylation and downregulation of *KRT13* could weaken the esophageal mucosal barrier, enhancing epithelial-to-mesenchymal transition (EMT) and increasing the invasive potential of tumor cells. Conversely, hypomethylation and upregulation of *Cdc42EP3* may activate cancer-associated fibroblasts (CAFs), leading to increased ECM stiffness and facilitating cancer cell invasion. Similarly, the upregulation of *TNC* and *COL6A3* could disrupt ECM homeostasis by promoting matrix rigidity and angiogenesis, fostering a pro-tumor microenvironment. These findings suggest that targeting ECM components and pathways, such as CAF activation, MMP activity, and collagen deposition, offers a therapeutic opportunity to mitigate ECM-driven ESCC progression, particularly when combined with conventional treatments targeting known mutations like *TP53*.

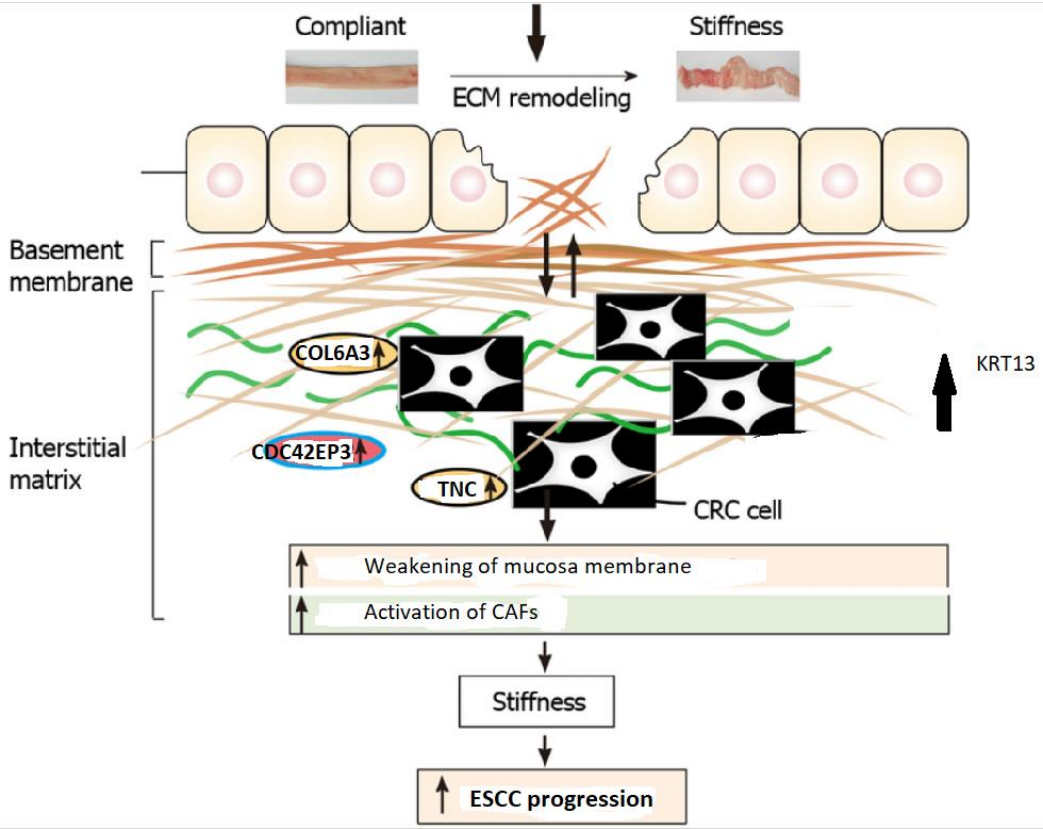


Figure 6-1: The figure illustrates the proposed model for ECM remodeling in ESCC. This model is driven by aberrant methylation of key ECM-related genes (*KRT13*, *Cdc42EP3*, *TNC*, and *COL6A3*). It highlights the effects of hypermethylation and hypomethylation events on gene expression and subsequent ECM changes, such as loss of mucosal protection, activation of cancer-associated fibroblasts (CAFs), increased matrix stiffness.

6.4 Limitations of the Study

The limitation of our study is that we tested a small cohort of ESCC samples. This was mainly due to the cost of sequencing to ensure that we could obtain high quality data with that could interrogate 3.3 million CpG sites at a depth of coverage >10, and to sequence matched transcriptome data with sufficient reads (~20 million reads per sample). Our findings must be tested on a larger African ESCC cohort. Another limitation is the integration of mutational data which may further inform on the sequence of events in ESCC carcinogenesis.

6.5 Future Studies

Our group intends to collect and perform methylation sequencing of our candidate genes on more ESCC samples once funding becomes available. Various initiatives are ongoing in our ESCC research group (Prof. C Mathew, Dr Pascale Willem). This includes exome sequencing on the same samples in our cohort to identify mutation signatures and driver genes and GWAS studies. Together, genomics studies would potentially allow us to contribute to precision medicine by defining the ESCC subtype more clearly in African populations.

6.6 Conclusion

In conclusion, the work presented in this thesis demonstrated important findings regarding ESCC in Southern Africa. First, we demonstrated the viability and feasibility of isolating high-quality nucleic acids for genomic studies in a limited resource setting.

We identified genes previously shown to have differential methylation levels in ESCC from other studies listed in Chapter 4. Our main findings were gene promoters and an enhancer whose methylation levels also inversely correlated with RNA expression that belonged to the ECM. Additional methylome studies with larger sample sizes are required to detect more of the genetic contribution to African ESCC and to target an epigenetic regulator of ESCC in African populations.

List of References

Akalin, A., Kormaksson, M., Li, S., *et al.*2012. MethylKit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. Genome Biol 13(10), R87. doi:10.1186/gb-2012-13-10-r87

Alaouna, M., Hull, R., Penny, C., *et al.*2019. Esophageal cancer genetics in south africa. Clin Exp Gastroenterol 12, 157–177. doi:10.2147/CEG.S182000

Alvarez, H., Opalinska, J., Zhou, L., *et al.*2011. Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. PLOS Genetics 7(3), e1001356. doi:10.1371/journal.pgen.1001356

Andersson, R.2015. Promoter or enhancer, what's the difference? deconstruction of established distinctions and presentation of a unifying model. Bioessays 37(3), 314–323. doi:10.1002/bies.201400162

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ao, R., Guan, L., Wang, Y., *et al.*2018. Silencing of col1a2, col6a3, and thbs2 inhibits gastric cancer cell proliferation, migration, and invasion while promoting apoptosis through the pi3k-akt signaling pathway. J Cell Biochem 119(6), 4420–4434. doi:10.1002/jcb.26524

Arafat, H., Lazar, M., Salem, K., *et al.*2011. Tumour-specific expression and alternative splicing of the col6a3 gene in pancreatic cancer. Surgery 150(2), 306–315. doi:10.1016/j.surg.2011.05.011

Arafat, H., Lazar, M., Salem, K., *et al.*2011. Tumour-specific expression and alternative splicing of the col6a3 gene in pancreatic cancer. Surgery 150(2), 306–315. doi:10.1016/j.surg.2011.05.011

Aran, D., Hellman, A.2013. DNA methylation of transcriptional enhancers and cancer predisposition. Cell 154(1), 11–13. doi:10.1016/j.cell.2013.06.018

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document] n.d. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 1.2.22).

Babraham Bioinformatics - Trim Galore! [WWW Document] n.d. URL https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed 1.2.22).

Balanis, N., Wendt, M.K., Schiemann, B.J., *et al.*2013. Epithelial to mesenchymal transition promotes breast cancer progression via a fibronectin-dependent stat3 signaling pathway. J Biol Chem 288(25), 17954–17967. doi:10.1074/jbc.M113.475277

Bashir, M., Damineni, S., Mukherjee, G., *et al.* 2015. Activin-a signaling promotes epithelial–mesenchymal transition, invasion, and metastatic growth of breast cancer. npj Breast Cancer 1(1), 1–13. doi:10.1038/npjbcancer.2015.7

Beggs, A.D., Mehta, S., Deeks, J.J., *et al.*2019. Validation of epigenetic markers to identify colitis associated cancer: results of module 1 of the endcap-c study. EBioMedicine 39, 265–271. doi:10.1016/j.ebiom.2018.11.034

Bell, R.E., Golan, T., Sheinboim, D., *et al.*2016. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. Genome Res 26(5), 601–611. doi:10.1101/gr.197194.115

Bergquist, H., Wenger, U., Johnsson, E., *et al.*2005. Stent insertion or endoluminal brachytherapy as palliation of patients with advanced cancer of the esophagus and gastroesophageal junction. results of a randomized, controlled clinical trial. Dis Esophagus 18(3), 131–139. doi:10.1111/j.1442-2050.2005.00467.x

Bitzer, M., Malek, N.P.2016. Chapter 42 - personalized epigenetics, in: Tollefsbol, T.O. (Ed.), Medical Epigenetics. Academic Press, Boston, pp. 843–858. doi:10.1016/B978-0-12-803239-8.00042-9

Bode, W., Maskos, K.2003. Structural basis of the matrix metalloproteinases and their physiological inhibitors, the tissue inhibitors of metalloproteinases. Biol Chem 384(6), 863–872. doi:10.1515/BC.2003.097

Bos, M.K., Deger, T., Sleijfer, S., *et al.*2022. ESR1 methylation measured in cell-free dna to evaluate endocrine resistance in metastatic breast cancer patients. International Journal of Molecular Sciences 23(10), 5631. doi:10.3390/ijms23105631

Bowcock, A.M.1993. Molecular cloning of brca1: a gene for early onset familial breast and ovarian cancer. Breast Cancer Res Treat 28(2), 121–135. doi:10.1007/BF00666425

Busch, S.M., Lorenzana, Z., Ryan, A.L.2021. Implications for extracellular matrix interactions with human lung basal stem cells in lung development, disease, and airway modeling. Front Pharmacol 12, 645858. doi:10.3389/fphar.2021.645858

Caldwell, R.I., Gaede, B., Aldous, C. 2015. Description of an internal medicine outreach consultant appointment in western kwazulu-natal, south africa, 2007 to mid-2014. S Afr Med J 105(5), 353–356. doi:10.7196/samj.9173

Calvo, F., Ranftl, R., Hooper, S., *et al.*2015. Cdc42EP3/borg2 and septin network enables mechano-transduction and the emergence of cancer-associated fibroblasts. Cell Reports 13(12), 2699–2714. doi:10.1016/j.celrep.2015.11.052

Calvo, F., Ranftl, R., Hooper, S., *et al.*2015. Cdc42EP3/borg2 and septin network enables mechano-transduction and the emergence of cancer-associated fibroblasts. Cell Reports 13(12), 2699–2714. doi:10.1016/j.celrep.2015.11.052

Cao, W., Lee, H., Wu, W., *et al.*2020. Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. Nat Commun 11(1), 3675. doi:10.1038/s41467-020-17227-z

Cawston, T.E., Young, D.A.2010. Proteinases involved in matrix turnover during cartilage and bone breakdown. Cell Tissue Res 339(1), 221–235. doi:10.1007/s00441-009-0887-6

Cawston, T.E., Young, D.A.2010. Proteinases involved in matrix turnover during cartilage and bone breakdown. Cell Tissue Res 339(1), 221–235. doi:10.1007/s00441-009-0887-6

Census SA . Provincial profile: KwaZulu-Natal. Available Census 2011. Provincial profile: KwaZulu-Natal. Available at www.statssa.gov.za (last accessed March 2, 2021).

ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualisation | Bioinformatics | Oxford Academic [WWW Document] n.d. URL https://academic.oup.com/bioinformatics/article/31/14/2382/255379 (accessed 2.15.22).

Chan, B.K.C.2018. Data analysis using r programming. Adv Exp Med Biol 1082, 47–122. doi:10.1007/978-3-319-93791-5_2

Chen, M.-F., Kuan, F.-C., Yen, T.-C., *et al.*2014. IL-6-stimulated cd11b+ cd14+ hla-dr- myeloid-derived suppressor cells, are associated with progression and poor prognosis in squamous cell carcinoma of the esophagus. Oncotarget 5(18), 8716–8728. doi:10.18632/oncotarget.2368

Chen, P., Cescon, M., Bonaldo, P.2013. Collagen vi in cancer and its biological mechanisms. Trends Mol Med 19(7), 410–417. doi:10.1016/j.molmed.2013.04.001

Chen, W., Zhu, Y., Zhang, W., *et al.*2021. CDC42EP3 is a key promoter involved in the development and progression of gastric cancer. Carcinogenesis 42(9), 1179–1188. doi:10.1093/carcin/bgab048

Chen, Y., Di, C., Zhang, X., *et al.*2020. Transforming growth factor β signaling pathway: a promising therapeutic target for cancer. J Cell Physiol 235(3), 1903–1914. doi:10.1002/jcp.29108

Cho, J.-W., Shim, H.S., Lee, C.Y., *et al.*2022. The importance of enhancer methylation for epigenetic regulation of tumourigenesis in squamous lung cancer. Exp Mol Med 54(1), 12–22. doi:10.1038/s12276-021-00718-4

Craene, B.D., Berx, G.2013. Regulatory networks defining emt during cancer initiation and progression. Nat Rev Cancer 13(2), 97–110. doi:10.1038/nrc3447

Cui, Y., Bi, M., Su, T., *et al.*2010. Molecular cloning and characterization of a novel esophageal cancer related gene. Int J Oncol 37(6), 1521–1528. doi:10.3892/ijo_00000805

Czaplinska, D., Ialchina, R., Andersen, H.B., *et al*.2023. Crosstalk between tumour acidosis, p53 and extracellular matrix regulates pancreatic cancer aggressiveness. International Journal of Cancer 152(6), 1210–1225. doi:10.1002/ijc.34367

Daley, W.P., Peters, S.B., Larsen, M.2008. Extracellular matrix dynamics in development and regenerative medicine. Journal of Cell Science 121(3), 255–264. doi:10.1242/jcs.006064

Dave, B., Mittal, V., Tan, N.M., *et al.*2012. Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. Breast Cancer Research 14(1), 202. doi:10.1186/bcr2938

Davies, M.A.2011. Regulation, role, and targeting of akt in cancer. J Clin Oncol 29(35), 4715–4717. doi:10.1200/JCO.2011.37.4751

Demir, E., Sabatelli, P., Allamand, V., *et al.*2002. Mutations in col6a3 cause severe and mild phenotypes of ullrich congenital muscular dystrophy. Am J Hum Genet 70(6), 1446–1458. doi:10.1086/340608

Ding, J., Li, D., Wang, X., *et al.*2008. Fibronectin promotes invasiveness and focal adhesion kinase tyrosine phosphorylation of human colon cancer cell. Hepatogastroenterology 55(88), 2072–2076

Dong, Y., Zhao, H., LI, H., *et al*. 2014. DNA methylation as an early diagnostic marker of cancer (review). Biomed Rep 2(3), 326–330. doi:10.3892/br.2014.237

DuFort, C.C., Paszek, M.J., Weaver, V.M.2011. Balancing forces: architectural control of mechanotransduction. Nature reviews. Molecular cell biology 12(5), 308. doi:10.1038/nrm3112

DuFort, C.C., Paszek, M.J., Weaver, V.M.2011. Balancing forces: architectural control of mechanotransduction. Nature reviews. Molecular cell biology 12(5), 308. doi:10.1038/nrm3112

Duffy, M.J.2004. The urokinase plasminogen activator system: role in malignancy. Curr Pharm Des 10(1), 39–49. doi:10.2174/1381612043453559

Egashira, A., Morita, M., Yoshida, R., *et al.*2011. Loss of p53 in esophageal squamous cell carcinoma and the correlation with survival: analyses of gene mutations, protein expression, and loss of heterozygosity in japanese patients. J Surg Oncol 104(2), 169–175. doi:10.1002/jso.21920

Ehrlich, M.2009. DNA hypomethylation in cancer cells. Epigenomics 1(2), 239–259. doi:10.2217/epi.09.33

Erkizan, H.V., Johnson, K., Ghimbovschi, S., *et al.*2017. African-american esophageal squamous cell carcinoma expression profile reveals dysregulation of stress response and detox networks. BMC Cancer 17(1), 426. doi:10.1186/s12885-017-3423-1

Feinberg, A.P., Vogelstein, B.1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301(5895), 89–92. doi:10.1038/301089a0

Ferndale, L., Ayeni, O., Chen, W., *et al.*2021a. Factors Influencing Survival Time in Patients Treated for Oesophageal Cancer with Palliative Intent in South Africa. doi:10.21203/rs.3.rs-1074694/v1

Ferndale, L., Moodley, M., Chen, W.C., *et al.*2021b. Processing and analysis of tissue samples from esophageal cancer patients in an african setting. Biopreserv Biobank. doi:10.1089/bio.2021.0030

Fitzgerald, A.A., Weiner, L.M.2020. The role of fibroblast activation protein in health and malignancy. Cancer Metastasis Rev 39(3), 783–803. doi:10.1007/s10555-020-09909-3

Frantz, C., Stewart, K.M., Weaver, V.M.2010. The extracellular matrix at a glance. J Cell Sci 123(Pt 24), 4195–4200. doi:10.1242/jcs.023820

Fu, J.-H., Wang, L.-Q., Li, T., *et al.*2015. RNA-sequencing based identification of crucial genes for esophageal squamous cell carcinoma. J Cancer Res Ther 11(2), 420–425. doi:10.4103/0973-1482.160122

Gaggioli, C., Hooper, S., Hidalgo-Carcedo, C., *et al.*2007. Fibroblast-led collective invasion of carcinoma cells with differing roles for rhogtpases in leading and following cells. Nat Cell Biol 9(12), 1392–1400. doi:10.1038/ncb1658

Gagliardi, A., Porter, V.L., Zong, Z., *et al*.2020. Analysis of ugandan cervical carcinomas identifies human papillomavirus clade-specific epigenome and transcriptome landscapes. Nat Genet 52(8), 800–810. doi:10.1038/s41588-020-0673-7

Gao, Y.-B., Chen, Z.-L., Li, J.-G., *et al.*2014. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet 46(10), 1097–1102. doi:10.1038/ng.3076

Gaston, J., Cheradame, L., Yvonnet, V., *et al.*2016. Intracellular sting inactivation sensitizes breast cancer cells to genotoxic agents. Oncotarget 7(47), 77205–77224. doi:10.18632/oncotarget.12858

Griffin, N., Rowe, C.W., Gao, F., *et al.*2020. Clinicopathological significance of nerves in esophageal cancer. The American Journal of Pathology 190(9), 1921–1930. doi:10.1016/j.ajpath.2020.05.012

Grolaux, R., Hardy, A., Olsen, C., et al. 2022. Identification of differentially methylated regions in rare diseases from a single-patient perspective. Clinical Epigenetics 14(1), 174. doi:10.1186/s13148-022-01403-7

Guo, W., Dong, Z., Cui, J., *et al.*2016. Aberrant hypermethylation of rassf2 in tumours and peripheral blood dna as a biomarker for malignant progression and poor prognosis of esophageal squamous cell carcinoma. Clin Exp Metastasis 33(1), 73–85. doi:10.1007/s10585-015-9759-5

Gutierrez, E., Cahatol, I., Bailey, C.A.R., *et al.*2019. Regulation of rhob gene expression during tumourigenesis and aging process and its potential applications in these processes. Cancers (Basel) 11(6), 818. doi:10.3390/cancers11060818

Hamada, T., McLean, W.H.I., Ramsay, M., *et al.*2002. Lipoid proteinosis maps to 1q21 and is caused by mutations in the extracellular matrix protein 1 gene (ecm1). Hum Mol Genet 11(7), 833–840. doi:10.1093/hmg/11.7.833

Harjunpää, H., Llort Asens, M., Guenther, C., *et al.*2019. Cell adhesion molecules and their roles and regulation in the immune and tumour microenvironment. Frontiers in Immunology 10

Hendrix, M.J., Seftor, E.A., Chu, Y.W., *et al.*1996. Role of intermediate filaments in migration, invasion and metastasis. Cancer Metastasis Rev 15(4), 507–525. doi:10.1007/BF00054016

Ho, M., Thompson, B., Fisk, J.N., *et al.*2022. Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders. Human Genomics 16(1), 1. doi:10.1186/s40246-021-00374-9

Hsu, P.-K., Kao, H.-L., Chen, H.-Y., *et al.*2014. Loss of crnn expression is associated with advanced tumour stage and poor survival in patients with esophageal squamous cell carcinoma. The Journal of Thoracic and Cardiovascular Surgery 147(5), 1612-1618.e4. doi:10.1016/j.jtcvs.2013.09.066

Huang, H., Zhang, G., Li, G., *et al.*2015. Circulating cd14(+)hla-dr(-/low) myeloid-derived suppressor cell is an indicator of poor prognosis in patients with escc. Tumour Biol 36(10), 7987–7996. doi:10.1007/s13277-015-3426-y

Huang, J., Koulaouzidis, A., Marlicz, W., *et al.*2021. Global burden, risk factors, and trends of esophageal cancer: an analysis of cancer registries from 48 countries. Cancers (Basel) 13(1), E141. doi:10.3390/cancers13010141

Huang, J., Wang, G., Tang, J., *et al.*2017. DNA methylation status of *PAX1* and *ZNF582* in esophageal squamous cell carcinoma. Int J Environ Res Public Health 14(2), E216. doi:10.3390/ijerph14020216

Huang, T.-X., Fu, L.2019. The immune landscape of esophageal cancer. Cancer Commun (Lond) 39, 79. doi:10.1186/s40880-019-0427-z

Huang, Y., Li, G., Wang, K., *et al.*2018. Collagen type vi alpha 3 chain promotes epithelial-mesenchymal transition in bladder cancer cells via transforming growth factor β (tgf-β)/smad pathway. Med Sci Monit 24, 5346–5354. doi:10.12659/MSM.909811

Hyman, G., Manglik, V., Rousch, J.M., *et al.*2015. Epigenetic approaches in glioblastoma multiforme and their implication in screening and diagnosis, in: Verma, M. (Ed.), Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis, Methods in Molecular Biology. Springer, New York, NY, pp. 511–521. doi:10.1007/978-1-4939-1804-1_26

Hyman, G., Manglik, V., Rousch, J.M., *et al.*2015. Epigenetic approaches in glioblastoma multiforme and their implication in screening and diagnosis, in: Verma, M. (Ed.), Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis, Methods in Molecular Biology. Springer, New York, NY, pp. 511–521. doi:10.1007/978-1-4939-1804-1_26

Hynes, R.O.2002. Integrins: bidirectional, allosteric signaling machines. Cell 110(6), 673–687. doi:10.1016/s0092-8674(02)00971-6

Ishiguro, S., Sasazuki, S., Inoue, M., *et al.*2009. Effect of alcohol consumption, cigarette smoking and flushing response on esophageal cancer risk: a population-based cohort study (jphc study). Cancer Lett 275(2), 240–246. doi:10.1016/j.canlet.2008.10.020

Iyengar, P., Espina, V., Williams, T.W., *et al.*2005. Adipocyte-derived collagen vi affects early mammary tumour progression in vivo, demonstrating a critical interaction in the tumour/stroma microenvironment. J Clin Invest 115(5), 1163–1176. doi:10.1172/JCI23424

Jaiswal, B.S., Janakiraman, V., Kljavin, N.M., *et al.*2009. Somatic mutations in p85α promote tumourigenesis through class ia pi3k activation. Cancer Cell 16(6), 463–474. doi:10.1016/j.ccr.2009.10.016

Jalkanen, J., Kotimäki, M., Huhtaniemi, I., *et al.*2006. Novel epididymal protease inhibitors with kazal or wap family domain. Biochem Biophys Res Commun 349(1), 245–254. doi:10.1016/j.bbrc.2006.08.023

Jiang, D., He, Z., Wang, C., *et al.*2018. Epigenetic silencing of *ZNF132* mediated by methylation-sensitive sp1 binding promotes cancer progression in esophageal squamous cell carcinoma. Cell Death Dis 10(1), 1–13. doi:10.1038/s41419-018-1236-z

Jin, Z., Olaru, A., Yang, J., *et al.*2007. Hypermethylation of tachykinin-1 is a potential biomarker in human esophageal cancer. Clin Cancer Res 13(21), 6293–6300. doi:10.1158/1078-0432.CCR-07-0818

Johnson, N.2001. Tobacco use and oral cancer: a global perspective. J Dent Educ 65(4), 328–339

Kalluri, R.2016. The biology and function of fibroblasts in cancer. Nat Rev Cancer 16(9), 582–598. doi:10.1038/nrc.2016.73

Kamangar, F., Diaw, L., Wei, W.-Q., *et al.*2009. Serum pepsinogens and risk of esophageal squamous dysplasia. Int J Cancer 124(2), 456–460. doi:10.1002/ijc.23918

Kamiya, A., Hayama, Y., Kato, S., *et al.*2019. Genetic manipulation of autonomic nerve fiber innervation and activity and its effect on breast cancer progression. Nat Neurosci 22(8), 1289–1305. doi:10.1038/s41593-019-0430-3

Kashyap, M.K., Marimuthu, A., Kishore, C.J.H., *et al.*2009. Genomewide mrna profiling of esophageal squamous cell carcinoma for identification of cancer biomarkers. Cancer Biol Ther 8(1), 36–46. doi:10.4161/cbt.8.1.7090

Kim, D., Paggi, J.M., Park, C., *et al.*2019. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nat Biotechnol 37(8), 907–915. doi:10.1038/s41587-019-0201-4

Kim, J., Bowlby, R., Mungall, A.J., *et al.*2017. Integrated genomic characterization of oesophageal carcinoma. Nature 541(7636), 169–175. doi:10.1038/nature20805

Kim, S., Wong, P., Coulombe, P.A.2006. A keratin cytoskeletal protein regulates protein synthesis and epithelial cell growth. Nature 441(7091), 362–365. doi:10.1038/nature04659

Kishino, T., Niwa, T., Yamashita, S., *et al.*2016. Integrated analysis of dna methylation and mutations in esophageal squamous cell carcinoma. Mol Carcinog 55(12), 2077–2088. doi:10.1002/mc.22452

Kishino, T., Niwa, T., Yamashita, S., *et al.*2016. Integrated analysis of dna methylation and mutations in esophageal squamous cell carcinoma. Mol Carcinog 55(12), 2077–2088. doi:10.1002/mc.22452

Kong, K.L., Kwong, D.L., Fu, L., *et al.*2010. Characterization of a candidate tumour suppressor gene uroplakin 1a in esophageal squamous cell carcinoma. Cancer Res 70(21), 8832–8841. doi:10.1158/0008-5472.CAN-10-0779

Korthauer, K., Chakraborty, S., Benjamini, Y., *et al.*2019. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics 20(3), 367–383. doi:10.1093/biostatistics/kxy007

Krueger, F. 2019. Trim Galore!

Available online at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Krueger, F., Andrews, S.R. 2011. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics 27(11), 1571–1572. doi:10.1093/bioinformatics/btr167

Kuroki, T., Trapasso, F., Yendamuri, S., *et al.*2003. Allele loss and promoter hypermethylation of vhl, RAR-β, rassf1a, and *FHIT* tumour suppressor genes on chromosome 3p in esophageal squamous cell carcinoma. Cancer Res 63(13), 3724–3728

Lee, C., Kim, M., Lee, J.H., *et al.*2019. COL6A3-derived endotrophin links reciprocal interactions among hepatic cells in the pathology of chronic liver disease. J Pathol 247(1), 99–109. doi:10.1002/path.5172

Lee, E.J., Lee, B.B., Han, J., *et al.*2008. CpG island hypermethylation of e-cadherin (CDH1) and integrin alpha4 is associated with recurrence of early stage esophageal squamous cell carcinoma. Int J Cancer 123(9), 2073–2079. doi:10.1002/ijc.23598

Lee, H.-Y., Li, C.-C., Huang, C.-N., *et al.* 2015. INHBA overexpression indicates poor prognosis in urothelial carcinoma of urinary bladder and upper tract. J Surg Oncol 111(4), 414–422. doi:10.1002/jso.23836

Lee, K.-D., Wang, T.-Y., Lu, C.-H., *et al.*2017. The bidirectional association between oral cancer and esophageal cancer: a population-based study in taiwan over a 28-year period. Oncotarget 8(27), 44567–44578. doi:10.18632/oncotarget.17818

Leuning, D.G., Beijer, N.R.M., du Fossé, N.A., *et al.*2018. The cytokine secretion profile of mesenchymal stromal cells is determined by surface structure of the microenvironment. Sci Rep 8(1), 7716. doi:10.1038/s41598-018-25700-5

Levental, K.R., Yu, H., Kass, L., *et al.*2009. Matrix crosslinking forces tumour progression by enhancing integrin signaling. Cell 139(5), 891–906. doi:10.1016/j.cell.2009.10.027

Li, B., Xu, W.W., Lam, A.K.Y., *et al.*2017. Significance of pi3k/akt signaling pathway in metastasis of esophageal squamous cell carcinoma and its potential as a target for anti-metastasis therapy. Oncotarget 8(24), 38755–38766. doi:10.18632/oncotarget.16333

Li, J.-S., Ying, J.-M., Wang, X.-W., *et al.*2013. Promoter methylation of tumour suppressor genes in esophageal squamous cell carcinoma. Chin J Cancer 32(1), 3–11. doi:10.5732/cjc.011.10381

Li, Jieling, Wang, X., Zheng, K., *et al.*2019. The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. PeerJ 7, e7705. doi:10.7717/peerj.7705

Lima, S.C.S., Hernández-Vargas, H., Simão, T., *et al.*2011. Identification of a dna methylome signature of esophageal squamous cell carcinoma and potential epigenetic biomarkers. Epigenetics 6(10), 1217–1227. doi:10.4161/epi.6.10.17199

Lin, D.-C., Hao, J.-J., Nagata, Y., *et al.*2014. Genomic and molecular characterization of esophageal squamous cell carcinoma. Nat Genet 46(5), 467–473. doi:10.1038/ng.2935

Lin, N., Liu, J., Castle, J., *et al.*2021. Genome-wide dna methylation profiling in human breast tissue by illumina truseq methyl capture epic sequencing and infinium methylationepic beadchip microarray. Epigenetics 16(7), 754–769. doi:10.1080/15592294.2020.1827703

Ling, Z.-Q., Li, P., Ge, M.-H., *et al.*2011. Hypermethylation-modulated down-regulation of CDH1 expression contributes to the progression of esophageal cancer. Int J Mol Med 27(5), 625–635. doi:10.3892/ijmm.2011.640

Litosh, V.A., Rochman, M., Rymer, J.K., *et al.*2017. Calpain-14 and its association with eosinophilic esophagitis. J Allergy Clin Immunol 139(6), 1762-1771.e7. doi:10.1016/j.jaci.2016.09.027

Liu, W., Snell, J.M., Jeck, W.R., *et al.* 2016. Subtyping sub-saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight 1(16), e88755. doi:10.1172/jci.insight.88755

Liu, G., Zhao, Y., Chen, H., *et al.*2020. Analysis of differentially expressed genes in a chinese cohort of esophageal squamous cell carcinoma. J Cancer 11(13), 3783–3793. doi:10.7150/jca.40850

Liu, H., Qin, Y.-R., Bi, J., *et al.*2012. Overexpression of matrix metalloproteinase 10 is associated with poor survival in patients with early stage of esophageal squamous cell carcinoma. Diseases of the Esophagus 25(7), 656–663. doi:10.1111/j.1442-2050.2011.01284.x

Liu, M., Hu, Y., Zhang, M.-F., *et al.*2016. MMP1 promotes tumour growth and metastasis in esophageal squamous cell carcinoma. Cancer Letters 377(1), 97–104. doi:10.1016/j.canlet.2016.04.034

Liu, W., Li, L., Ye, H., *et al.*2018. Role of col6a3 in colorectal cancer. Oncol Rep 39(6), 2527–2536. doi:10.3892/or.2018.6331

Liu, W., Snell, J.M., Jeck, W.R., *et al.*n.d. Subtyping sub-saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight 1(16), e88755. doi:10.1172/jci.insight.88755

Liu, Y., Fang, Y., Chen, X., *et al.*2020. Gasdermin e-mediated target cell pyroptosis by car t cells triggers cytokine release syndrome. Sci Immunol 5(43), eaax7969. doi:10.1126/sciimmunol.aax7969

Liu, Y., Zhang, Y., Zhao, Y., *et al.*2016. High parp-1 expression is associated with tumour invasion and poor prognosis in gastric cancer. Oncol Lett 12(5), 3825–3835. doi:10.3892/ol.2016.5169

Liu, Z., Yu, S., Ye, S., *et al.*2020. Keratin 17 activates akt signalling and induces epithelial-mesenchymal transition in oesophageal squamous cell carcinoma. J Proteomics 211, 103557. doi:10.1016/j.jprot.2019.103557

Loots, E., Ferndale, L., Clarke, D.L.2019. Oesophageal cancer in south africa: the need for a public health response. South African Journal of Surgery 57(2), 2–2

Loots, E., Sartorius, B., Madiba, T.E., *et al.*2017. Is clinical research in oesophageal cancer in south africa in crisis? a systematic review. World J Surg 41(3), 810–816. doi:10.1007/s00268-016-3778-5

Lotem, J., Netanely, D., Domany, E., *et al.*2005. Human cancers overexpress genes that are specific to a variety of normal human tissues. Proceedings of the National Academy of Sciences 102(51), 18556–18561. doi:10.1073/pnas.0509360102

Lu, P., Takai, K., Weaver, V.M., *et al.*2011. Extracellular matrix degradation and remodeling in development and disease. Cold Spring Harb Perspect Biol 3(12), a005058. doi:10.1101/cshperspect.a005058

Lu, P., Weaver, V.M., Werb, Z.2012. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol 196(4), 395–406. doi:10.1083/jcb.201102147

Lu, T., Chen, D., Wang, Y., *et al.*2019. Identification of dna methylation-driven genes in esophageal squamous cell carcinoma: a study based on the cancer genome atlas. Cancer Cell International 19(1), 52. doi:10.1186/s12935-019-0770-9

Lu, T., Chen, D., Wang, Y., *et al.*2019. Identification of dna methylation-driven genes in esophageal squamous cell carcinoma: a study based on the cancer genome atlas. Cancer Cell International 19(1), 52. doi:10.1186/s12935-019-0770-9

Lyu, S., Jiang, C., Xu, R., *et al.*2018. INHBA upregulation correlates with poorer prognosis in patients with esophageal squamous cell carcinoma. Cancer Manag Res 10, 1585–1596. doi:10.2147/CMAR.S160186

Ma, L., Yu, H., Ni, Z., *et al.*2013. Spink13, an epididymis-specific gene of the kazal-type serine protease inhibitor (spink) family, is essential for the acrosomal integrity and male fertility. J Biol Chem 288(14), 10154–10165. doi:10.1074/jbc.M112.445866

Ma, Y., He, S., Gao, A., *et al.*2020. Methylation silencing of tgf-β receptor type ii is involved in malignant transformation of esophageal squamous cell carcinoma. Clinical Epigenetics 12(1), 25. doi:10.1186/s13148-020-0819-6

Mabi, B.2017. Esophageal cancer in the sudan: focus on the screening for ebv and the surgical management. MOJ Surgery 4(6)

Maekawa, R., Sato, S., Yamagata, Y., *et al.*2013. Genome-wide dna methylation analysis reveals a potential mechanism for the pathogenesis and development of uterine leiomyomas. PLoS One 8(6), e66632. doi:10.1371/journal.pone.0066632

Maekawa, R., Sato, S., Yamagata, Y., *et al.*2013. Genome-wide dna methylation analysis reveals a potential mechanism for the pathogenesis and development of uterine leiomyomas. PLoS One 8(6), e66632. doi:10.1371/journal.pone.0066632

Magnon, C., Hall, S.J., Lin, J., *et al.*2013. Autonomic nerve development contributes to prostate cancer progression. Science 341(6142), 1236361. doi:10.1126/science.1236361

Maitra, A., Biswas, N. K., Amin, K., *et al*. (2013). Mutational landscape of gingivo-buccal oral squamous cell carcinoma reveals new recurrently-mutated genes and molecular subgroups. Nature Communications 4(1), 3873

Mariette, C., Finzi, L., Piessen, G., *et al.*2005. Esophageal carcinoma: prognostic differences between squamous cell carcinoma and adenocarcinoma. World J Surg 29(1), 39–45. doi:10.1007/s00268-004-7542-x

Marsit, C.J., Houseman, E.A., Christensen, B.C., *et al.*2010. Identification of methylated genes associated with aggressive bladder cancer. PLoS One 5(8), e12334. doi:10.1371/journal.pone.0012334

Mehmood, A., Laiho, A., Elo, L.L.2021. Exon-level estimates improve the detection of differentially expressed genes in rna-seq studies. RNA Biology 18(11), 1739–1746. doi:10.1080/15476286.2020.1868151

Mendy, M., Caboux, E., Lawlor, R.T., *et al.*2017. Common Minimum Technical Standards and Protocols for Biobanks Dedicated to Cancer Research, IARC Technical Publications. International Agency for Research on Cancer, Lyon (FR)

Midwood, K.S., Chiquet, M., Tucker, R.P., *et al.*2016. Tenascin-c at a glance. Journal of Cell Science 129(23), 4321–4327. doi:10.1242/jcs.190546

Miroshnikova, Y.A., Mouw, J.K., Barnes, J.M., *et al.*2016. Tissue mechanics promote idh1-dependent hif1α–tenascin c feedback to regulate glioblastoma aggression. Nat Cell Biol 18(12), 1336–1345. doi:10.1038/ncb3429

Mj, S., Bk, R., B, C., *et al.*1994. Molecular cloning of fibroblast activation protein alpha, a member of the serine protease family selectively expressed in stromal fibroblasts of epithelial cancers. Proceedings of the National Academy of Sciences of the United States of America 91(12). doi:10.1073/pnas.91.12.5657

Moll, R., Franke, W.W., Schiller, D.L., *et al.*1982. The catalog of human cytokeratins: patterns of expression in normal epithelia, tumours and cultured cells. Cell 31(1), 11–24. doi:10.1016/0092-8674(82)90400-7

Moody, S., Senkin, S., Islam, S.M.A., *et al*.2021. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. Nat Genet 53(11), 1553–1563. doi:10.1038/s41588-021-00928-6Moran, S., Arribas, C., Esteller, M.2016. Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. Epigenomics 8(3), 389–399. doi:10.2217/epi.15.114

Naba, A., Clauser, K.R., Ding, H., *et al.*2016. The extracellular matrix: tools and insights for the “omics” era. Matrix Biology 49, 10–24. doi:10.1016/j.matbio.2015.06.003

Naganuma, K., Hatta, M., Ikebe, T., *et al.*2014. Epigenetic alterations of the keratin 13 gene in oral squamous cell carcinoma. BMC Cancer 14(1), 988. doi:10.1186/1471-2407-14-988

Nagata, H., Kozaki, K.-I., Muramatsu, T., *et al.*2017. Genome-wide screening of dna methylation associated with lymph node metastasis in esophageal squamous cell carcinoma. Oncotarget 8(23), 37740–37750. doi:10.18632/oncotarget.17147

Naito, Y., Yamamoto, Y., Sakamoto, N., *et al.*2019. Cancer extracellular vesicles contribute to stromal heterogeneity by inducing chemokines in cancer-associated fibroblasts. Oncogene 38(28), 5566–5579. doi:10.1038/s41388-019-0832-4

Nanda, A., Carson-Walter, E.B., Seaman, S., *et al.*2004. TEM8 interacts with the cleaved c5 domain of collagen alpha 3(vi). Cancer Res 64(3), 817–820. doi:10.1158/0008-5472.can-03-2408

Nishiyama, A., Nakanishi, M.2021. Navigating the dna methylation landscape of cancer. Trends in Genetics 37(11), 1012–1027. doi:10.1016/j.tig.2021.05.002

Nishiyama, A., Nakanishi, M.2021. Navigating the dna methylation landscape of cancer. Trends in Genetics 37(11), 1012–1027. doi:10.1016/j.tig.2021.05.002

Okano, M., Yamamoto, H., Ohkuma, H., et al. 2013. Significance of inhba expression in human colorectal cancer. Oncol Rep 30(6), 2903–2908. doi:10.3892/or.2013.2761

Orend, G., Chiquet-Ehrismann, R.2006. Tenascin-c induced signaling in cancer. Cancer Lett 244(2), 143–163. doi:10.1016/j.canlet.2006.02.017

Orend, G., Chiquet-Ehrismann, R.2006. Tenascin-c induced signaling in cancer. Cancer Lett 244(2), 143–163. doi:10.1016/j.canlet.2006.02.017

Oshlack, A., Wakefield, M.J.2009. Transcript length bias in rna-seq data confounds systems biology. Biology Direct 4(1), 14. doi:10.1186/1745-6150-4-14

Palumbo, A., Meireles Da Costa, N., Pontes, B., et al.2020. Esophageal cancer development: crucial clues arising from the extracellular matrix. Cells 9(2), 455. doi:10.3390/cells9020455

Pan, T.-C., Zhang, R.-Z., Pericak-Vance, M.A., et al.1998. Missense mutation in a von willebrand factor type a domain of the α3(vi) collagen gene (col6a3) in a family with bethlem myopathy. Human Molecular Genetics 7(5), 807–812. doi:10.1093/hmg/7.5.807

Peng, X., Xue, H., Lü, L., et al.2016. Accumulated promoter methylation as a potential biomarker for esophageal cancer. Oncotarget 8(1), 679–691. doi:10.18632/oncotarget.13510

Peng, X., Xue, H., Lü, L., *et al.*2016. Accumulated promoter methylation as a potential biomarker for esophageal cancer. Oncotarget 8(1), 679–691. doi:10.18632/oncotarget.13510

Pfeifer, G.P.2018. Defining driver dna methylation changes in human cancer. International Journal of Molecular Sciences 19(4). doi:10.3390/ijms19041166

Procopio, M.-G., Laszlo, C., Al Labban, D., *et al*.2015. Combined csl and p53 downregulation promotes cancer-associated fibroblast activation. Nat Cell Biol 17(9), 1193–1204. doi:10.1038/ncb3228

Qiao, Y., Liu, B., Bai, R., *et al.*2022. White sponge nevus caused by keratin 4 gene mutation: a case report. Genes (Basel) 13(12), 2184. doi:10.3390/genes13122184

Ramos, G. de O., Bernardi, L., Lauxen, I., *et al.*2016. Fibronectin modulates cell adhesion and signaling to promote single cell migration of highly invasive oral squamous cell carcinoma. PLOS ONE 11(3), e0151338. doi:10.1371/journal.pone.0151338

Rawlings, N.D., Tolle, D.P., Barrett, A.J.2004. Evolutionary families of peptidase inhibitors. Biochem J 378(Pt 3), 705–716. doi:10.1042/BJ20031825

Reed, P.I., Johnston, B.J.1993. The changing incidence of oesophageal cancer. Endoscopy 25(9), 606–608. doi:10.1055/s-2007-1010414

Ren, G., Roberts, A.I., Shi, Y.2011. Adhesion molecules. Cell Adh Migr 5(1), 20–22. doi:10.4161/cam.5.1.13491

Rezwan, F.I., Docherty, L.E., Poole, R.L., et al. 2015. A statistical method for single sample analysis of humanmethylation450 array data: genome-wide methylation analysis of patients with imprinting disorders. Clinical Epigenetics 7(1), 48. doi:10.1186/s13148-015-0081-5

Ridley, A.J.2015. Rho gtpase signalling in cell migration. Current Opinion in Cell Biology 36, 103–112. doi:10.1016/j.ceb.2015.08.005

Risso, D., Schwartz, K., Sherlock, G., *et al.*2011. GC-content normalization for rna-seq data. BMC Bioinformatics 12, 480. doi:10.1186/1471-2105-12-480

Robinson, M.D., McCarthy, D.J., Smyth, G.K.2010. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1), 139–140. doi:10.1093/bioinformatics/btp616

Robinson, M.D., Oshlack, A.2010. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol 11(3), R25. doi:10.1186/gb-2010-11-3-r25

Rowe, R.G., Weiss, S.J.2009. Navigating ecm barriers at the invasive front: the cancer cell–stroma interface. Annual Review of Cell and Developmental Biology 25(1), 567–595. doi:10.1146/annurev.cellbio.24.110707.175315

Rozario, T., DeSimone, D.W.2010. The extracellular matrix in development and morphogenesis: a dynamic view. Developmental Biology 341(1), 126–140. doi:10.1016/j.ydbio.2009.10.026

Saito, T., Mitomi, H., Imamhasan, A., *et al.*2014. Downregulation of sfrp-2 by epigenetic silencing activates the β-catenin/wnt signaling pathway in esophageal basaloid squamous cell carcinoma. Virchows Arch 464(2), 135–143. doi:10.1007/s00428-014-1538-1

Saldin, L.T., Patel, S., Zhang, L., *et al.*2019. Extracellular matrix degradation products downregulate neoplastic esophageal cell phenotype. Tissue Eng Part A 25(5–6), 487–498. doi:10.1089/ten.TEA.2018.0105

Saloman, J.L., Albers, K.M., Li, D., *et al.*2016. Ablation of sensory neurons in a genetic model of pancreatic ductal adenocarcinoma slows initiation and progression of cancer. Proc Natl Acad Sci U S A 113(11), 3078–3083. doi:10.1073/pnas.1512603113

Samantaray, S., Sharma, R., Chattopadhyaya, T.K., *et al.*2004. Increased expression of mmp-2 and mmp-9 in esophageal squamous cell carcinoma. J Cancer Res Clin Oncol 130(1), 37–44. doi:10.1007/s00432-003-0500-4

Sammon, A.M.2007. 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. Medical Hypotheses 69(1), 125–131. doi:10.1016/j.mehy.2006.11.011

Sankalecha, T.H., Gupta, S.J., Gaikwad, N.R., *et al*.2017. Yield of p53 expression in esophageal squamous cell cancer and its relationship with survival. Saudi J Gastroenterol 23(5), 281–286. doi:10.4103/sjg.SJG_56_17

Sawada, G., Niida, A., Uchi, R., *et al.*2016. Genomic landscape of esophageal squamous cell carcinoma in a japanese population. Gastroenterology 150(5), 1171–1182. doi:10.1053/j.gastro.2016.01.035

Seder, C.W., Hartojo, W., Lin, L., *et al.* 2009. Upregulated inhba expression may promote cell proliferation and is associated with poor survival in lung adenocarcinoma. Neoplasia 11(4), 388–396. doi:10.1593/neo.81582

Segal, I., Reinach, S.G., de Beer, M.1988. Factors associated with oesophageal cancer in soweto, south africa. Br J Cancer 58(5), 681–686

Seiz, L., Dorn, J., Kotzsch, M., *et al.*2012. Stromal cell-associated expression of kallikrein-related peptidase 6 (klk6) indicates poor prognosis of ovarian cancer patients. Biol Chem 393(5), 391–401. doi:10.1515/hsz-2011-0264

Shah, J.N., Shao, G., Hei, T.K., *et al.*2008. Methylation screening of the tgfbi promoter in human lung and prostate cancer by methylation-specific pcr. BMC Cancer 8, 284. doi:10.1186/1471-2407-8-284

Sherman-Baust, C.A., Weeraratna, A.T., Rangel, L.B.A., *et al.*2003. Remodeling of the extracellular matrix through overexpression of collagen vi contributes to cisplatin resistance in ovarian cancer cells. Cancer Cell 3(4), 377–386. doi:10.1016/S1535-6108(03)00058-8

Shi, X., Chen, Z., Hu, X., *et al.*2016. AJUBA promotes the migration and invasion of esophageal squamous cell carcinoma cells through upregulation of mmp10 and mmp13 expression. Oncotarget 7(24), 36407–36418. doi:10.18632/oncotarget.9239

Shibuya, Y., Zhang, J., Yokoo, S., *et al.*2003. Constitutional mutation of keratin 13 gene in familial white sponge nevus. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 96(5), 561–565. doi:10.1016/s1079-2104(03)00372-x

Shin, W.-S., Hong, Y., Lee, H.W., *et al.*2016. Catalytically defective receptor protein tyrosine kinase ptk7 enhances invasive phenotype by inducing mmp-9 through activation of ap-1 and nf-κb in esophageal squamous cell carcinoma cells. Oncotarget 7(45), 73242–73256. doi:10.18632/oncotarget.12303

Shull, A.Y., Latham-Schwark, A., Ramasamy, P., *et al.*2012. Novel somatic mutations to pi3k pathway genes in metastatic melanoma. PLoS One 7(8), e43369. doi:10.1371/journal.pone.0043369

Silva, C., Machado, M., Ferrão, J., *et al.*2021. Whole human genome 5'-mc methylation analysis using long read nanopore sequencing. doi:10.1101/2021.05.20.444035

Singh, P., Schwarzbauer, J.E.2012. Fibronectin and stem cell differentiation - lessons from chondrogenesis. J Cell Sci 125(Pt 16), 3703–3712. doi:10.1242/jcs.095786

Soares-Lima, S.C., Gonzaga, I.M., Camuzi, D., *et al.*2021. IL6 and bcl3 expression are potential biomarkers in esophageal squamous cell carcinoma. Frontiers in Oncology 11

Somdyala, N.I., Bradshaw, D., Gelderblom, W.C., *et al.*2010. Cancer incidence in a rural population of south africa, 1998-2002. Int J Cancer 127(10), 2420–2429. doi:10.1002/ijc.25246

Song, Y., van den Berg, P.R., Markoulaki, S., *et al.*2019. Dynamic enhancer dna methylation as basis for transcriptional and cellular heterogeneity of escs. Molecular Cell 75(5), 905-920.e6. doi:10.1016/j.molcel.2019.06.045

Stone, A., Zotenko, E., Locke, W.J., *et al.*2015. DNA methylation of oestrogen-regulated enhancers defines endocrine sensitivity in breast cancer. Nat Commun 6, 7758. doi:10.1038/ncomms8758

Subramanian, A., Tamayo, P., Mootha, V.K., *et al.*2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102(43), 15545–15550. doi:10.1073/pnas.0506580102

Sumeruk, R., Segal, I., Te Winkel, W., *et al.*1992. Oesophageal cancer in three regions of south africa. S Afr Med J 81(2), 91–93

Sun, Z., Schwenzer, A., Rupp, T., *et al.*2018. Tenascin-c promotes tumour cell migration and metastasis through integrin α9β1-mediated yap inhibition. Cancer Res 78(4), 950–961. doi:10.1158/0008-5472.CAN-17-1597

Sung, H., Ferlay, J., Siegel, R.L., *et al.*2021. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 71(3), 209–249. doi:10.3322/caac.21660

Supplitt, S., Karpinski, P., Sasiadek, M., *et al.*2021. Current achievements and applications of transcriptomics in personalized cancer medicine. Int J Mol Sci 22(3), 1422. doi:10.3390/ijms22031422

Suzuki, M., Liao, W., Wos, F., *et al.*2018. Whole-genome bisulfite sequencing with improved accuracy and cost. Genome Res 28(9), 1364–1371. doi:10.1101/gr.232587.117

Takemoto, A., Tanimoto, K., Mori, S., *et al.*2021. Integrative genome-wide analyses reveal the transcriptional aberrations in japanese esophageal squamous cell carcinoma. Cancer Sci 112(10), 4377–4392. doi:10.1111/cas.15063

Talukdar, F.R., Soares Lima, S.C., Khoueiry, R., *et al.*2021. Genome-wide dna methylation profiling of esophageal squamous cell carcinoma from global high-incidence regions identifies crucial genes and potential cancer markers. Cancer Res 81(10), 2612–2624. doi:10.1158/0008-5472.CAN-20-3445

Talukdar, F.R., Soares Lima, S.C., Khoueiry, R., *et al.*2021. Genome-wide dna methylation profiling of esophageal squamous cell carcinoma from global high-incidence regions identifies crucial genes and potential cancer markers. Cancer Res 81(10), 2612–2624. doi:10.1158/0008-5472.CAN-20-3445

Tanaka, T., Ikari, K., Furushima, K., *et al.*2003. Genomewide linkage and linkage disequilibrium analyses identify col6a1, on chromosome 21, as the locus for ossification of the posterior longitudinal ligament of the spine. Am J Hum Genet 73(4), 812–822

Thorpe, L.M., Yuzugullu, H., Zhao, J.J.2015. PI3K in cancer: divergent roles of isoforms, modes of activation, and therapeutic targeting. Nat Rev Cancer 15(1), 7–24. doi:10.1038/nrc3860

Thurman, R.E., Rynes, E., Humbert, R., *et al.*2012. The accessible chromatin landscape of the human genome. Nature 489(7414), 75–82. doi:10.1038/nature11232

TruSeq Methyl Capture EPIC Library Prep Kit Documentation [WWW Document] n.d. URL https://support.illumina.com/sequencing/sequencing_kits/TruSeq-methyl-capture-epic-kit/documentation.html (accessed 3.27.22).

TruSeq Methyl Capture EPIC Library Prep Kit Documentation [WWW Document] n.d. URL https://support.illumina.com/sequencing/sequencing_kits/truseq-methyl-capture-epic-kit/documentation.html (accessed 3.27.22).

Vakonakis, I., Campbell, I.D.2007. Extracellular matrix: from atomic resolution to ultrastructure. Current Opinion in Cell Biology 19(5), 578–583. doi:10.1016/j.ceb.2007.09.005

Waalkes, S., Atschekzei, F., Kramer, M.W., *et al.*2010. Fibronectin 1 mrna expression correlates with advanced disease in renal cancer. BMC Cancer 10, 503. doi:10.1186/1471-2407-10-503

Walker, C., Mojares, E., Del Río Hernández, A.2018. Role of extracellular matrix in development and cancer progression. Int J Mol Sci 19(10), E3028. doi:10.3390/ijms19103028

Wang, J., Pan, W.2020. The biological role of the collagen alpha-3 (vi) chain and its cleaved c5 domain fragment endotrophin in cancer. Onco Targets Ther 13, 5779–5793. doi:10.2147/OTT.S256654

Wang, J., Pan, W.2020. The biological role of the collagen alpha-3 (vi) chain and its cleaved c5 domain fragment endotrophin in cancer. Onco Targets Ther 13, 5779–5793. doi:10.2147/OTT.S256654

Wang, Q., Wen, Y.-G., Li, D.-P., *et al.* 2012. Upregulated inhba expression is associated with poor survival in gastric cancer. Med Oncol 29(1), 77–83. doi:10.1007/s12032-010-9766-y Wang, N., Zhang, H., Yao, Q., *et al.*2012. TGFBI promoter hypermethylation correlating with paclitaxel chemoresistance in ovarian cancer. Journal of Experimental & Clinical Cancer Research 31(1), 6. doi:10.1186/1756-9966-31-6

Wang, Q., Armenia, J., Zhang, C., *et al.*2018. Unifying cancer and normal rna sequencing data from different sources. Sci Data 5(1), 180061. doi:10.1038/sdata.2018.61

Wang, W., Sun, Q.-K., He, Y.-F., *et al.*2014. Overexpression of periostin is significantly correlated to the tumour angiogenesis and poor prognosis in patients with esophageal squamous cell carcinoma. Int J Clin Exp Pathol 7(2), 593–601

Wang, Z., Ren, B., Huang, J., *et al.*2018. LncRNA duxap10 modulates cell proliferation in esophageal squamous cell carcinoma through epigenetically silencing p21. Cancer Biol Ther 19(11), 998–1005. doi:10.1080/15384047.2018.1470723

Wapenaar, M.C., Monsuur, A.J., Poell, J., *et al.*2007. The spink gene family and celiac disease susceptibility. Immunogenetics 59(5), 349–357. doi:10.1007/s00251-007-0199-5

Wei, R., Liu, S., Zhang, S., *et al.*2020. Cellular and extracellular components in tumour microenvironment and their application in early diagnosis of cancers. Anal Cell Pathol (Amst) 2020, 6283796. doi:10.1155/2020/6283796

Wight, T.N., Kinsella, M.G., Qwarnström, E.E.1992. The role of proteoglycans in cell adhesion, migration and proliferation. Current Opinion in Cell Biology 4(5), 793–801. doi:10.1016/0955-0674(92)90102-I

Winkler, J., Abisoye-Ogunniyan, A., Metcalf, K.J., *et al.*2020. Concepts of extracellular matrix remodelling in tumour progression and metastasis. Nat Commun 11(1), 5120. doi:10.1038/s41467-020-18794-x

Worst, T.S., Reiner, V., Gabriel, U., *et al.*2014. IL1RN and krt13 expression in bladder cancer: association with pathologic characteristics and smoking status. Adv Urol 2014, 184602. doi:10.1155/2014/184602

Wreczycka, K., Gosdschan, A., Yusuf, D., *et al.*2017. Strategies for analyzing bisulfite sequencing data. Journal of Biotechnology 261, 105–115. doi:10.1016/j.jbiotec.2017.08.007

Wu, L., Herman, J.G., Brock, M.V., *et al.*2014. Silencing dach1 promotes esophageal cancer growth by inhibiting tgf-β signaling. PLoS One 9(4), e95509. doi:10.1371/journal.pone.0095509

Xi, Y., Lin, Y., Guo, W., *et al*.2022. Multi-omic characterization of genome-wide abnormal dna methylation reveals diagnostic and prognostic markers for esophageal squamous-cell carcinoma. Signal Transduct Target Ther 7(1), 53. doi:10.1038/s41392-022-00873-8

Xing, S., Zheng, X., Wei, L.-Q., *et al.*2017. Development and validation of a serum biomarker panel for the detection of esophageal squamous cell carcinoma through rna transcriptome sequencing. J Cancer 8(12), 2346–2355. doi:10.7150/jca.19465

Xu, G., Ou, L., Liu, Y., *et al.*2020. Upregulated expression of mmp family genes is associated with poor survival in patients with esophageal squamous cell carcinoma via regulation of proliferation and epithelial-mesenchymal transition. Oncol Rep 44(1), 29–42. doi:10.3892/or.2020.7606

Yang, C.S.1980. Research on esophageal cancer in china: a review. Cancer Res 40(8 Pt 1), 2633–2644

Yang, F., Strand, D.W., Rowley, D.R.2008. Fibroblast growth factor-2 mediates transforming growth factor-beta action in prostate cancer reactive stroma. Oncogene 27(4), 450–459. doi:10.1038/sj.onc.1210663

Yang, Z.-T., Yeo, S.-Y., Yin, Y.-X., *et al.*2016. Tenascin-c, a prognostic determinant of esophageal squamous cell carcinoma. PLoS One 11(1), e0145807. doi:10.1371/journal.pone.0145807

Yi, L., Pimentel, H., Bray, N.L., *et al.*2018. Gene-level differential analysis at transcript-level resolution. Genome Biology 19(1), 53. doi:10.1186/s13059-018-1419-z

Yilmaz, A., Loustau, T., Salomé, N., *et al.*2022. Advances on the roles of tenascin-c in cancer. J Cell Sci 135(18), jcs260244. doi:10.1242/jcs.260244

Yu, G., He, Q.-Y.2016. ReactomePA: an r/bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 12(2), 477–479. doi:10.1039/c5mb00663e

Yu, Y., Li, Z., Huang, C., *et al.*2019. Integrated analysis of genomic and transcriptomic profiles identified a prognostic immunohistochemistry panel for esophageal squamous cell cancer. Cancer Med 9(2), 575–585. doi:10.1002/cam4.2744

Yun, T., Liu, Y., Gao, D., *et al.*2015. Methylation of *CHFR* sensitizes esophageal squamous cell cancer to docetaxel and paclitaxel. Genes Cancer 6(1–2), 38–48

Zanussi, S., Doliana, R., Segat, D., *et al.*1992. The human type vi collagen gene. mrna and protein variants of the alpha 3 chain generated by alternative splicing of an additional 5-end exon. Journal of Biological Chemistry 267(33), 24082–24089. doi:10.1016/S0021-9258(18)35949-0

Zhang, H., Shi, Q., Yang, Z., *et al.*2021. An extracellular matrix-based signature associated with immune microenvironment predicts the prognosis and therapeutic responses of patients with oesophageal squamous cell carcinoma. Frontiers in Molecular Biosciences 8

Zhang, J., Zhi, H., Zhou, C., *et al.*2005. Up-regulation of fibronectin in oesophageal squamous cell carcinoma is associated with activation of the erk pathway. J Pathol 207(4), 402–409. doi:10.1002/path.1846

Zhang, L., Gao, Y., Zhang, X., *et al.*2020. TSTA3 facilitates esophageal squamous cell carcinoma progression through regulating fucosylation of lamp2 and erbb2. Theranostics 10(24), 11339–11358. doi:10.7150/thno.48225

Zhang, L.-Y., Wu, J.-L., Qiu, H.-B., *et al.*2016. PSCA acts as a tumour suppressor by facilitating the nuclear translocation of rb1cc1 in esophageal squamous cell carcinoma. Carcinogenesis 37(3), 320–332. doi:10.1093/carcin/bgw010

Zhang, W., Hubin, G., Endam, L.M., *et al.*2012. Expression of the extracellular matrix gene periostin is increased in chronic rhinosinusitis and decreases following successful endoscopic sinus surgery. Int Forum Allergy Rhinol 2(6), 471–476. doi:10.1002/alr.21056

Zheng, S., Liu, T., Middottuersun, A., *et al.*2011. [Expression of erk1/2 mapk signaling transduction pathway in esophageal cancers in kazakh patients]. Zhonghua Zhong Liu Za Zhi 33(6), 421–425

Zheng, Y., Gao, Q., Su, X., *et al.*2022. Genome-wide dna methylation and gene expression profiling characterizes molecular subtypes of esophagus squamous cell carcinoma for predicting patient survival and immunotherapy efficacy. Cancers (Basel) 14(20), 4970. doi:10.3390/cancers14204970

Zheng, Y., Zhang, Y., Huang, X., *et al.*2011. Analysis of the *RUNX3* gene methylation in serum dna from esophagus squamous cell carcinoma, gastric and colorectal adenocarcinoma patients. Hepatogastroenterology 58(112), 2007–2011. doi:10.5754/hge10016

Zhong, L., Li, H., Chang, W., *et al*.2023. TP53 mutations in esophageal squamous cell carcinoma. FBL 28(9), 219. doi:10.31083/j.fbl2809219

Zhou, Y., Guo, S., Li, Y., *et al.*2022. METTL3 is associated with the malignancy of esophageal squamous cell carcinoma and serves as a potential immunotherapy biomarker. Front Oncol 12, 824190. doi:10.3389/fonc.2022.824190

Zhou, Z., He, H., Wang, K., *et al.*2020. Granzyme a from cytotoxic lymphocytes cleaves gsdmb to trigger pyroptosis in target cells. Science 368(6494), eaaz7548. doi:10.1126/science.aaz7548

Zhu, Y., Cheung, A.L.M.2021. Proteoglycans and their functions in esophageal squamous cell carcinoma. World J Clin Oncol 12(7), 507–521. doi:10.5306/wjco.v12.i7.507

Ziller, M.J., Hansen, K.D., Meissner, A., *et al.*2015. Coverage recommendations for methylation analysis by whole genome bisulfite sequencing. Nat Methods 12(3), 230–232. doi:10.1038/nmeth.3152

Zito, A., Lualdi, M., Granata, P., *et al.*2021. Gene set enrichment analysis of interaction networks weighted by node centrality. Frontiers in Genetics 12

Supplementary Information

Supplementary Information: Chapter 2

The commands used for each step for NGS analysis for the entire project can be found using this link below.

Due to the size of the file Supplementary document 2.1 is in a digital word document file.

Link:

https://docs.google.com/document/d/1SV0XU7J8AMpaNjO0JO6nn8XgJmkEoZvE/edit

Raw data for methylome data in the form of COV files can be downloaded using this link:

Supplementary folder 2.1

Link:
https://drive.google.com/drive/folders/1SkF6CHirQzyyXnOhEc3fgrjaBw4iCwaO?usp=share_link

Raw data for RNA sequencing in the form of count files can be downloaded using this link:

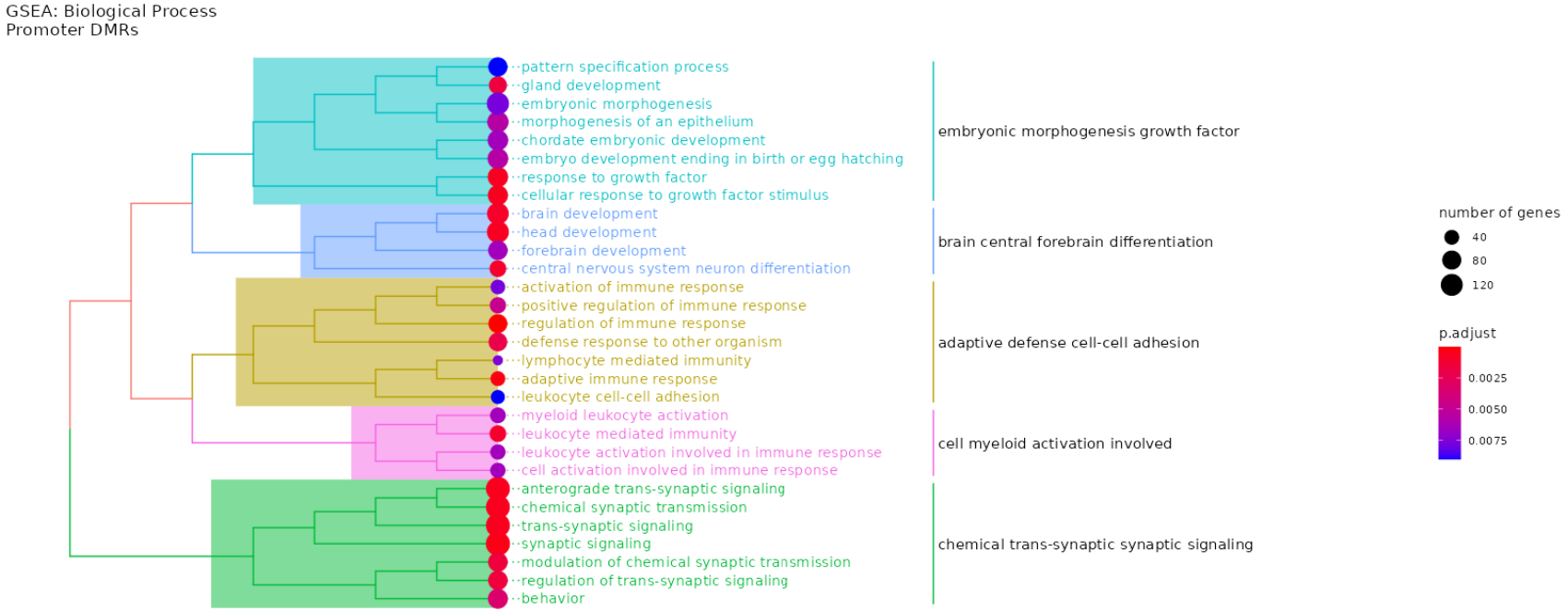
Supplementary folder 2.2

Link:
https://drive.google.com/drive/folders/1dIHV3HSdoriI3GXxzeOPiJP9kplZWnB1?usp=share_link

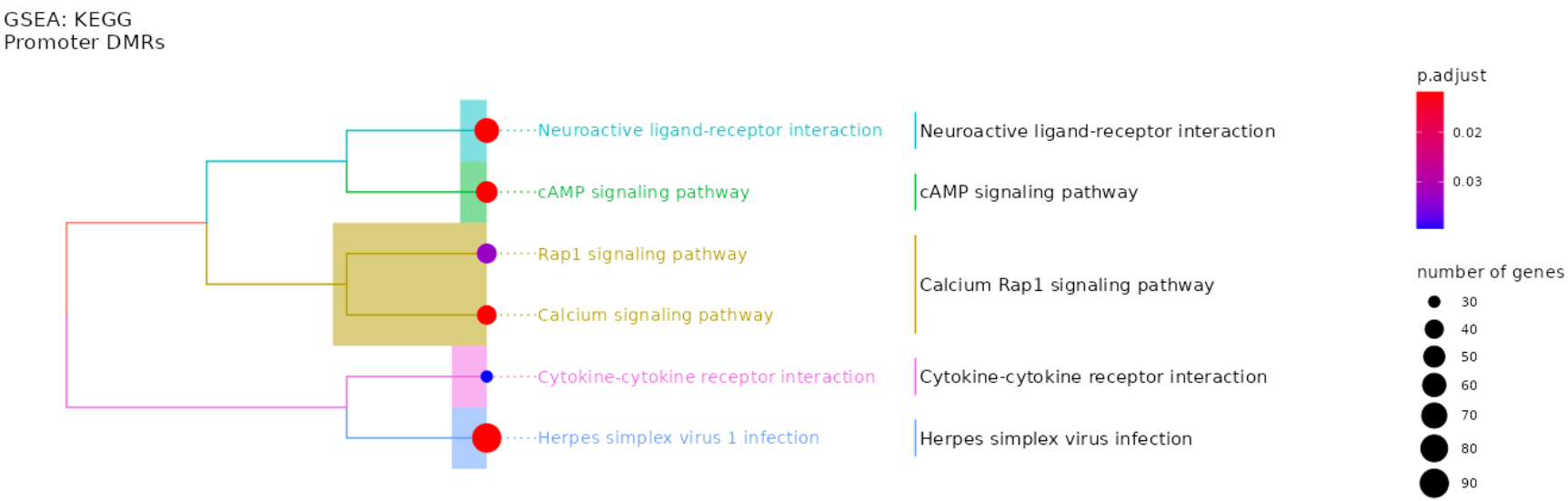
Supplementary Information: Chapter 4

Supplementary Figures

A



B


GSEA: KEGG
Promoter DMRs
Neuroactive ligand-receptor interaction
cAMP signaling pathway
Rap1 signaling pathway
Calcium signaling pathway
Cytokine-cytokine receptor interaction
Herpes simplex virus 1 infection
Neuroactive ligand-receptor interaction
cAMP signaling pathway
Calcium Rap1 signaling pathway
Cytokine-cytokine receptor interaction
Herpes simplex virus infection
p.adjust
0.02
0.03
number of genes
30
40
50
60
70
80
90

C

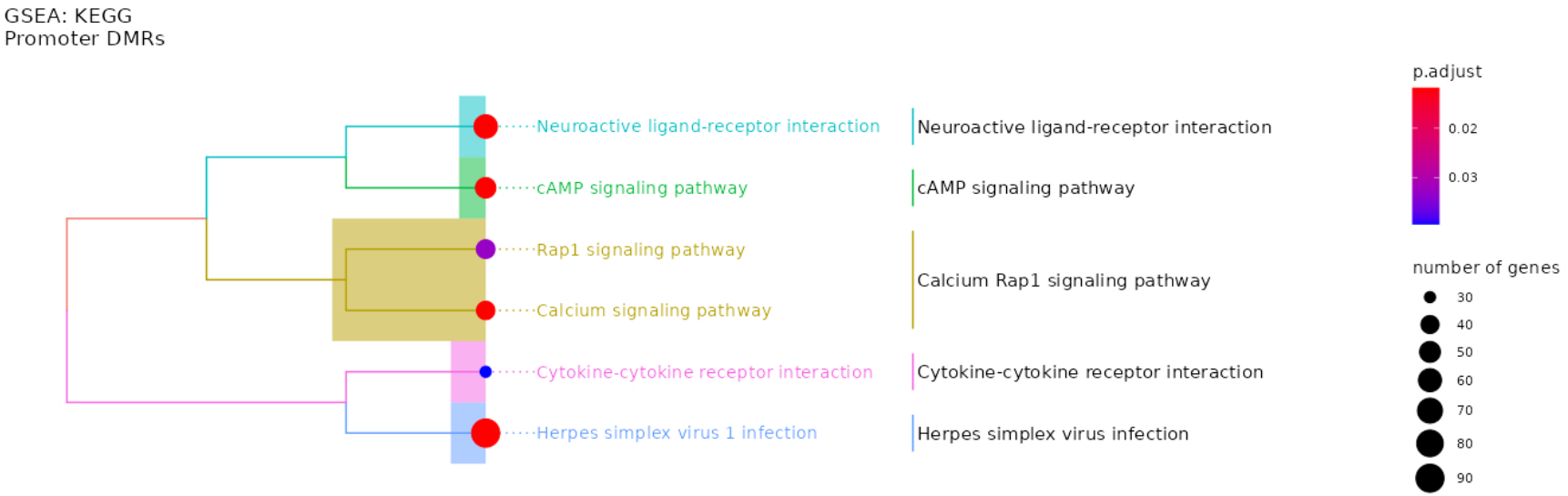
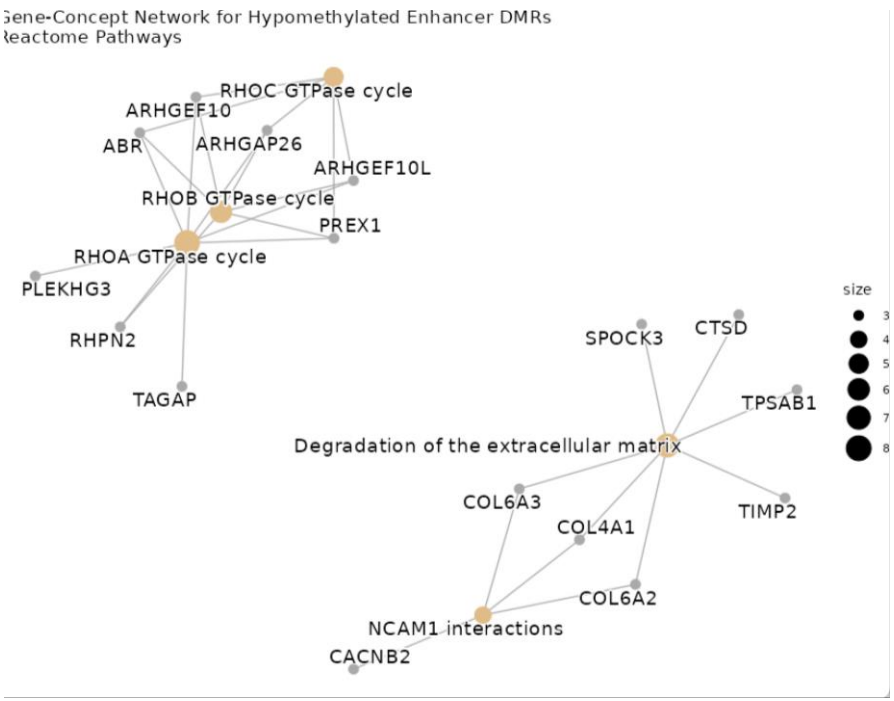


Figure S4- 1: Pathways analysis from methylome profiling Pathways associated with the most relevant BPs (A), KEGG (B) and Reactome (C) using differentially methylated promoters (20% methylation difference and q-value <0.01)

A



B

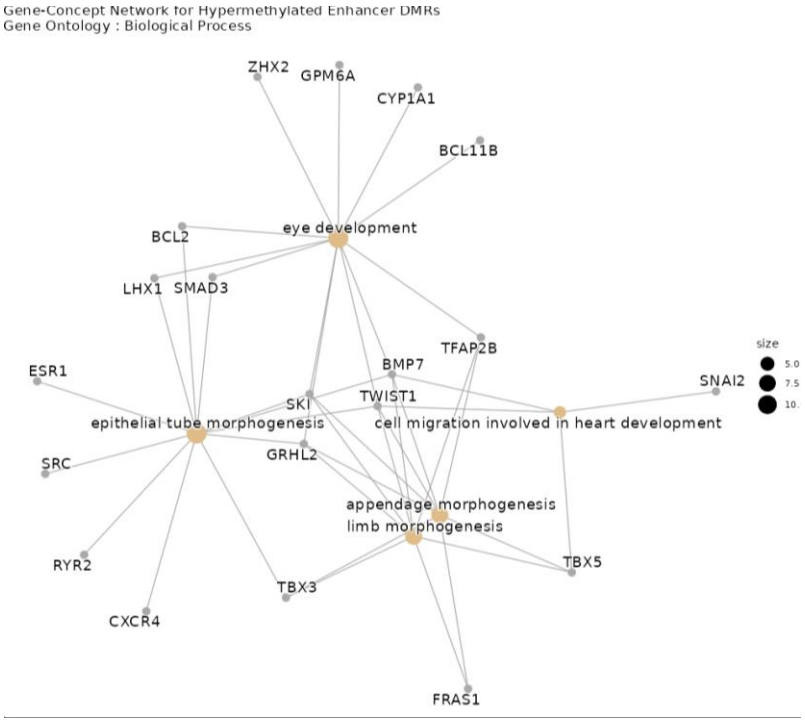
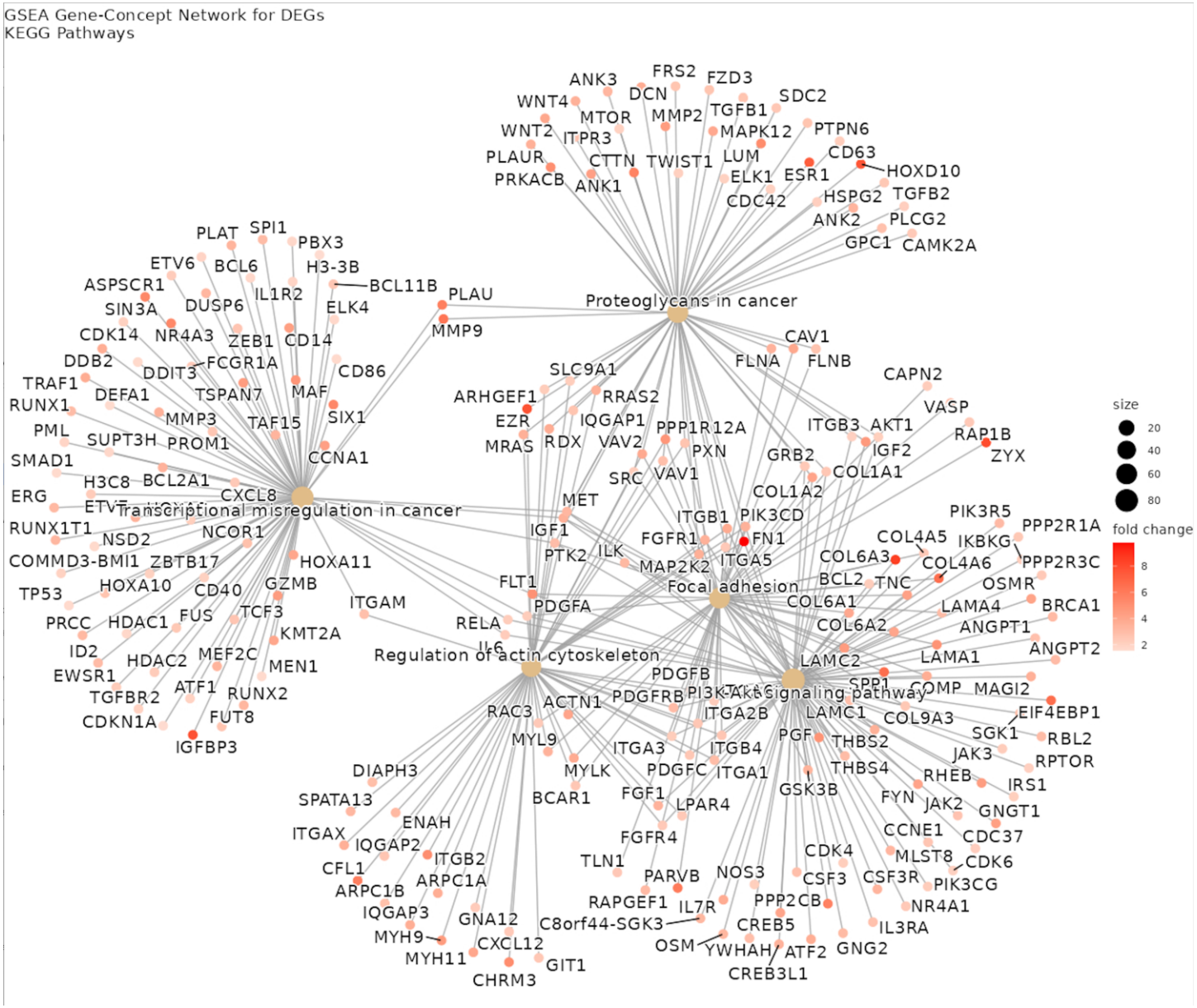


Figure S4- 2: Enhancer Pathway Analysis Supplementary Figure S4- 2 Enhancer DMR pathway enrichment. Most enriched pathways from hypomethylated (A) enhancers identified with Reactome. Most enriched pathways from hypermethylated (B) enhancers identified with Gene Ontology: Biological Process.

A


GSEA Gene-Concept Network for DEGs
KEGG Pathways
Proteoglycans in cancer
Transcriptional misregulation in cancer
Focal adhesion
Regulation of actin cytoskeleton
PI3K-Akt signaling pathway
size
20
40
60
80
fold change
8
6
4
2

B

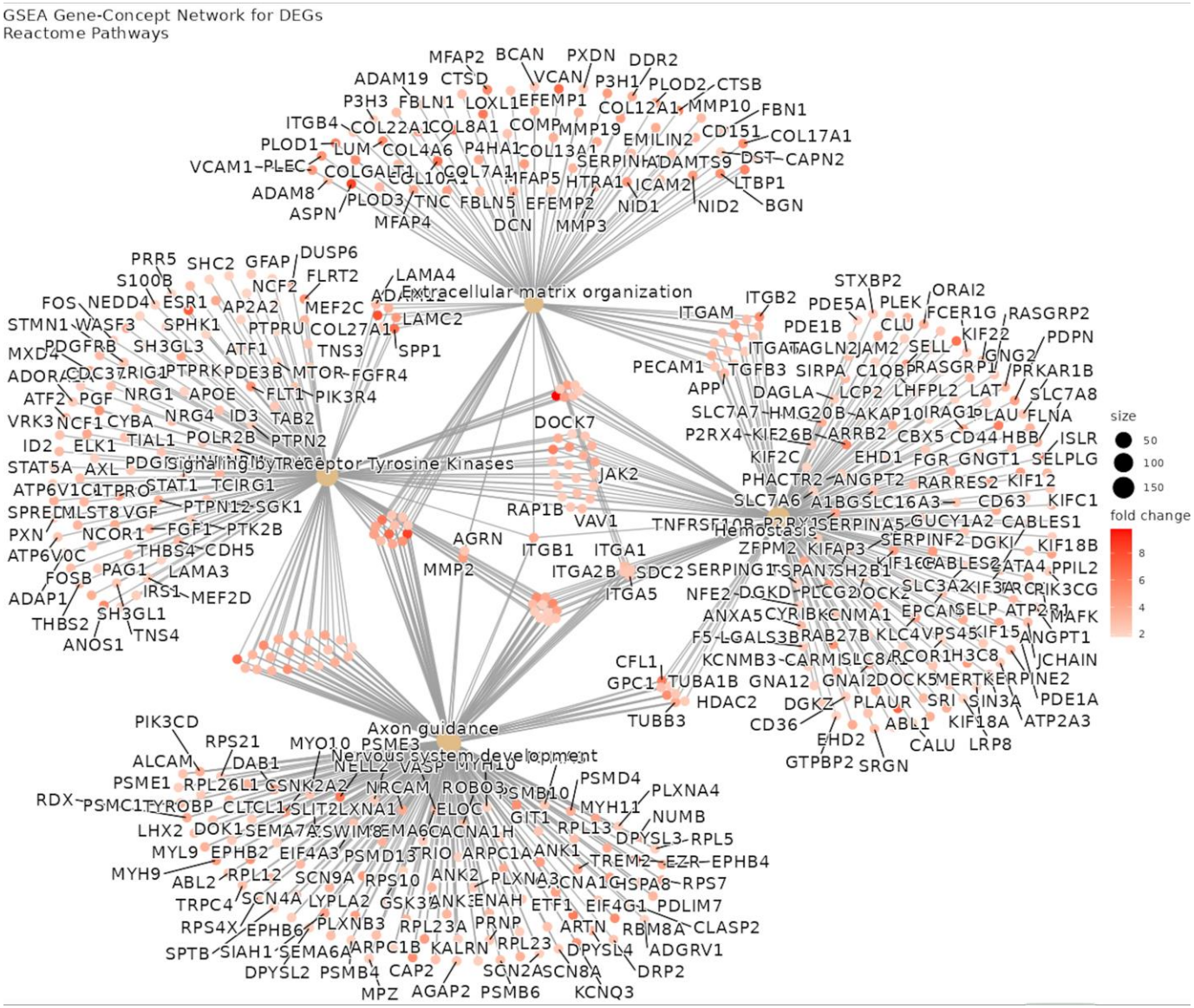


Figure S4- 3: Pathways enriched after transcriptome analysis A KEGG and B Reactome

Supplementary Tables

Supplementary table 4.1:

Enhancers in the vicinity of genes which displayed ≥10% difference in methylation between normal and tumour group.

Meth.diff = methylation difference

n.cpgs = number of CpGs

UTR = Untranslated Region

Due to the large size of these Supplementary tables, the tables are provided in digital link (excel file).

Link: https://drive.google.com/drive/folders/1dIHV3HSdoriI3GXxzeOPiJP9kplZWnB1?usp=share_link

Supplementary table 4.2:

Tab 1: Integration of gene promoter methylation levels with RNA expression between normal and tumour group.

Tab 2: Integration of gene enhancer methylation levels with RNA expression between normal and tumour group.

Seg.mean = segment mean methylation (percentage)

Seg.group = segment group (hypomethylated or hypermethylated)

abs.diff = absolute difference in methylation

fdr = false discovery rate

Link: https://docs.google.com/spreadsheets/d/1hLceNRDsSK1-A9oYKeJoFCQZmOZHqGnH/edit?usp=share_link&ouid=115380236071882997506&rtpof=true&sd=true

Supplementary table 4.3:

Tab 1: Hypo-up promoters.

Tab 2: Hyper-down promoters.

Tab 3: Hypo-up enhancers.

Tab 4: Hyper-down enhancers.

Link:
https://docs.google.com/spreadsheets/d/1OVd2IHLZSgWlMIid_U3b9ifY5CWC6xxk/edit?usp=share_link&ouid=115380236071882997506&rtpof=true&sd=true

Supplementary Information: Chapter 5

Supplementary Figure

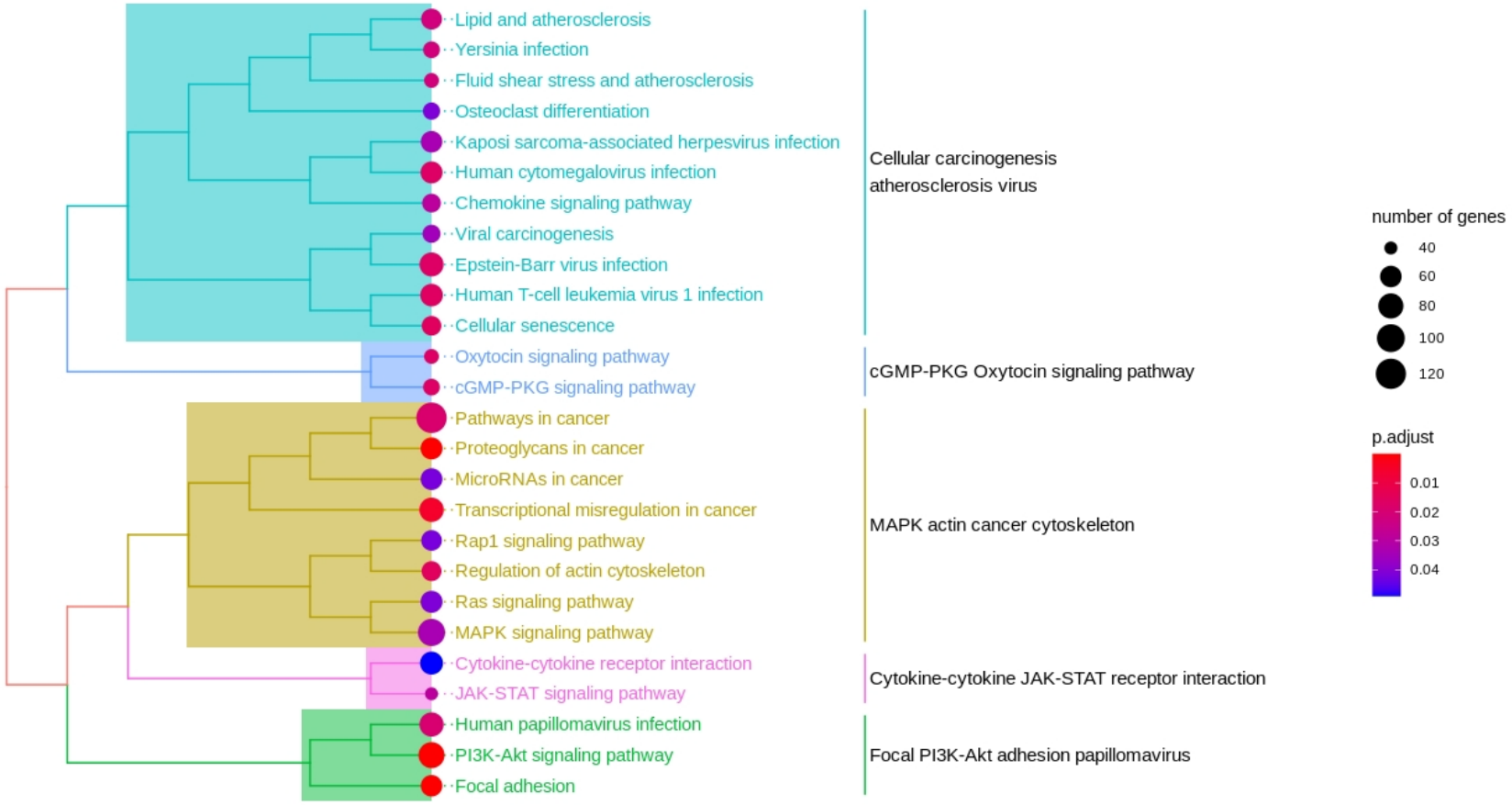


Figure S5- 1: Pathways enriched from transcriptome data across KEGG

Supplementary table

Supplementary table 5.1:
List of all genes which had a 2-fold change in expression level (up or down) and a significant FDR (≤0.05)

Due to the large size of this table, it is provided in digital link (excel file).

Link:
https://docs.google.com/spreadsheets/d/1gFW4D9MRJO_f9zLDcEJTN2DUKn1mFma9/edit?usp=share_link&ouid=115380236071882997506&rtpof=true&sd=true

Data availability

The data produced in this study will be made accessible through the Gene Expression Omnibus (GEO) repository.

Appendices

Appendix A – Ethics Certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Mr Mishalan Moodley et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M181188

NAME: (Principal Investigator)	Mr Mishalan Moodley et al
DEPARTMENT:	Haematology and Molecular Medicine Charlotte Maxeke Johannesburg Academic Hospital
PROJECT TITLE:	Analysis of the Oesophageal Squamous cell Carcinoma Methylome in South African Patients
DATE CONSIDERED:	30/11/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr P Willem and Prof C Matthew
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	15/02/2019

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

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in November and will therefore be due in the month of November each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

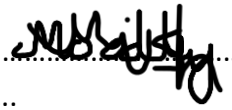
Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix B – Co-author Agreement form/ Supervisor Declaration

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

I, Mishalan Moodley, student number 0108860t, declare that this thesis is my own work and that I contributed adequately towards research findings published in the article(s) stated below which are included in my thesis.

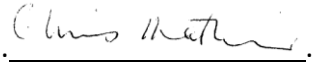
Signature of Student ……

Date…18/12/2023…….

Name of Primary Supervisor: Dr. Pascale Willem …….

Signature of Primary Supervisor ……**Date: 18/12/2023…..**……..

Name of Co Supervisor: Professor Christopher Mathew…….

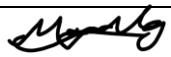
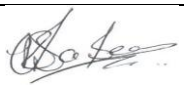
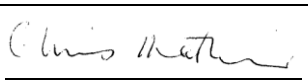
Signature of Co Supervisor … **Date: 18/12/2023**……

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of his thesis. In cases where the student is not the 1st author of a published article, the primary supervisor must explain (under comments) why the student is entitled to use the paper for his degree purposes.

Article 1:

Title: Processing and analysis of blood and tissue samples from oesophageal cancer patients in an African setting.

Journal name, year, volume and page numbers: Biopreservation and Biobanking. 2022;20(2):185-94.

Author	Name	Signature	Date
Ferndale	Lucien		23/11/2023
Moodley	Mishalan		18/12/2023
Chen	Wenlong		23/11/2023
Wadee	Reubina		21/22/2023
Wright	Colleen		22/11/2023
Parker	Mohamed Iqbal		
Willem	Pascale		18/12/2023
Mathew	Christopher		18/12/2023

Appendix C – Published Manuscripts

BIOPRESERVATION AND BIOBANKING
Volume 20, Number 2, 2022
© Mary Ann Liebert, Inc.
DOI: 10.1089/bio.2021.0030

Processing and Analysis of Tissue Samples from Esophageal Cancer Patients in an African Setting

Lucien Ferndale,[1,2,*] Mishalan Moodley,[3,*] Wenlong C. Chen,[4,5] Reubina Wadee,[6] Colleen A. Wright,[6,7] Mohamed Iqbal Parker,[8] Pascale Willem,[3] and Christopher G. Mathew[5,9]

Although infectious diseases continue to present a major health care problem in Africa, the incidence of cancer is increasing rapidly on the African continent and this merits an increased investment in cancer research in low to medium resource settings. Esophageal squamous cell carcinoma (ESCC) has a high incidence in Eastern and Southern Africa, with late clinical presentation and a very poor prognosis. There is limited research on the molecular pathology of this cancer in Africa, partly as a result of a lack of infrastructure for biobanking and sample processing in many African countries. The aim of this study was to establish a practical and robust workflow to collect, store, and process esophageal cancer samples such that both the tissue architecture and quality of the samples would be preserved and suitable for future genomic research. We developed a workflow that allows storage of fresh biopsy tissue in sterile Eppendorf tubes containing RNAlater, an efficient RNAse inhibitor. We collected 142 ESCC biopsy samples and showed that storage in RNAlater for up to 18 months did not alter tissue morphology, thus allowing histologic assessment by experienced pathologists and determination of tumor content in each biopsied sample. DNA and RNA extracted from tissue samples was assessed for purity, molecular size, and yield. The quantity and quality of nucleic acids obtained were suitable for genomic applications, and whole-exome sequencing of DNA from tumor tissues produced sequence data with a high proportion of both usable reads and correct base calling. We conclude that this workflow may be applicable to a wide range of malignancies for future genomic research in low-resource settings.

Keywords: esophageal, cancer, sample, processing, analysis

Introduction

Cancer is a growing health care problem worldwide and particularly in Africa, which is predicted to be the continent with the fastest increase in cancer incidence over the next decade.[1,2] Despite this, the focus of research in Africa has been mainly on communicable diseases, such as Acquired Human Immune Deficiency Syndrome and tuberculosis.[3,4] Noncommunicable diseases in Africa have enjoyed less attention, although this appears to be changing.[5,6] Health care systems in Africa are largely geared toward addressing the impact of communicable diseases, leaving them ill equipped to deal with the ever-increasing burden of chronic diseases.[7] Even among publications on

[1]Department of Surgery, Grey's Hospital, Pietermaritzburg, South Africa.
[2]Department of Surgery, College of Health Sciences, School of Clinical Medicine, University of KwaZulu-Natal, Durban, South Africa.
[3]Department of Molecular Medicine and Haematology, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg and the National Health Laboratory Service, Johannesburg, South Africa.
[4]National Cancer Registry, National Health Laboratory Service, Johannesburg, South Africa.
[5]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
[6]Department of Anatomical Pathology, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg and the National Health Laboratory Service, Johannesburg, South Africa.
[7]Lancet Laboratories, Johannesburg, South Africa.
[8]Division of Medical Biochemistry and Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa.
[9]Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
*Equal contribution.

noncommunicable diseases, cancer-related research is less prolific than cardiovascular-related research.[8]

Esophageal cancer, particularly esophageal squamous cell carcinoma (ESCC), is considered to be endemic in certain parts of Africa.[9,10] A corridor, which extends along the eastern side of Africa from Ethiopia to South Africa, is known as a high-incidence area of esophageal cancer, with ESCC as the predominant subtype.[11] Despite the high prevalence, there is limited literature on clinical studies of ESCC and of genomic profiles of ESCC tumors in Africa.[12–14] In contrast, significant progress has been made in genomic research in ESCC in non-African countries, particularly in China.[15] The success of these studies is based on (1) large sample sizes, which have the power to identify the potential driver genes, that is, genes which have significant rates of somatic mutations in tumors, and (2) application of high-throughput genomic technology such as whole-exome and RNA sequencing of tumors.

South Africa (SA) and other African countries, on the other hand, face major challenges which stem largely from resource constraints. These challenges include a lack of human resources dedicated to research, inadequate infrastructure for biobanking, and most importantly, very limited funding.[16] The establishment of a biobank with nationally accepted, validated protocols, and quality control measures is still an emerging concept in Africa, and as a result, there is a lack of protocols for biobanking that are validated in this setting.[17,18]

Following protocols established by facilities in first-world countries, while desirable, may not be feasible in an African setting. There is, therefore, a need to develop research protocols that are functional but also feasible to follow in a resource-constrained environment. Establishing robust workflows for biobanking of esophageal cancer tissue samples in South Africa would provide essential tools to facilitate good-quality translational research and contribute to improving the quality of care for patients afflicted with this devastating disease. It could also serve as a template for the collection and processing of tissue samples for genomic research in other African cancers.

We set out to assess the feasibility of performing genomic studies in patients with esophageal cancer from centers with significant resource constraints in two provinces in South Africa by establishing procedures for collection, storage, transport, and processing of tumor samples collected from patients with squamous cell carcinoma of the esophagus. We also assessed the quality of the genetic material obtained from tissue samples with the aim to develop a workflow for sampling and processing that is feasible in this setting.

Materials and Methods

Patient recruitment

Patients were recruited from two centers in two provinces of SA, namely Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) in Gauteng, which is a high-resource setting, and Grey's Hospital (GH) in KwaZulu-Natal, a low-resource setting. CMJAH offers a range of tertiary and specialized services and also serves patients from neighboring provinces. ESCC patients recruited from CMJAH mainly reside in the urban areas of Gauteng. GH is a tertiary hospital providing health care to one of three areas in KwaZulu-Natal. The area covers the western part of KwaZulu-Natal and has a population of about three million, of which approximately two-thirds are from rural areas.[19,20]

Potential ESCC patients arriving for endoscopy at CMJAH and at GH were eligible for recruitment into this study. Only patients with histologically confirmed ESCC were then enrolled. The purpose of the study was explained to each patient in their native language. Informed consent was obtained from each patient, who then completed a study questionnaire before their consultation with the surgeons. A research nurse facilitated the patient enrollment, consenting, and compiling the study questionnaire at CMJAH, while the study clinician (L.F.) and research assistant facilitated the same processes at GH. Ethics clearance for this study was obtained from the Human Research Ethics Committee (Medical) at the University of the Witwatersrand (Wits) (Certificate number M170871) and from the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (UKZN) (Certificate number BF270/15).

Sample collection and storage

The International Agency for Research and Cancer (IARC) protocol was used as a guideline for collecting and processing ESCC biospecimens at CMJAH and GH.[21]

Figure 1 illustrates the procedure used to collect, process, and store ESCC biopsies from CMJAH and GH.

Routine biopsies of esophageal tissue were taken from patients suspected to have ESCC for diagnostic purposes. For this study, additional tissue specimen biopsies were obtained during the routine procedure. At CMJAH, an additional biopsy specimen in the esophagus of each suspected tumor was obtained. At GH 2–3 additional biopsy specimens from the suspected tumor were obtained.

Histology preparation for FFPE and Frozen section

Biopsy specimens from the CMJAH samples were placed in a 2-mL Cryotube immediately after extraction and the Cryotube was then immersed in a vacuum flask containing liquid nitrogen. The biopsy was then placed in optimal cutting temperature (OCT) compound (Tissue-Tek®) and stored at −80°C. Biopsies were transported in batches of 10, in liquid nitrogen, to a histopathology laboratory off-site for histological processing. Each tissue specimen was divided into two pieces (25%:75%), with the 25% portion being cut from the OCT embedded frozen block and used for histology. The remaining biopsy (75% of the OCT embedded frozen block) was stored at −80°C until nucleic acid extraction.

A different protocol was used for the GH samples. Tissue specimens were placed immediately into 0.75–1 mL of RNAlater (Qiagen) and kept at room temperature for 30 minutes to 6 hours and then stored overnight in RNAlater at 5°C before being transferred to a −20°C freezer. The specimens were stored at −20°C for 1 to 18 months, then batched, and transported at 5°C (temperature controlled) to CMJAH. Upon arrival at CMJAH, the specimens were stored at −20°C. Before processing, the specimens were taken out of the −20°C freezer, kept at room temperature for 24 hours, after which the RNAlater was removed. Each biopsy specimen was cut into two equal (50%:50%) portions in sterile conditions. One-half of the specimens were placed in 10% buffered

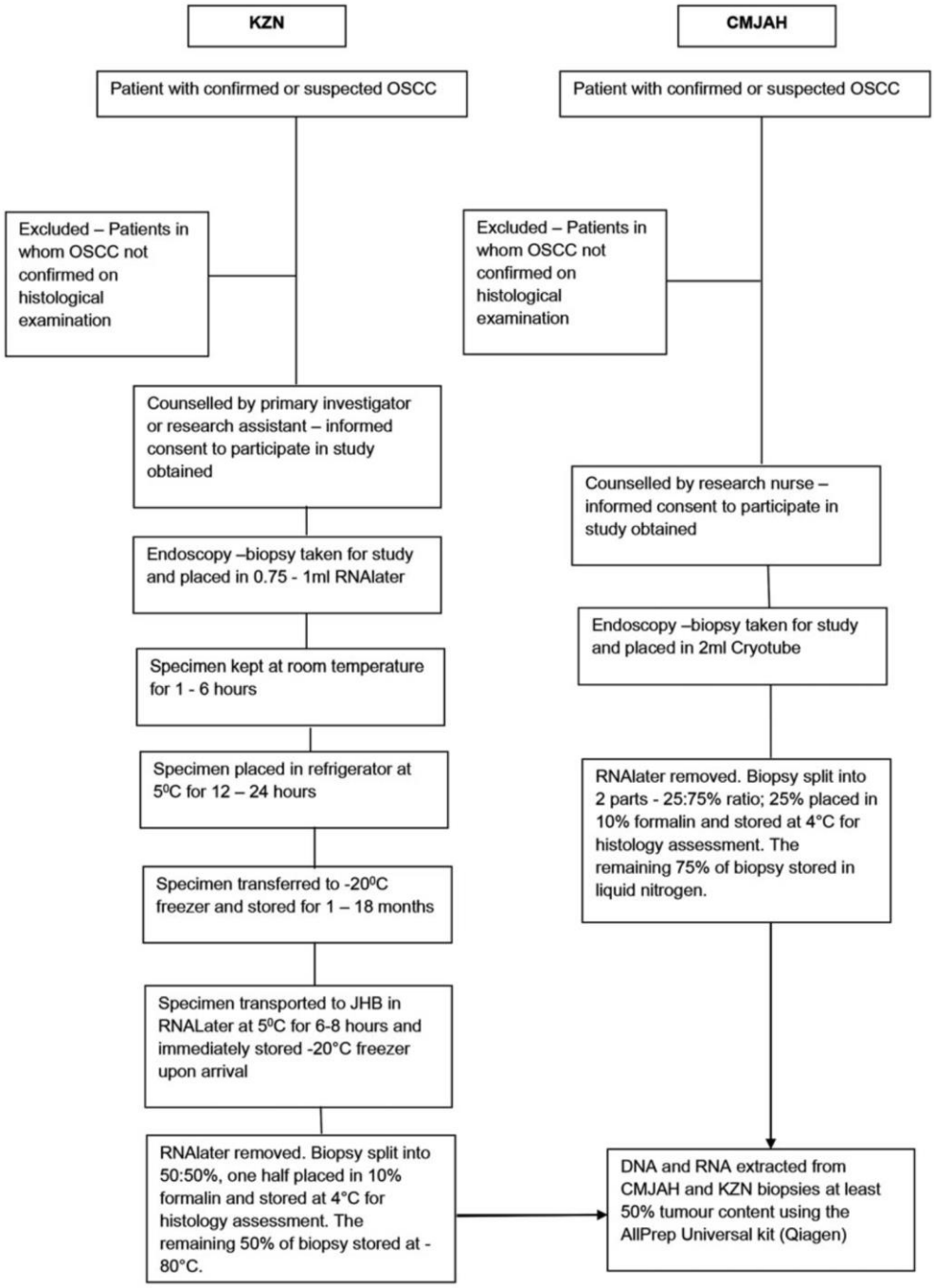


FIG. 1. Protocols followed by the two sites for collection, storage, and processing of samples.

formalin and batched for transportation to an offsite histopathology laboratory for histological processing. The other half was immediately placed in a 1.5-mL Eppendorf tube and stored at −80°C for nucleic acid extraction.

For histological processing, all tissue specimens fixed in formalin (at 6–8°C) were washed repeatedly in different levels of ethanol of increasing concentrations from 70% to 100%. Xylene was then added to each specimen, followed by paraffin wax at 60°C. The tissue specimen was then cooled to allow solidification. The tissue was embedded into a cassette and serial sections cut at 5 µm and placed on a single glass slide and stained with Hematoxylin and Eosin

(H&E). The ESCC tissue specimens were sectioned in a Leica microtome cryostat using OCT compound to embed the tissue samples. Four sections of 10 μm each were mounted onto a single glass slide and stained with H&E.

All tissue sections were then examined microscopically by two experienced histopathologists. The tissue sections were evaluated for the following parameters: tissue type, tumor percentage, necrosis, inflammation, ulceration, and the presence or absence of overlying surface epithelial dysplasia. The percentage of surface epithelial dysplasia was quantified. The tumor was assessed for the degree of differentiation and the presence or absence of lymphovascular invasion. A consensus assessment was rendered for each case.

All biopsy specimens from both sites were stored in an upright position. All refrigeration equipment at the CMJAH site was monitored daily with an alarm-linked notification system. Backup refrigeration equipment was available if primary refrigeration equipment failure occurred. At GH there was no alarm system or back-up refrigeration equipment, but the refrigeration equipment was connected to a backup generator and uninterrupted power supply device.

Tissue DNA and RNA isolation

Tissue DNA (tsDNA) and tissue RNA (tsRNA) was extracted from a subset of tissue specimens with substantial tumor content. All tissue specimens were extracted using the Qiagen AllPrep® Universal Kit (Qiagen) according to the manufacturer's protocol. This kit was chosen for its ability to extract high-quality nucleic acid, which is needed for molecular applications such as next-generation sequencing (NGS). The entire remaining tissue biopsies were used for tsDNA and tsRNA extraction.

Tissue from each specimen was added to 350 μL Qiagen buffer RLT (Qiagen) and lysed at 30 mhz for 45 seconds using the TissueLyser II (Qiagen®). The lysate was pipetted onto an AllPrep DNA spin column (Qiagen) for tsDNA binding. The column was centrifuged at 8000×RCF for 1 minute to collect the flow-through tsRNA in the collection tube. Since RNA is more sensitive to degradation, the tsRNA was extracted before the tsDNA. RNA flow through was transferred to an RNeasy spin column. Both nucleic acid specimens were extracted according to the Qiagen Allprep DNA/RNA/miRNA Universal Kit (Qiagen). The tsDNA was eluted in 35 μL of elution buffer and the tsRNA was eluted in 60 μL of buffer.

Assessment of tsDNA and tsRNA quality

Nucleic acid quantity and quality were assessed with the Qubit (Thermo Scientific), the 2100 Bioanalyser (Agilent Technologies, CA), and the NanoDrop™ 1000 (Thermo Scientific) spectrophotometer. Qubit was used to determine DNA and RNA concentrations. The NanoDrop 1000 was used to determine the absorbance ratios at 260 and 280 nm. A 260/280 ratio of >1.8 for DNA and >2 for RNA was used to indicate good purity. The Agilent 2100 Bioanalyser was used to determine RNA quality. RNA was extracted from a total of 49 biopsies, which were prioritized as having substantial proportions of tumor or dysplasia content. RNA quality was measured using the LabChip GX RNA Assay (PerkinElmer) according to the manufacturer's instructions, using the RNA integrity number equivalent (RINe) score, which has a scale of 1–10 with 10 being the highest quality.[22] tsRNA and tsDNA samples were stored at −80°C.

Whole-exome sequencing

Exome sequencing of tumor DNA with matched germline DNA was performed on samples from six ESCC patients. A total amount of 500 ng of DNA from each sample was used as the starting material. DNA was fragmented using the Covaris shearing system (Covaris). Fragments that were generated ranged between 180 and 280 bp. Exonucleases removed 5′ and 3′ ends to ensure that each fragment had a blunt end. DNA was end-repaired after adenylation of the 3′ ends. Adapter molecules were ligated to both ends to ensure that DNA fragments were enriched in a polymerase chain reaction (PCR).

The exonic region of each sample was captured, enriched for PCR, and indexed to prepare for hybridization using the SureSelect Human All Exon V6 (Agilent Technologies). NGS of the library was performed on the Illumina HiSeq 2500 sequencer (Novogene, United Kingdom). The quality of the sequence data was assessed by calculation of the percentage of usable reads and the Phred Q score. A sequenced fragment is considered "usable" if it maps uniquely to the genome and remains after removing PCR duplicates. The Phred Q score is the quality score of a base, which is an integer value representing the estimated probability of an error. If P is the error probability, then $P = 10^{-Q/10}$, and $Q = -10 \log10 (P)$.

Personnel resources

GH, KwaZulu-Natal: Patients were consented and biopsies were collected and stored by one of the authors (L.F.) or one of two additional endoscopists (see Acknowledgments section). Biopies for research were collected at the same time as diagnostic biopsies and placed in RNALater in the endoscopy suite. Samples were then placed in a refrigerator overnight and in a −20°C freezer the next morning by L.F. Thereafter, the specimens were transported in batches to the laboratory in Gauteng.

CMJAH, Gauteng: Patients were consented by one research nurse and biopsies were collected by one of three endoscopists (see Acknowledgments section). Samples were processed for histology and DNA/RNA extracted and subjected to quality control by a medical laboratory scientist (M.M.). Slides were assessed for tissue morphology and tumor content by two consultant histopathologists (R.W. and C.A.W.).

Statistical analysis

The Kruskal–Wallis nonparametric test was used to assess for differences in tsDNA and tsRNA quality (A260/280) between study sites, and between tissue storage media (RNAlater vs. liquid nitrogen). Spearman's rank-sum test was used to assess if there was any correlation between the number of days of sample storage and the quality of the tsDNA and tsRNA. The Mann–Whitney U test was used to assess differences in RNA quality score by sample storage method.

Results

A total of 142 patients were recruited into the study from the two study sites. Histological examination was performed

on tissue samples from all patients, and nucleic acids (DNA and RNA) were extracted from a total of 49 samples with substantial content of tumor or dysplastic tissue as part of our ongoing study on the genomic analysis of African esophageal cancer. The total number of biopsies collected at each site with the storage method and histology profile is shown in Table 1. Important differences between protocols followed at the two different sites included storage time, temperature, and storage medium. Samples at the GH were stored for long periods (several months) in RNAlater before downstream processing, while biopsies collected at the CMJAH were stored for shorter periods (minutes to hours) in liquid nitrogen before downstream processing.

Histological analysis of the tumor content of all biopsy samples was obtained and evaluated as described in the Methods. The quality of cellular morphology from all 142 ESCC specimens from CMJAH and GH and both the frozen section and RNAlater protocols was sufficient to allow assessment of all the required parameters, including tissue type, tumor percentage, and the percentage of surface epithelial dysplasia. As shown in Table 1, the GH protocol (RNAlater) produced a higher percentage of samples with histology positive for squamous cell carcinoma compared with the protocol followed at CMJAH (frozen section/liquid nitrogen). A substantial proportion of the research biopsy samples from CMJAH (50%) were assessed as dysplasia rather than invasive carcinoma, although all patients had a clinical diagnosis of ESCC.

Table 1. Number of Biopsies Collected at Each Site with Histology Profile, Nucleic Acid Yield, and Quality for Tissue DNA and Tissue RNA Isolated from Tissue Specimens

	CMJAH (Frozen section/Liquid nitrogen)	*GH (RNAlater)*
Number of biopsies collected at each site	50	92
Stored as frozen section	26	0
Stored in RNAlater	24	92
Squamous cell carcinoma	16 (32%)	60 (63%)
Dysplasia	25 (50%)	18 (19%)
Dysplasia and squamous cell carcinoma	1 (2%)	5 (5%)
Other	8 (16%)	9 (13%)
Number of biopsies that were used for nucleic acid extraction	37	12
Tissue size (cm)	0.3–0.5	0.5–1
Mean tsDNA yield/amount (μg)	4.5 μg	9.2 μg
Mean A260/280 tsDNA (standard deviation)	1.91 (0.13)	1.96 (0.11)
Mean tsRNA yield/amount (μg)	5.7 μg	11.5 μg
Mean A260/280 tsRNA (standard deviation)	2.02 (0.09)	2.03 (0.04)
Mean tissue storage time (days) (standard deviation)	92.8 (64.2)	650 (84.4)
Mean RINe score (standard deviation)	8.29 (1.67)	8.61 (1.59)

CMJAH, Charlotte Maxeke Johannesburg Academic Hospital; GH, Grey's Hospital; RINe, RNA integrity number equivalent; tsDNA, Tissue DNA; tsRNA, tissue RNA.

Representative photomicrographs of biopsy tissues from both centers are shown in Figure 2. The panels on the left half of the image (A, C, E, G) show a squamous cell carcinoma from a frozen section tissue biopsy while the panels on the right half (B, D, F, H) show, at the same set of magnifications, a squamous cell carcinoma that was placed in RNAlater. An extensive freeze artefact is apparent in images on the left panel, which restrict areas suitable for histopathological assessment. A freeze artefact with distortion of cellular detail is best shown in image E (3 thin arrows). The freeze artefact causes blurring of the epithelial/stromal interface, making assessment of invasion difficult. Also, there is no clarity of nuclear detail, which is critical for a definitive diagnosis of malignancy, and identification of mitoses is problematic. The panels on the right show islands of tumor cells with clear cellular morphology and well-formed keratin pearls (2 thin arrows in H).

Table 1 shows the yield (amount) and quality of the genetic material obtained from samples from the two centers. The yield of nucleic acids obtained from the GH samples was more than twice that of the CMJAH samples. This is likely due to the size of the biopsies obtained from GH, approximately twice the size of biopsies obtained from the CMJAH.

There were no statistically significant differences between the quality (A260/280) of tsDNA ($p=0.072$) and tsRNA ($p=0.625$) from the two study sites. The tsDNA from both sites was of good quality, with a mean A260/280 ratio of 1.91 and 1.96 for CMJAH and GH, respectively, and mean A260/280 ratios of 2.02 and 2.03 for tsRNA. The mean RINe scores for RNA samples for both protocols were over 8 and were not significantly different, with 43 of 49 (88%) of samples meeting the generally accepted metric (>7) required to produce high-quality RNA sequencing libraries.[22] An example of the gel electrophoresis of the tsDNA samples is shown in Figure 3. High-molecular-weight tsDNA was obtained from all samples from both study sites.

DNA extracted from a subset of six tissue samples and their paired germline DNA was submitted for whole-exome sequencing to determine whether the samples were of sufficient quality for genomic studies. The quality control analysis of the sequence data is shown in Table 2. The average number of usable sequence reads across the six tissue samples was 97.4%. The average proportion of bases called with a Phred score of Q30, which indicates a 99.9% probability that the base was called correctly, was 92.6%. These metrics indicate good-quality next-generation sequence data,[23,24] which was similar between samples sequenced from both sites. Furthermore, the quality of the sequence data from the tissue samples was very similar to DNA extracted from paired blood samples (data not shown).

The personnel required to execute the two protocols are described in the Methods. The additional effort required over and above that needed for a standard diagnostic workflow is taking additional biopsies at the same time as the diagnostic biopsies (endoscopist), then either freezing the research biopsy in liquid nitrogen and storing it at −80°C or placing it in RNAlater. The freezing protocol requires a trained laboratory assistant. Preparation and analysis of slides is required for both protocols and is done by histopathologists. Extraction of DNA and RNA from the tissue

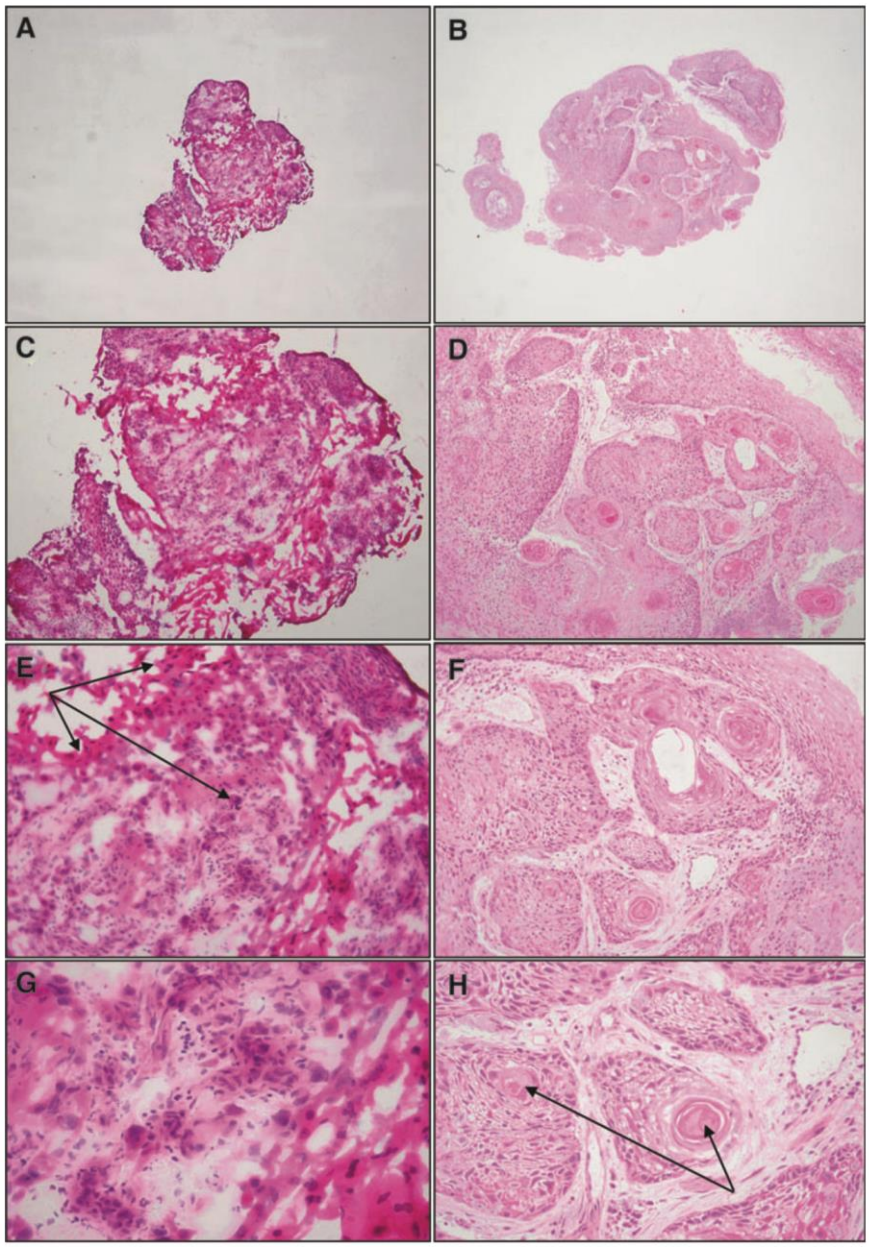


FIG. 2. Composite photomicrograph of esophageal invasive squamous cell carcinomas at increasing magnifications. Panels **(A, C, E, G)** squamous cell carcinoma from a frozen section tissue biopsy; Panels **(B, D, F, H)** squamous cell carcinoma placed in RNAlater. Extensive freeze artefacts noted in **(A, C, E, G)** with freeze artefact and distortion of cellular detail best shown in image **E** (thhree *thin arrows*). Panels **(B, D, F, H)** show islands of tumor cells with clear cellular morphology and well-formed keratin pearls (two *thin arrows* in **H**). Original magnification: **A** and **B**:×40, **C** and **D**:×100, **E** and **F**:×200, **G** and **H**:×400.

samples is done by a laboratory scientist. An estimate of the costs of reagents, personnel, and equipment for the two protocols in our two institutions is shown in Table 3. This shows that the running costs per sample are very similar for the two protocols. However, the frozen section protocol requires a substantial initial investment for equipment. The methods for extraction of DNA/RNA and subsequent quality control of the extracted nucleic acids are identical for the two protocols.

Discussion

This study was undertaken to ascertain whether genomic research could be carried out successfully in centers and study sites with significant resource constraints. In our study, one of the centers had no on-site genetic laboratory or staff, no on-site histopathology service and relied on a refrigerator/freezer, a supply of RNAlater, specimen collection tubes, and clinical staff to collect and store specimens. Despite these limitations, substantial proportions of tumor tissue could be identified in the majority of specimens on histological assessment, and genetic material (both DNA and RNA) could be extracted with acceptable quality and quantity, with the DNA being used successfully for NGS.

The first important consideration when performing genomic studies on tumor tissue is ensuring that tissue samples taken are representative of the tumor. Histopathological confirmation of tumor content in biopsy specimens is an

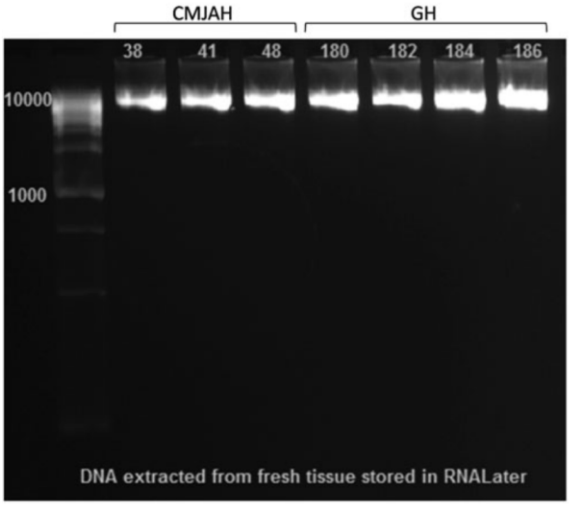


FIG. 3. DNA samples extracted from tumor tissues on a 1% agarose gel.

essential step in successful biobanking.[25] Since a wide disparity in tumor content may occur between biopsies, each specimen should be evaluated by a pathologist to confirm the presence of a substantial fraction of tumor tissue before using samples for costly genomic research.[26] This is further illustrated by our study, where tumor content varied from 15% to 100% between different samples. One of our initial concerns was that biopsies stored in RNAlater for long periods (>12 months) would lose the integrity of the tissue and thus their morphology and would not be suitable for histological analysis. However, histology results were obtained successfully from all 142 ESCC specimens from CMJAH (high resource) and GH (low resource) indicating that esophageal biopsies stored for up to 18 months did not affect the integrity of the biopsies so they could still be used for histological assessment. A previous study showed that storing tissue biopsies in RNAlater or OCT compound for 24 hours did not compromise histological, immunochemical, and genomic qualities.[27] However, another study reported that 45% of RNAlater samples had poor nuclear morphology preservation, rendering pathological evaluation challenging.[28] The protocol followed for samples collected at GH, where specimens were immersed in RNAlater and kept at room temperature for several hours, then at 5°C overnight before being stored at −20°C, still in RNAlater, may offer improved protection of tissue morphology.[26,29]

TABLE 2. QUALITY CONTROL STATISTICS FOR DNA EXOME SEQUENCING DATA FROM THE TISSUE DNA SAMPLES

Sample	*Center*	*Raw reads*	*Usable reads (%)*	*Q30 score (%)*
180 tDNA	GH	28,178,349	98.29	92.54
182 tDNA		33,810,584	97.15	92.62
186 tDNA		32,985,596	97.79	92.63
41 tDNA	CMJAH	37,575,723	96.89	92.34
48 tDNA		31,801,242	96.91	92.44
38 tDNA		37,587,499	97.26	92.89

TABLE 3. COST PER SAMPLE BREAKDOWN FOR THE TWO PROTOCOLS. EXCHANGE RATE OF 1 USD = 13.7400 ZAR WAS USED

Running cost	*CMJAH (Frozen section/Liquid nitrogen)*	*GH (RNAlater)*
Consumables per sample: (RNAlater, cryovials for GH; Liquid nitrogen, cryovials for CMJAH)	USD 1.69	USD 3.51
Histology costs per sample	USD 27.50	USD 27.50
Reagents for tsDNA/RNA extraction	USD 16.50	USD 16.50
Staff time for sample processing and extraction	USD 18.00	USD 18.00
Total running cost per sample	USD 63.69	USD 65.51
Fixed costs for equipment		
Liquid nitrogen 21 L storage tank	USD 1 515.47	
Liquid nitrogen flask (Dewar)	USD 283.17	
−80°C Ultra low temperature freezer	USD 5789.12	
Cryo gloves (1 pair for long-term usage)	USD 806.53	
−20°C chest freezer		USD 174.66
Total fixed costs	USD 6 245.50	USD 174.66
Sample shipment		
Dry ice shipment (100 samples)		USD 365.07

Sample shipment when relevant is estimated at USD 365 per batch of 100 samples.

Another important consideration when using RNAlater for storage is the size of the biopsy specimen since RNAlater has been shown to interfere with histological assessment in smaller samples.[25,28]

A major difference between the two centers was the method of specimen storage, with liquid nitrogen being preferred at CMJAH while RNAlater was used exclusively at GH. In our experience, the logistics used for storing and extracting nucleic acids from liquid nitrogen was more challenging compared with RNAlater. Liquid nitrogen must be available at the biopsy site, careful handling is needed to prevent injury to research staff, and a long-term liquid nitrogen storage facility must be available to maintain frozen sections. Somewhat surprisingly, we found that the quality of the nuclear morphology was superior in the RNAlater samples compared with the samples prepared in liquid nitrogen. The use of liquid nitrogen can make histological assessment more difficult since specimens are fragile and

more difficult to section.[26] In our study, we found that certain samples had freeze artefacts, which can lead to an inconclusive interpretation.[30] RNAlater is also easier to use within a busy clinical environment and a low-resource setting where clinicians would not always have immediate access to refrigerators or freezers.

There was a difference in the percentage tumor content between the two centers. As shown in Table 1, specimens from GH had a substantially higher rate of squamous cell carcinoma-positive cases (63%) compared with patients from CMJAH (32%). The reason for the higher tumor content from GH samples may be related to the biopsy size, since samples from GH were larger than those from CMJAH despite the same-size forceps being used at both centers (Table 1). Another possible reason may be the experience of the endoscopist performing the biopsy. At GH, the biopsies were collected either by the primary investigator or by one of two other endoscopists, all of whom have several years of endoscopy experience. In contrast, at CMJAH, biopsies were collected by several different clinicians and in most instances by trainee surgeons with less experience, who might have been less effective in deep sampling of the core of the tumor.

Quality control of DNA and RNA samples extracted from tumor tissue showed similar A260/280 ratios between the two centers, while the mean yield of DNA and RNA extracted from GH cases was about twice as high as compared with yields obtained from CMJAH cases. This could also be explained by the fact that biopsies from GH were larger (0.5–1 cm for KZN compared with 0.3–0.5 cm for CMJAH). As mentioned previously, this is most likely due to the fact that the majority of biopsies were taken by a single very experienced endoscopist at GH (L.F.) in contrast to less experienced endoscopists at CMJAH. However, the quality and quantity of DNA and RNA from both sites were adequate for genomic studies such as NGS.

There was a marked difference in the storage time of the tissue from the two different sites (Table 1), which was necessary since samples at GH were batched for shipment to CMJAH. Our study demonstrates that despite a storage time of several months, the quality of genetic material is not compromised. This is an important consideration in centers without on-site laboratories such as GH. The absence of these facilities does not, therefore, preclude genomic research. The quality scores from the exome sequencing data are an indication that DNA extracted from tissue samples at both sites generated high-quality sequences whether biopsies were stored in RNAlater or liquid nitrogen.

The personnel resources required for the two protocols are very similar, except that a trained laboratory assistant is required for freezing samples in liquid nitrogen, whereas the biopsies can be placed in tubes containing RNAlater by the endoscopist. Also, although the use of liquid nitrogen has a slightly lower consumable cost than RNAlater, it requires a substantial initial investment for equipment. The only equipment required for storage in RNAlater by GH is a –20°C freezer, thus the RNAlater protocol has a significant advantage in settings, where there is a lack of any existing storage facilities. This is particularly common in resource-constrained settings. However, these estimates come with the *caveat* that they are based on our institutions and may vary considerably in other settings.

In conclusion, our results indicate that it is possible to produce good-quality samples for genomic research in an environment lacking on-site laboratory facilities geared toward this type of research without compromising important preanalytical variables that may influence results. The use of RNAlater, the availability of a standard refrigerator, and –20°C freezer, as well as collaboration with a center with the appropriate laboratory facilities and staff are essential. A recommended workflow for tissue biobanking in centers with limited resources is shown in Figure 4. We suggest that this workflow is both practical and cost-effective, and thus feasible for use in resource-constrained settings. It may also be applicable to other types of tissues and tumors, subject to the outcome of the kind of quality control assessments outlined in this study.

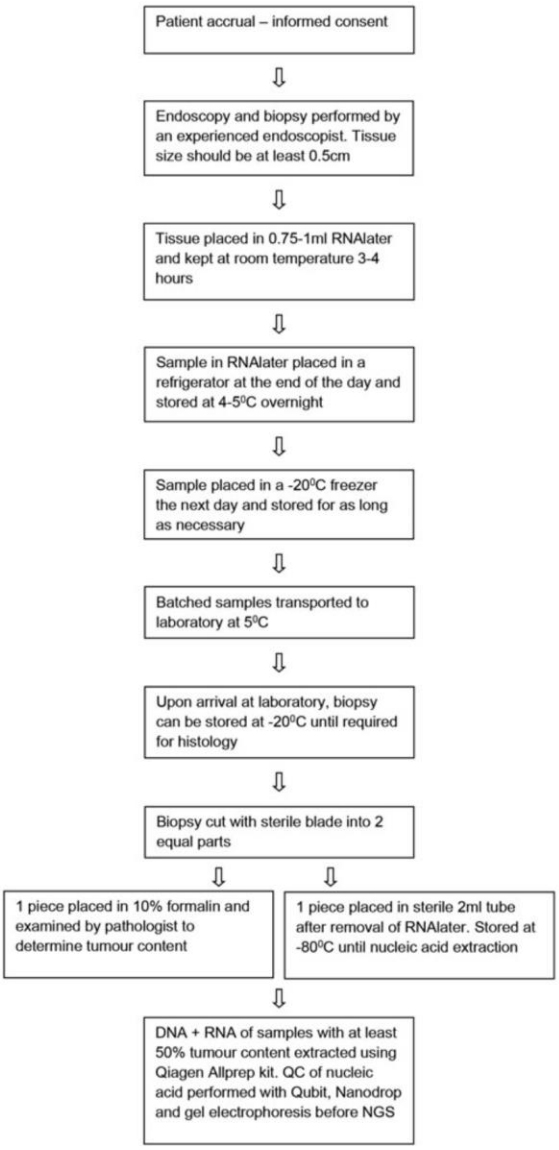


FIG. 4. Recommended operational workflow for tissue collection in a low-resource setting.

Acknowledgments

The authors thank Sister Gloria Mokwatle and Surgeons Dr. U Khan, Dr. LM Prodehl, and Dr. TK Marumo for recruitment of patients at the Charlotte Maxeke Johannesburg Academic Hospital. They thank Dr. M Govender and Dr. V Nair for assistance with obtaining tissue samples at Grey's Hospital.

Authors' Contributions

M.M. received and processed samples for nucleic acid extraction, curated sample information, and contributed to the article; L.F. recruited patients, performed biopsies, and wrote the article; W.C.C. assisted with the logistics of sample collection, transport and processing, statistical analyses, and the editing of the article; R.W. and C.A.W. reviewed the histopathology of samples and the article; P.W. oversaw sample quality, database curation, and contributed to the article; M.I.P. contributed to study design, funding, and review of the article; C.G.M. contributed to the study design, supervision of the study, and the writing of the article.

Author Disclosure Statement

No conflicting financial interests exist.

Funding Information

This work was supported by grant #046 (M.I.P., C.G.M., M.M., P.W.) from the South African Medical Research Council (SAMRC) with funds received from the South African National Department of Health, the Medical Research Council (United Kingdom) with funds from the UK Government's Newton Fund and GSK; Self-initiated research grant from the SAMRC (L.F.); National Research Foundation (SA) Thuthuka grant (W.C.C.); and the Cancer Association of South Africa—CANSA (C.G.M.).

References

1. Sylla BS, Wild CP. A million africans a year dying from cancer by 2030: What can cancer research and control offer to the continent? Int J Cancer 2012;130:245–250.
2. Global Burden of Disease Study 2013 Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015; 386:743–800.
3. Adedokun BO, Olopade CO, Olopade OI. Building local capacity for genomics research in Africa: Recommendations from analysis of publications in Sub-Saharan Africa from 2004 to 2013. Glob Health Action 2016;9:31026.
4. Hofman KJ, Kanyengo CW, Rapp BA, Kotzin S. Mapping the health research landscape in Sub-Saharan Africa: A study of trends in biomedical publications. J Med Libr Assoc 2009;97:41–44.
5. Olesen OF, Parker MI. Health research in Africa: Getting priorities right. Trop Med Int Health 2012;17:1048–1052.
6. Kane J, Landes M, Carroll C, Nolen A, Sodhi S. A systematic review of primary care models for noncommunicable disease interventions in Sub-Saharan Africa. BMC Fam Pract 2017;18:46.
7. de-Graft Aikins A, Unwin N, Agyemang C, Allotey P, Campbell C, Arhinful D. Tackling Africa's chronic disease burden: From the local to the global. Global Health 2010;6:5.
8. Hofman K, Ryce A, Prudhomme W, Kotzin S. Reporting of non-communicable disease research in low- and middle-income countries: A pilot bibliometric analysis. J Med Libr Assoc 2006;94:415–420.
9. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136:E359–E386.
10. Parkin DM, Bray F, Ferlay J, Jemal A. Cancer in Africa 2012. Cancer Epidemiol Biomarkers Prev 2014;23:953–966.
11. Schaafsma T, Wakefield J, Hanisch R, et al. Africa's Oesophageal Cancer Corridor: Geographic variations in incidence correlate with certain micronutrient deficiencies. PLoS One 2015;10:e0140107.
12. Loots E, Sartorius B, Madiba TE, et al. Is clinical research in oesophageal cancer in South Africa in crisis? A systematic review. World J Surg 2017;41:810–816.
13. Liu W, Snell JM, Jeck WR, Hoadley KA, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight 2017;2: e98457.
14. Brown J, Stepien AJ, Willem P. Landscape of copy number aberrations in esophageal squamous cell carcinoma from a high endemic region of South Africa. BMC Cancer 2020; 20:281.
15. Lin DC, Wang MR, Koeffler HP. Genomic and epigenomic aberrations in esophageal squamous cell carcinoma and implications for patients. Gastroenterology 2018;154:374–389.
16. Vaught J, Abayomi A, Peakman T, et al. Critical issues in international biobanking. Clin Chem 2014;60:1368–1374.
17. Abayomi A, Christoffels A, Grewal R, et al. Challenges of biobanking in South Africa to facilitate indigenous research in an environment burdened with human immunodeficiency virus, tuberculosis, and emerging noncommunicable diseases. Biopreserv Biobank 2013;11:347–354.
18. Soo CC, Mukomana F, Hazelhurst S, Ramsay M. Establishing an academic biobank in a resource-challenged environment. S Afr Med J 2017;107:486–492.
19. Census 2011. Provincial profile: KwaZulu-Natal. Available at www.statssa.gov.za (last accessed March 2, 2021).
20. Caldwell RI, Gaede B, Aldous C. Description of an internal medicine outreach consultant appointment in Western KwaZulu-Natal, South Africa, 2007 to mid-2014. S Afr Med J 2015;105:353–356.
21. Mendy M, Caboux E, Lawlor RT, et al. *Common Minimum Technical Standards and Protocols for Biobanks Dedicated to Cancer Research*. Lyon, FR: International Agency for Research on Cancer; 2017.
22. Matsubara T, Soh J, Morita M, et al. DV200 index for assessing RNA integrity in next-generation sequencing. Biomed Res Int 2020;2020:9349132.
23. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–185.
24. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–194.
25. Basik M, Aguilar-Mahecha A, Rousseau C, et al. Biopsies: Next-generation biospecimens for tailoring therapy. Nat Rev Clin Oncol 2013;10:437–450.
26. Diaz Z, Aguilar-Mahecha A, Paquet ER, et al. Next-generation biobanking of metastases to enable multidi-

mensional molecular profiling in personalized medicine. Mod Pathol 2013;26:1413–1424.
27. Florell SR, Coffin CM, Holden JA, et al. Preservation of RNA for functional genomic studies: A multidisciplinary tumor bank protocol. Mod Pathol 2001;14:116–128.
28. Aguilar-Mahecha A, Lafleur J, Pelmus M, et al. The identification of challenges in tissue collection for biomarker studies: The Q-CROC-03 neoadjuvant breast cancer translational trial experience. Mod Pathol 2017;30:1567–1576.
29. Ellis M, Davis N, Coop A, et al. Development and validation of a method for using breast core needle biopsies for gene expression microarray analyses. Clin Cancer Res 2002;8:1155–1166.
30. Chatterjee S. Artefacts in histopathology. J Oral Maxillofac Pathol 2014;18(Suppl 1):S111–S116.

Address correspondence to:
Christopher G. Mathew, PhD, FRCPath
Sydney Brenner Institute for Molecular Bioscience
Faculty of Health Sciences
University of the Witwatersrand
9 Jubilee Road
Johannesburg 2193
South Africa

E-mail: christopher.mathew@kcl.ac.uk

Appendix D – Plagiarism Declaration



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Mishalan Moodley (Student number: 0108860t) am a student registered for the degree of PhD (Molecular Medicine) in the academic year 2019.

I hereby declare the following:

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

Signature: _____ Date: 2023

1

Appendix E – Turn-It-In Report

Research and Postgraduate Office
Faculty of Health Sciences
University of the Witwatersrand
Phillip V Tobias Health Sciences Building
29 Princess of Wales Terrace, Parktown
Johannesburg, South Africa, 2193

December 2023

RE: MMOODLEY 01018860T: Turn-It-In report clarification for PhD Submission

Dear Colleagues,

This letter is in support of Mr M Moodley's PhD submission. The Turn-It-In Originality report for his thesis returned a similarity index of 30%. This is due to the inclusion of his original published manuscripts as part of the thesis submission:

1. Ferndale, L., **Moodley, M**., Chen, W.C., et al.2021b. Processing and analysis of tissue samples from esophageal cancer patients in an african setting. Biopreserv Biobank. doi:10.1089/bio.2021.0030

Please refer to the Originality report attached. The top two matches, accounting for 50% of the 30% of the similarity index, are all from his original work.

Yours sincerely,

Dr Pascale Willem, MD and PhD

Medical Geneticist in Human Genetics and Oncology

Wits Diagnostic Innovation Hub, University of the Witwatersrand, South Africa

Email: pwillem@witsdih.ac.za

Mishalan_Moodley_Thesis_PhD.pdf

ORIGINALITY REPORT

30%	26%	27%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Lucien Ferndale, Mishalan Moodley, Wenlong C. Chen, Reubina Wadee et al. "Processing and Analysis of Tissue Samples from Esophageal Cancer Patients in an African Setting", Biopreservation and Biobanking, 2021 Publication	9%
2	www.lancet.co.za Internet Source	6%
3	hdl.handle.net Internet Source	1%
4	www.frontiersin.org Internet Source	<1%
5	Submitted to University of Witwatersrand Student Paper	<1%
6	res.mdpi.com Internet Source	<1%
7	hal.archives-ouvertes.fr Internet Source	<1%

8	link.springer.com Internet Source	<1%
9	"Handbook of Nutrition, Diet, and Epigenetics", Springer Science and Business Media LLC, 2019 Publication	<1%
10	www.medrxiv.org Internet Source	<1%
11	www.wits.ac.za Internet Source	<1%
12	www.nature.com Internet Source	<1%
13	"Encyclopedia of Signaling Molecules", Springer Science and Business Media LLC, 2018 Publication	<1%
14	Wenjin Liu, Jeff M. Snell, William R. Jeck, Katherine A. Hoadley et al. "Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis", JCI Insight, 2016 Publication	<1%
15	www.ncbi.nlm.nih.gov Internet Source	<1%
16	Yao, Lijing. "Identification and Characterization of Cancer-Associated	<1%

	Enhancers.", University of Southern California, 2019 Publication	
17	Fernando Calvo, Romana Ranftl, Steven Hooper, AaronJ. Farrugia et al. "Cdc42EP3/BORG2 and Septin Network Enables Mechano-transduction and the Emergence of Cancer-Associated Fibroblasts", Cell Reports, 2015 Publication	<1 %
18	www.researchgate.net Internet Source	<1 %
19	www.biorxiv.org Internet Source	<1 %
20	academic.oup.com Internet Source	<1 %
21	www.scribd.com Internet Source	<1 %
22	ebin.pub Internet Source	<1 %
23	Cristina A. Martinez, Ina Marteinsdottir, Ann Josefsson, Gunilla Sydsjö, Elvar Theodorsson, Heriberto Rodriguez-Martinez. "Epigenetic modifications appear in the human placenta following anxiety and depression during pregnancy", Placenta, 2023 Publication	<1 %

24	research-repository.griffith.edu.au Internet Source	<1%
25	orca.cf.ac.uk Internet Source	<1%
26	aacrjournals.org Internet Source	<1%
27	edepot.wur.nl Internet Source	<1%
28	vdoc.pub Internet Source	<1%
29	www.mdpi.com Internet Source	<1%
30	Veronika N. Laine, Bernice Sepers, Melanie Lindner, Fleur Gawehns, Suvi Ruuskanen, Kees Oers. "An ecologist's guide for studying DNA methylation variation in wild vertebrates", Molecular Ecology Resources, 2022 Publication	<1%
31	d-nb.info Internet Source	<1%
32	api.research-repository.uwa.edu.au Internet Source	<1%
33	www.dovepress.com Internet Source	<1%

34	"Melanoma", Springer Science and Business Media LLC, 2019 Publication	<1 %
35	Manoj K. Kashyap, Arivusudar Marimuthu, Charles Jacob Harrys Kishore, Suraj Peri et al. "Genomewide mRNA profiling of esophageal squamous cell carcinoma for identification of cancer biomarkers", Cancer Biology & Therapy, 2014 Publication	<1 %
36	discovery.ucl.ac.uk Internet Source	<1 %
37	pubmed.ncbi.nlm.nih.gov Internet Source	<1 %
38	www.lecb.ncifcrf.gov Internet Source	<1 %
39	Md. Abdul Aziz, Mohammad Sarowar Uddin, Md. Shalahuddin Millat, Mohammad Safiqul Islam. "Vascular endothelial growth factor A (VEGFA) promoter rs2010963 polymorphism and cancer risk: An updated meta-analysis and trial sequential analysis", Meta Gene, 2022 Publication	<1 %
40	Yi, Lynn. "Statistical Methods for Gene Differential Expression Analysis of RNA-	<1 %

	Sequencing.", California Institute of Technology, 2023 Publication	
41	Submitted to St George's Hospital Medical School Student Paper	<1%
42	spiral.imperial.ac.uk Internet Source	<1%
43	wiredspace.wits.ac.za Internet Source	<1%
44	www.illumina.com Internet Source	<1%
45	"Breast Cancer: From Bench to Personalized Medicine", Springer Science and Business Media LLC, 2022 Publication	<1%
46	wlv.openrepository.com Internet Source	<1%
47	Hua, Jun Jie. "Functional Analysis of Risk Single Nucleotide Polymorphism and Long Noncoding RNA in Prostate Cancer.", University of Toronto (Canada), 2019 Publication	<1%
48	Nan Lin, Jinpeng Liu, James Castle, Jun Wan, Aditi Shendre, Yunlong Liu, Chi Wang, Chunyan He. "Genome-wide DNA methylation	<1%

profiling in human breast tissue by illumina TruSeq methyl capture EPIC sequencing and infinium methylationEPIC beadchip microarray", Epigenetics, 2020
Publication

49	docksci.com Internet Source	<1%
50	medsci.org Internet Source	<1%
51	www.researchsquare.com Internet Source	<1%
52	Pickett-Leonard, Michael D.. "Identification of Novel Genes and Compounds for the Development of Precision Therapeutics for Dystrophic Epidermolysis Bullosa and Associated Cutaneous Squamous Cell Carcinoma", University of Minnesota, 2021 Publication	<1%
53	oa.las.ac.cn Internet Source	<1%
54	"Molecular Pathology of Neoplastic Gastrointestinal Diseases", Springer Science and Business Media LLC, 2013 Publication	<1%
55	Harsh Pawar, Manoj Kumar Kashyap, Nandini A. Sahasrabuddhe, Santosh Renuse et al. "Quantitative tissue proteomics of	<1%

esophageal squamous cell carcinoma for novel biomarker discovery", Cancer Biology & Therapy, 2014
Publication

56	etheses.bham.ac.uk Internet Source	<1%
57	Submitted to Anglia Ruskin University Student Paper	<1%
58	Arfa Mehmood, Asta Laiho, Laura L. Elo. "Exon-level estimates improve the detection of differentially expressed genes in RNA-seq studies", RNA Biology, 2021 Publication	<1%
59	Methods in Molecular Biology, 2015. Publication	<1%
60	Submitted to Universiti Sains Malaysia Student Paper	<1%
61	Wenlong Carl Chen, Jean-Tristan Brandenburg, Ananyo Choudhury, Mahtaab Hayat et al. "Genome-wide association study of esophageal squamous cell cancer identifies shared and distinct risk variants in African and Chinese populations", The American Journal of Human Genetics, 2023 Publication	<1%
62	erc.bioscientifica.com Internet Source	<1%

63	repository.upenn.edu Internet Source	<1%
64	Katarzyna Wreczycka, Alexander Gosdschan, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, Altuna Akalin. "Strategies for analyzing bisulfite sequencing data", Journal of Biotechnology, 2017 Publication	<1%
65	Schwaner, Caroline. "Costs and Mechanisms Associated with Resilience to Acidification in Marine Bivalves", State University of New York at Stony Brook, 2023 Publication	<1%
66	cshperspectives.cshlp.org Internet Source	<1%
67	eprints.bournemouth.ac.uk Internet Source	<1%
68	Kravitz, Stephanie Nicole. "Random Allelic Expression in the Human Body: Implications for Phenotypic Variation, Adaptive Traits, and Disease Variability", The University of Utah, 2023 Publication	<1%
69	Oliver, James. "A Personal 'Omics' Approach to Understand the Molecular Mechanism of Response to Biologic Therapy in Rheumatoid	<1%

	Arthritis", The University of Manchester (United Kingdom), 2020 Publication	
70	Handbook of Disease Burdens and Quality of Life Measures, 2010. Publication	<1 %
71	Menezes, Prema. "Is there a trial effect in HIV clinical trials? Identifying who participates in clinical trials and assessing the effect of trial participation on the response to highly active antiretroviral therapy", Proquest, 20111108 Publication	<1 %
72	Thorpe, Lauren M., Haluk Yuzugullu, and Jean J. Zhao. "PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting", Nature Reviews Cancer, 2014. Publication	<1 %
73	Translational Bioinformatics, 2016. Publication	<1 %
74	apps.dtic.mil Internet Source	<1 %
75	tel.archives-ouvertes.fr Internet Source	<1 %
76	www.jcancer.org Internet Source	<1 %
77	www.science.gov Internet Source	<1 %

78	Submitted to Aspen University Student Paper	<1 %
79	Submitted to University of Edinburgh Student Paper	<1 %
80	Submitted to University of South Australia Student Paper	<1 %
81	harvest.usask.ca Internet Source	<1 %
82	jultika.oulu.fi Internet Source	<1 %
83	kclpure.kcl.ac.uk Internet Source	<1 %
84	ndl.ethernet.edu.et Internet Source	<1 %
85	nova.newcastle.edu.au Internet Source	<1 %
86	raw.githubusercontent.com Internet Source	<1 %
87	repository.up.ac.za Internet Source	<1 %
88	unsworks.unsw.edu.au Internet Source	<1 %
89	Submitted to Aston University Student Paper	<1 %

90	Qingguo Wang, Joshua Armenia, Chao Zhang, Alexander V. Penson et al. "Unifying cancer and normal RNA sequencing data from different sources", Scientific Data, 2018 Publication	<1 %
91	Submitted to University of Sussex Student Paper	<1 %
92	Wu, Liang, James G. Herman, Malcolm V. Brock, Kongming Wu, Gaoping Mao, Wenji Yan, Yan Nie, Hao Liang, Qimin Zhan, Wen Li, and Mingzhou Guo. "Silencing DACH1 Promotes Esophageal Cancer Growth by Inhibiting TGF-β Signaling", PLoS ONE, 2014. Publication	<1 %
93	bioinfo.cd-genomics.com Internet Source	<1 %
94	jhoonline.biomedcentral.com Internet Source	<1 %
95	researchspace.ukzn.ac.za Internet Source	<1 %
96	ueaeprints.uea.ac.uk Internet Source	<1 %
97	www.coherentmarketinsights.com Internet Source	<1 %
98	Phillips, Chelsea M.. "The 20 Kilodalton Isoform of Connexin 43: A Multifaceted	<1 %

	Contributor to Cerebral Cavernous Malformation Type 3 Pathogenesis", University of Michigan, 2023 Publication	
99	Submitted to University of Exeter Student Paper	<1 %
100	publications.polymtl.ca Internet Source	<1 %
101	qmro.qmul.ac.uk Internet Source	<1 %
102	Wenbin Guo, Nikoleta Tzioutziou, Gordon Stephen, Iain Milne, Cristiane Calixto, Robbie Waugh, John WS Brown, Runxuan Zhang. "3D RNA-seq - a powerful and flexible tool for rapid and accurate differential expression and alternative splicing analysis of RNA-seq data for biologists", Cold Spring Harbor Laboratory, 2020 Publication	<1 %
103	neuro.unboundmedicine.com Internet Source	<1 %
104	www.theses.fr Internet Source	<1 %
105	"Biomechanics in Oncology", Springer Science and Business Media LLC, 2018 Publication	<1 %

106	"Cancer Genetics and Psychotherapy", Springer Science and Business Media LLC, 2017 Publication	<1 %
107	9pdf.net Internet Source	<1 %
108	Camille Ravel-Godreuil, Olivia Massiani-Beaudoin, Philippe Mailly, Alain Prochiantz, Rajiv L. Joshi, Julia Fuchs. "Perturbed DNA methylation by Gadd45b induces chromatin disorganization, DNA strand breaks and dopaminergic neuron death", iScience, 2021 Publication	<1 %
109	Fleischer, Thomas, Arnoldo Frigessi, Kevin C Johnson, Hege Edvardsen, Nizar Touleimat, Jovana Klajic, Margit Riis, Vilde D Haakensen, Fredrik Wärnberg, Bjørn Naume, Åslaug Helland, Anne-Lise Børresen-Dale, Jörg Tost, Brock C Christensen, and Vessela N Kristensen. "Genome-wide DNA methylation profiles in progression to", Genome Biology, 2014. Publication	<1 %
110	Jill Alldredge, Leslie Randall. "Germline and Somatic Tumor Testing in Gynecologic Cancer Care", Obstetrics and Gynecology Clinics of North America, 2019 Publication	<1 %

111	Kocsis, Tori. "Labeling Melanoma Cells with Black Microspheres for Improved Sensitivity in Detection Via Photoacoustic Flow Cytometry", Duquesne University, 2023 Publication	<1 %
112	Submitted to The University of Manchester Student Paper	<1 %
113	Zvonko Magic, Gordana Supic, Mirjana Brankovic-Magic, Nebojsa Jovic. "Chapter 8 DNA Methylation in the Pathogenesis of Head and Neck Cancer", IntechOpen, 2012 Publication	<1 %
114	www.research.manchester.ac.uk Internet Source	<1 %
115	www.spandidos-publications.com Internet Source	<1 %
116	Arnaud Stigliani, Renata Ialchina, Jiayi Yao, Dominika Czaplinska et al. "Adaptation to an acid microenvironment promotes pancreatic cancer organoid growth and drug resistance in a p53-dependent manner", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
117	Submitted to CUNY, John Jay College of Criminal Justice Student Paper	<1 %

118	Sameer Ullah Khan, Kaneez Fatima, Fayaz Malik, Halime Kalkavan, Abubakar Wani. "Cancer metastasis: Molecular mechanisms and clinical perspectives", Pharmacology & Therapeutics, 2023 Publication	<1 %
119	Thadeous J. Kacmarczyk, Mame P. Fall, Xihui Zhang, Yuan Xin, Yushan Li, Alicia Alonso, Doron Betel. ""Same difference": comprehensive evaluation of four DNA methylation measurement platforms", Epigenetics & Chromatin, 2018 Publication	<1 %
120	Submitted to University of Sheffield Student Paper	<1 %
121	expeditiorepositorio.utadeo.edu.co Internet Source	<1 %
122	nvlu.repo.nii.ac.jp Internet Source	<1 %
123	open.uct.ac.za Internet Source	<1 %
124	"DNA Methyltransferases - Role and Function", Springer Science and Business Media LLC, 2022 Publication	<1 %

125	"Handbook of Statistical Genomics", Wiley, 2019 Publication	<1 %
126	"The Extracellular Matrix", Springer Science and Business Media LLC, 2019 Publication	<1 %
127	Lim, Jongpyo Joe. "Acute and Persistent Effects from Environmental Toxicant Exposure on the Gut-Liver Axis.", University of Washington, 2020 Publication	<1 %
128	Po-Kuei Hsu, Hua-Ling Kao, Hsuan-Yu Chen, Chueh-Chuan Yen, Yu-Chung Wu, Wen-Hu Hsu, Teh-Ying Chou. "Loss of CRNN expression is associated with advanced tumor stage and poor survival in patients with esophageal squamous cell carcinoma", The Journal of Thoracic and Cardiovascular Surgery, 2014 Publication	<1 %
129	Ying Zhang, Camila Bruna de Lima, Rémi Labrecque, Marc-André Sirard. "Whole-genome DNA methylation analysis of the sperm in relation to bull fertility", Reproduction, 2023 Publication	<1 %

130	Zitnay, Rebecca Suzanne Goldstein. "Strain and Extracellular Tenascin-C in the Progression of Early Lung Adenocarcinoma to Invasive Disease", The University of Utah, 2023 Publication	<1 %
131	clinicalepigeneticsjournal.biomedcentral.com Internet Source	<1 %
132	www.tandfonline.com Internet Source	<1 %
133	"Compendium of Inflammatory Diseases", Springer Science and Business Media LLC, 2016 Publication	<1 %
134	Esophageal Squamous Cell Carcinoma, 2015. Publication	<1 %
135	Fuqiang Dai, Longyong Mei, Shenglan Meng, Zheng Ma, Wei Guo, Jinghai Zhou, Jingge Zhang. "The global expression profiling in esophageal squamous cell carcinoma", Genomics, 2017 Publication	<1 %
136	J. Lotem, D. Netanely, E. Domany, L. Sachs. "Human cancers overexpress genes that are specific to a variety of normal human tissues", Proceedings of the National Academy of Sciences, 2005	<1 %

Publication

137	K. L. Kong, D. L. Kwong, L. Fu, T. H. M. Chan, L. Chen, H. Liu, Y. Li, Y.-H. Zhu, J. Bi, Y.-R. Qin, S. Y. K. Law, X.-Y. Guan. "Characterization of a Candidate Tumor Suppressor Gene Uroplakin 1A in Esophageal Squamous Cell Carcinoma", Cancer Research, 2010 Publication	<1 %
138	Liu, Y.. "Identification of novel retinal target genes of thyroid hormone in the human WERI cells by expression microarray analysis", Vision Research, 200708 Publication	<1 %
139	Submitted to UC, San Diego Student Paper	<1 %
140	Virendra Singh, Laishram Chandreshwor Singh, Madavan Vasudevan, Indranil Chattopadhyay et al. "Esophageal Cancer Epigenomics and Integrome Analysis of Genome-Wide Methylation and Expression in High Risk Northeast Indian Population", OMICS: A Journal of Integrative Biology, 2015 Publication	<1 %
141	Xianzhen Peng, Hengchuan Xue, Lingshuang Lü, Peiyi Shi, Jianping Wang, Jianming Wang. "Accumulated promoter methylation as a potential biomarker for esophageal cancer", Oncotarget, 2016	<1 %

	Publication	
142	Yahui Qiao, Binjie Liu, Ruiqi Bai, Jingwen Cai, Qian Peng. "White Sponge Nevus Caused by Keratin 4 Gene Mutation: A Case Report", Genes, 2022 Publication	<1 %
143	Yasushi Sasaki, Miyuki Tamura, Ryota Koyama, Takafumi Nakagaki, Yasushi Adachi, Takashi Tokino. "Genomic characterization of esophageal squamous cell carcinoma: Insights from next-generation sequencing", World Journal of Gastroenterology, 2016 Publication	<1 %
144	cris.maastrichtuniversity.nl Internet Source	<1 %
145	open.library.ubc.ca Internet Source	<1 %
146	rcastoragev2.blob.core.windows.net Internet Source	<1 %
147	"Engineering Translational Models of Lung Homeostasis and Disease", Springer Science and Business Media LLC, 2023 Publication	<1 %
148	"Metabolism and Epigenetic Regulation: Implications in Cancer", Springer Science and Business Media LLC, 2022 Publication	<1 %

149	Advances in Experimental Medicine and Biology, 2015. Publication	<1 %
150	Basavaraj Vastrad, Chanabasayya Vastrad. "Screening and identification of potential prognostic biomarkers in bladder urothelial carcinoma: Evidence from bioinformatics analysis", Gene Reports, 2020 Publication	<1 %
151	Fan Guo, Lin Li, Jingyun Li, Xinglong Wu, Boqiang Hu, Ping Zhu, Lu Wen, Fuchou Tang. "Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells", Cell Research, 2017 Publication	<1 %
152	Kai Fu Jhuang, Man Lun Hsu, Yu-Chia Chen, Jan-Growth Chang, Moncef Zouali. " methylation trajectories during innate and adaptive immune responses of human B lymphocytes ", Immunology, 2023 Publication	<1 %
153	Kervarrec, Thibault. "Histogenesis of Merkel Cell Carcinoma.", Bayerische Julius-Maximilians-Universitaet Wuerzburg (Germany), 2020 Publication	<1 %
154	Meiqi Wang, Dan Liu, Yunchuanxiang Huang, Ziyi Jiang, Feng Wu, Yu Cen, Lan Ma.	<1 %

"Identification of Key Genes Related to the Prognosis of Esophageal Squamous Cell Carcinoma Based on Chip Re-Annotation", Applied Sciences, 2021
Publication

155	Perron, Gabrielle. "Deciphering the Regulatory Programs that Govern mRNA Stability in Cancer.", McGill University (Canada), 2023 Publication	<1 %
156	Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélène Denis, Christos Sotiriou, François Fuks. "Evaluation of the Infinium Methylation 450K technology", Epigenomics, 2011 Publication	<1 %
157	Submitted to University of Surrey Student Paper	<1 %
158	Venkataraman, Yaamini R.. "Ocean Acidification Influences on Physiology and Epigenetics in the Pacific Oyster (Crassostrea gigas)", University of Washington, 2021 Publication	<1 %
159	Yang, Zhao-Ting, So-Young Yeo, Yong-Xue Yin, Zhen-Hua Lin, Hak-Min Lee, Yan-Hua Xuan, Yan Cui, and Seok-Hyung Kim. "Tenascin-C, a Prognostic Determinant of Esophageal Squamous Cell Carcinoma", PLoS ONE, 2016.	<1 %

	Publication	
160	Yue Yu, Zhihua Li, Chenjun Huang, Haisheng Fang et al. "Integrated analysis of genomic and transcriptomic profiles identified a prognostic immunohistochemistry panel for esophageal squamous cell cancer", Cancer Medicine, 2019 Publication	<1 %
161	Zheng, Jianyong, Chunsheng Xu, Dake Chu, Xiaoying Zhang, Jipeng Li, Gang Ji, Liu Hong, Quanxin Feng, Xiaohua Li, Guosheng Wu, Jianjun Du, and Qingchuan Zhao. "Human leukocyte antigen G is associated with esophageal squamous cell carcinoma progression and poor prognosis", Immunology Letters, 2014. Publication	<1 %
162	atrium.lib.uoguelph.ca Internet Source	<1 %
163	biblio.ugent.be Internet Source	<1 %
164	coek.info Internet Source	<1 %
165	dipot.ulb.ac.be Internet Source	<1 %
166	ir.uiowa.edu Internet Source	<1 %

167	jcancer.org Internet Source	<1 %
168	mediatum.ub.tum.de Internet Source	<1 %
169	opinvisindi.is Internet Source	<1 %
170	os.unil.cloud.switch.ch Internet Source	<1 %
171	Alameri, Alia. "Microfluidic-Based Organoid Vascularization in PDMS Devices Replicated from 3D Printed Molds", McGill University (Canada), 2023 Publication	<1 %
172	Alsiwiehri, Naif O.. "Identification and Characterisation of Novel Cancer Testis Antigens in Human Cancer Cells.", Bangor University (United Kingdom), 2020 Publication	<1 %
173	Andrea Mathe, Michelle Wong-Brown, Warwick J. Locke, Clare Stirzaker et al. "DNA methylation profile of triple negative breast cancer-specific genes comparing lymph node positive patients to lymph node negative patients", Scientific Reports, 2016 Publication	<1 %

174	Brown, Jacqueline. "Genomic imbalances in esophageal carcinoma cell lines involve Wnt pathway genes", World Journal of Gastroenterology, 2011. Publication	<1 %
175	Caiyu Guo, Fanye Zeng, Hui Liu, Jianlin Wang, Xue Huang, Judong Luo. "Establish immune-related gene prognostic index for esophageal cancer", Frontiers in Genetics, 2022 Publication	<1 %
176	Chao Tan, Chenyu Shi, Yin Li, Wen Teng, Yongjing Li, Huiru Fu, Liting Ren, Hong Yu, Qi Li, Shikai Liu. "Comparative Methylome Analysis Reveals Epigenetic Signatures Associated with Growth and Shell Color in the Pacific Oyster, Crassostrea gigas", Marine Biotechnology, 2022 Publication	<1 %
177	Diego E. Andrade-Brito, Diana L. Núñez-Ríos, José Jaime Martínez-Magaña, Sheila T. Nagamatsu et al. "Neuronal-specific methylome and hydroxymethylome analysis reveal replicated and novel loci associated with alcohol use disorder", Cold Spring Harbor Laboratory, 2023 Publication	<1 %
178	Dou, Wei, Guang-Mao Shen, Jin-Zhi Niu, Tian-Bo Ding, Dan-Dan Wei, and Jin-Jun Wang.	<1 %

"Mining Genes Involved in Insecticide Resistance of Liposcelis bostrychophila Badonnel by Transcriptome and Expression Profile Analysis", PLoS ONE, 2013.
Publication

---

179 Jianzhou Cui, Karishma Sachaphibulkij, Wen Shiun Teo, Hong Meng Lim et al. " Annexin-A1 deficiency attenuates stress-induced tumor growth fatty acid metabolism in mice: an Integrated multiple omics analysis on the stress- microbiome-metabolite-epigenetic-oncology (SMMEO) axis ", Theranostics, 2022 <1 %
Publication

---

180 Li Wang, Wan Jiang, Wuhua Zhang, Weitong Chen, Yu Mo, Tong Zhou, Jinzhu Zhang, Daidi Che. " Molecular evidence to the day length in regulating the short shoot phenomenon during summer in roses (sp) ", Journal of Plant Interactions, 2022 <1 %
Publication

---

181 Lianghua Bin. "Inhibition of Transcription Factor Specificity Protein 1 Alters the Gene Expression Profile of Keratinocytes Leading to Upregulation of Kallikrein-Related Peptidases and Thymic Stromal Lymphopoietin", Journal of Investigative Dermatology, 11/2011 <1 %
Publication

182	Natasha Tilikj, Alejandro Martínez Navarro, Marta Novo. "Terrestrial Zen master: Transcriptomics in long-term aestivation and arousal of Mediterranean earthworms", Journal of Experimental Zoology Part A: Ecological and Integrative Physiology, 2023 Publication	<1 %
183	Pryhoda, Moira Kate. "A Longitudinal Examination of Biomechanical Balance and Quantitative Multidomain Assessments During Recovery Following Sport-Related Concussion.", University of Denver, 2020 Publication	<1 %
184	Rau, A., M. Gallopin, G. Celeux, and F. Jaffrezic. "Data-based filtering for replicated high-throughput transcriptome sequencing experiments", Bioinformatics, 2013. Publication	<1 %
185	Sebastian Moran, Carles Arribas, Manel Esteller. "Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences", Epigenomics, 2016 Publication	<1 %
186	Shavannor M Smith, Yinan Yuan, Andrew N Doust, Jeffrey L Bennetzen. " Haplotype Analysis and Linkage Disequilibrium at Five Loci in ", G3 Genes|Genomes|Genetics, 2012	<1 %

Publication

187	Sheila C.S. Lima. "Identification of a DNA methylome signature of esophageal squamous cell carcinoma and potential epigenetic biomarkers", Epigenetics, 10/01/2011 Publication	<1 %
188	Weghorst, Forrest. "Mechanisms of Non-Apoptotic Roles for Apoptotic Proteins in Development of the Chick Auditory Brainstem", University of California, Irvine, 2023 Publication	<1 %
189	Wu, Jieyu. "Targeting Stromal Components in the Tumor Microenvironment for Cancer Therapy", Karolinska Institutet (Sweden), 2023 Publication	<1 %
190	Yikun Geng, Min Wang, Zhouying Wu, Jianchao Jia, Tingyu Yang, Lan Yu. "Research progress of circRNA in malignant tumour metabolic reprogramming", RNA Biology, 2023 Publication	<1 %
191	Yuki Okawa, Shota Sasagawa, Hiroaki Kato, Todd A. Johnson et al. "Immuno-genomic analysis reveals eosinophilic feature and favorable prognosis of female non-smoking	<1 %

esophageal squamous cell carcinomas", Cancer Letters, 2023
Publication

192	Zhongqiu Wang, Binhui Ren, Jianfeng Huang, Rong Yin, Feng Jiang, Qin Zhang. "LncRNA DUXAP10 modulates cell proliferation in esophageal squamous cell carcinoma through epigenetically silencing p21", Cancer Biology & Therapy, 2018 Publication	<1 %
193	cancerci.biomedcentral.com Internet Source	<1 %
194	crimsonpublishers.com Internet Source	<1 %
195	docplayer.net Internet Source	<1 %
196	dspace.mit.edu Internet Source	<1 %
197	epdf.pub Internet Source	<1 %
198	eprints.nottingham.ac.uk Internet Source	<1 %
199	eprints.soton.ac.uk Internet Source	<1 %
200	export.arxiv.org Internet Source	<1 %

201	hal-agrosup-dijon.archives-ouvertes.fr Internet Source	<1 %
202	healthjade.net Internet Source	<1 %
203	issuu.com Internet Source	<1 %
204	mobt3ath.com Internet Source	<1 %
205	ourarchive.otago.ac.nz Internet Source	<1 %
206	peerj.com Internet Source	<1 %
207	pubs.rsc.org Internet Source	<1 %
208	scholar.sun.ac.za Internet Source	<1 %
209	www.autismparentingmagazine.com Internet Source	<1 %
210	www.fedoa.unina.it Internet Source	<1 %
211	www.hal.inserm.fr Internet Source	<1 %
212	www.hindawi.com Internet Source	<1 %

213	www.scirp.org Internet Source	<1 %
214	www.whba1990.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 8 words
Exclude bibliography	On		