ADAM9 Isoform Switching in Oesophageal Cancer

By

Noor Us Subah

Fall-2018-MSBI-3-00000273477

Supervised by:

Dr Mehak Rafiq

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE in Bioinformatics

Research Centre for Modelling and Simulation (RCMS)
National University of Sciences & Technology (NUST)
September, 2021

Dedication

This dissertation is wholeheartedly dedicated to my beloved family, especially to my parents Noordad and Nisbat Zahra for their sacrifices, trust, constant support and endless love.

Certificate of Originality

I Noor Us Subah, hereby declare that the results presented in this research work titled "ADAM9 Isoform switching in Oesophageal Cancer" are generated by myself. Moreover, none of its contents is plagiarised nor set forth for any kind of evaluation or higher education purposes. I have acknowledged/referenced all the literary content used for support in this research work.

_____

Name Noor Us Subah

Fall-2018-MSBI-3-00000273477

September, 2021

Acknowledgement

First and foremost, I would like to thank Almighty Allah, the most compassionate and the most merciful, the author of knowledge and wisdom, for his countless blessing. who has bestowed upon me the power and ability to think and grow, empowering me to play my role in conveying a little share of my knowledge.

I am extremely grateful and wish to convey my sincere gratitude to my respected supervisor Dr. Mehak Rafiq, Assistant professor, RCMS, NUST, for her continuous supervision throughout my MS degree and research work. Her trust in me kept me focused and determined. I would like to thank her for her constant assistance, for the invaluable guidance she provided me with and the tremendous effort to offer every possible help to finish this thesis.

I would like to present cordial gratitude to my respected GEC members, Dr Rehan Zafar Paracha, Dr Maria Shabbir and Engr. Fawad Khan for their honest advises, sincere guidance. Who has always been available for their humble assistance at various stages of my study and provided me with their valuable feedback and opinion.

I present my deepest gratitude to my parents Noordad and Nisbat Zahra, for their endless love, guidance, trust, moral and financial support throughout my education. I would also like to thank my sisters, Kainaat Noor, Kinza Noor and Masooma Maryam for their love, encouragement, guidance and support.

I am grateful to my dear friend Mehar Masood for her constant encouragement, continuous support and assistance in my studies, trusting my abilities and turning me into a positive and strong individual. I greatly acknowledge my friends and colleagues Aqsa Qureshi, Sameen Fatima, Tayyaba Alvi and for their love, support, guidance and encouragement.In the end I want to pay my deepest gratitude to all the faculty members, lab engineers and other staff members at RCMS, NUST, for having a positive and supportive role in this beautiful journey.

Table of Contents

List of Abbreviations

mRNA	Messenger RNA
aTTS	alternative Transcription Termination Sites
aTSS	alternative Transcription Start Sites
DGE	Differential Gene Expression
DTE	Differential Transcript Expression
DTU	Differential Transcript Usage
ALK	Anaplastic lymphoma kinase
MCL	Induced Myeloid Leukemia cell differentiation protein
HISAT2	Hierarchical Indexing for Spliced Alignment of Transcripts
ORF	Open Reading Frame
AS	Alternative Splicing
EC	Oesophageal Carcinoma
AC	Adenomcarcinoma
SCC	Squamous Cell Carcinoma
GERD	Gastroesophageal reflux disease
TNM	Tumour, nodes, and metastases
ADAM	A Disintegrin and Metalloproteinase
ADAMTS	A Disintegrin and Metalloproteinase with Thrombospondin motifs
TACE	Tumour necrosis factor-α-converting enzyme
MDC	Human macrophage-derived chemokine
EGFR	Epidermal Growth Factor Receptor
EGF	Epidermal Growth Factor
TNF-alpha	Tumour Necrosis Factor-alpha
HB-EGF	Heparin Binding-Epidermal growth factor
ECM	Extracellular matrix
VCAM	Vascular cell Adhesion protein
CD40	Cluster of differentiation 40

EMT	Epithelial to Mesenchymal Transition
CDCP1	CUB-domain containing protein 1
VEGFA	Vascular endothelial growth factor-A
siRNA	Small Inhibitory RNA
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
TCGA	The Cancer Genome Atlas
EMBL-EBI	European Molecular Biology Laboratory's European Bioinformatics Institute
SAM	Sequence Alignment Map
BAM	Binary Alignment Map
KEGG	Kyoto Encyclopedia of Genes and Genomes
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
PPI	Protein-protein interactions
FPKM	Fragment per kilo million
FC	Fold Change
IF	Isoform Fraction
dIF	Differential Isoform Fraction

List of Figures

List of Tables

Abstract

Alternative splicing (AS) generates various structurally and functionally different protein isoforms. AS plays an important role in cancers by triggering hallmarks of cancer from a progression of primary tumour cells (tumorigenesis) to metastasis of secondary tumour cells to distant organs. Oesophageal cancer (EC) is one of the deadliest and least studied cancers worldwide because of its aggressive nature and low mortality rate. It remains a public health concern worldwide (Holmes and Vaughan, 2007). ADAM9 is a membrane-anchored protein that is involved in various physiological and regulatory functions. Proteolytically, ADAM9 is involved in EGFR signalling by processing EGFR ligands HB-EGF) whereas non-proteolytically interacts with integrins and is involved in cell adhesion and cancer invasion. Expression levels of the two alternatively spliced transcripts of ADAM9 have an opposing role in breast cancer. This study was designed to provide a clear understanding if switching exists between L and S forms and ADAM9 enrichment in oesophageal cancer. Transcriptome assembly and reconstruction analysis revealed that ADAM9 is significantly upregulated, L transcript has high coverage than S transcript in oesophageal cancer. Dexseq statistical analysis revealed the differential transcript usage of L and S -form, which shows the increased usage of L-form compared to S-form.

Furthermore, interacting partners for ADAM9 were identified. Moreover, enrichment analysis was performed, which revealed focal adhesion as the enriched process in all three datasets. This shows that ADAM9 is involved in ECM interactions occurring at specialized zones called focal adhesions. These focal adhesions are rich in integrin adhesion receptors which play an essential role in bi-directional transmembrane communication by connecting cell cytoskeletons to the extracellular membrane matrix in response to these focal adhesion signalling, the cell initiates diverse processes, including cell growth or death, cell motility and cytoskeleton reorganisation. Thus, this study enhanced the understanding of the proteolytic and non-proteolytic roles of ADAM9 and its isoforms in oesophageal cancer; however, a specific pathway in which ADAM9 is involved still needs to be discovered. Moreover, the stage-specific role of S and L-form is yet to be studied using stage-specific data.

Chapter 1
Introduction

1.1 Background

The central dogma of molecular biology illustrates the flow of genetic information within the cell. Initially, specific genes on DNA are transcribed to messenger RNA (mRNA) transcript, which excluded the intronic region from the gene and joined the exonic regions, further translated to a single functional protein. Soon, this one gene-one protein theory was challenged in the mid-1970s by some researchers. They evaluated that a single gene might produce more than one functional protein via an alternative splicing mechanism of mRNA transcript Figure ***1-1***.

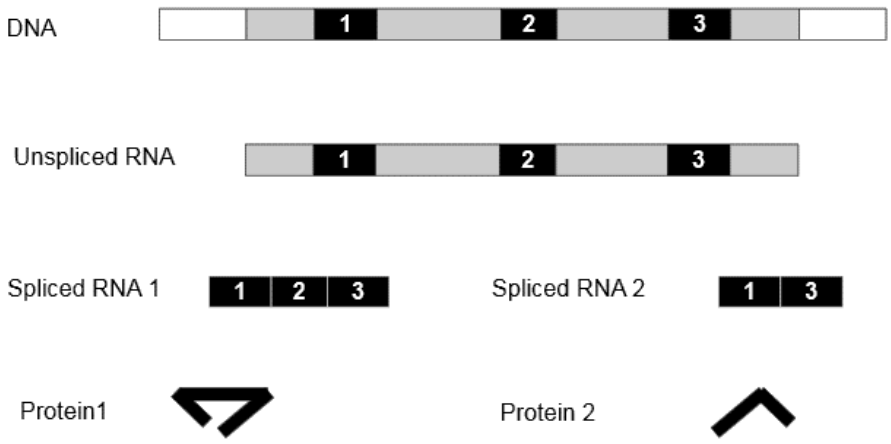

Figure 1-1:Alternative splicing mechanism: after removal of non-coding introns, rearrangement of exons into alternatively spliced isoforms that code for different functional proteins

Since then, the number of known isoforms/alternatively spliced transcripts has increased drastically. It is now reported that the majority of multi-exon genes show alternative splicing mechanisms, including approximately 95% of genes in humans having multiple isoforms/transcripts [1].

1.2 Alternative Splicing

The tendency of multi-exon genes to produce different alternatively spliced isoforms from the same gene using: different alternative transcription termination sites (aTTS), alternative transcription start site (aTSS), polyadenylation sites, and alternative promoter usage serves as a primary determining factor for increased proteome complexity in higher vertebrates. Around 70% of genes have multiple polyadenylation sites, whereas more than 50% of the gene has aTSS [2]. According to the Encyclopedia of DNA Elements ENCODE project statistics, on average, each gene encodes approximately 6.3 isoforms and 3.9 different functional protein isoforms [3]. This protein diversity arises from the usage of different splicing cites during alternative splicing, almost all events of alternative splicing result from the use of four primary events or modules that are [4] shown in Figure ***1-2***:

a) Alternative 5′ splice-site choice
b) Alternative 3′ splice-site choice
c) Cassette-exon inclusion or skipping
d) Intron retention

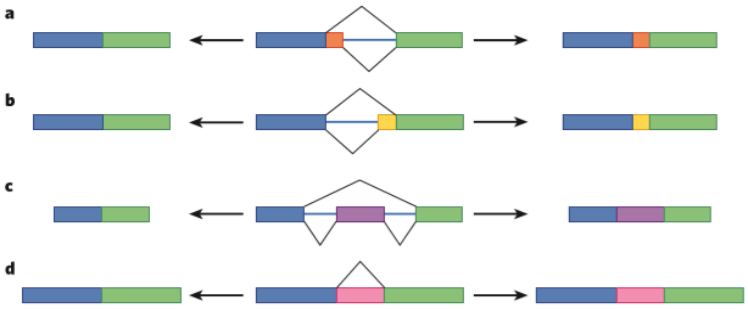

Figure 1-2: Four basic modules of alternative splicing (a) alternative 5′ splice site selection, (b) alternative 3′ splice-site selection, (c) cassette-exon inclusion or skipping and (d) intron retention. (Nilsen & Graveley, 2010).

These protein isoforms generated from single-parent genes regulate similar functions in the body and play a significant role in genetic diversity [5]. The majority of isoforms regulate similar functions in closely related metabolic pathways [6]. However, in some aberrant conditions, the functions of two isoforms from the same protein can have opposing effects on a cellular process. Isoforms may differ in structure, function, localisation, and other properties [7]. Considering that most multi-exon genes are responsible for different

functional proteins, assessing the isoform expression and gene-level expression is essential. Previously many gene expression studies have been carried out using Microarray and Ribonucleic acid sequencing (RNA-seq) technologies and received much attention but did not evaluate the importance of these technologies at the isoform level [8].

1.2.1 Differential Gene Expression and Differential Transcript Expression

Initially, in many high throughputs sequencing studies, the primary focus was to check differential gene expression (DGE). However, in eukaryotes, various transcripts/isoforms arise from each gene due to the alternative splicing mechanisms. Since gene expression comprises a collective sum of all transcripts (union of exons), differential transcript expression (DTE) analysis is preferred over DGE. Therefore, to check the gene and isoform expression, a considerable number of methods are developed, which are broadly divided into two categories a) Union-exon based approach, also known as DGE b) transcript-based approach, also known as DTE.

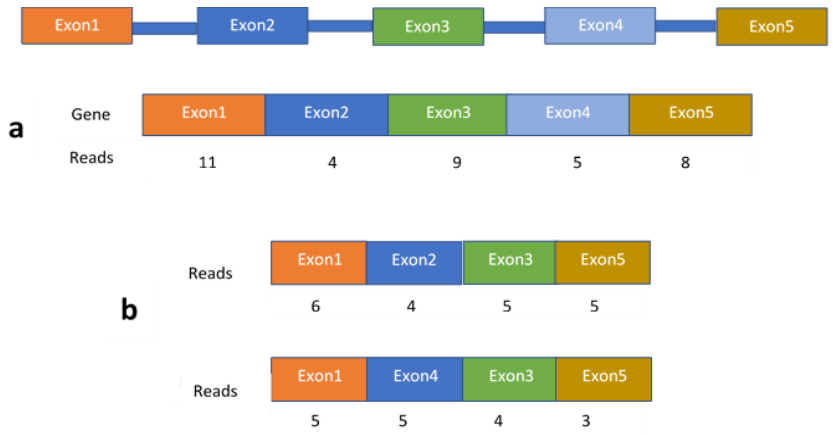

Figure 1-3: a) Union-exon based approach merges all overlapping exons from transcripts b) In DTE reads are assigned based on higher confidence to the gene.

Union exon-based approach merges all the overlapping exons from the gene into union exons; this gives a differential gene expression between two conditions shown in Figure ***1-3(a)***. Whereas in transcript quantification absolute abundance of each transcript is calculated irrespective of the gene, reads are assigned to gene-based on higher confidence shown in Figure ***1-3(b)***. However, the union exon-based approach was simple but failed to distinguish isoforms from the same gene. Furthermore, as most genes are expressed in more

than one transcript, the transcript-based approach gives more biologically meaningful information than the union-exon-based approach [9].

Unfortunately, transcript-level analysis is complex and expensive. To achieve highly significant results, higher sequencing depth (generating more reads that will increase statistical power to detect genes with the lowest expression) is required as gene expression is split among other isoforms. Moreover, high genomic similarity in isoforms complicates the assignment of reads among them. Despite these challenges, several studies have shown that isoforms have distinct functions and that shifts in individual isoform expression represent an actual level of gene regulation [10]. It is possible to identify differential transcript expression (DTE) even when there is little to no significant change in gene expression, introducing a new concept of differential transcript usage (DTU), which checks the relative abundance of isoforms. Different alternative splicing tools are available to differentiate between DGE, DTE and DTU.

1.2.2 Differential Transcript Expression and Differential Transcript Usage

Differential transcript expression (DTE) calculates the absolute expression of individual transcripts irrespective of the gene of origin. It is possible to identify differential transcript expression DTE even when there is little to no significant change in gene expression, introducing a new concept of differential transcript usage (DTU) and calculates the individual abundance of each isoform relative to the gene, where the dominance completely shifts from one isoform to another also known as isoform switching. In contrast, these important events like minor isoform expression change and isoform switches are disguised at the gene level, as it fails to distinguish isoforms [10]. Figure ***1-4*** illustrates the DTE and DTU among two conditions for genes with two isoforms. DTE states that the expression of at least one isoform changes among two conditions. However, the change in the proportion of transcripts expression remains the same, so this DTE does not infer DTU [11]. On the other hand, in DTU, there is a relative change in the expression of isoforms between two conditions while gene expression may or may not change, as the expression of one isoform is changed, so it also implies the DTE.

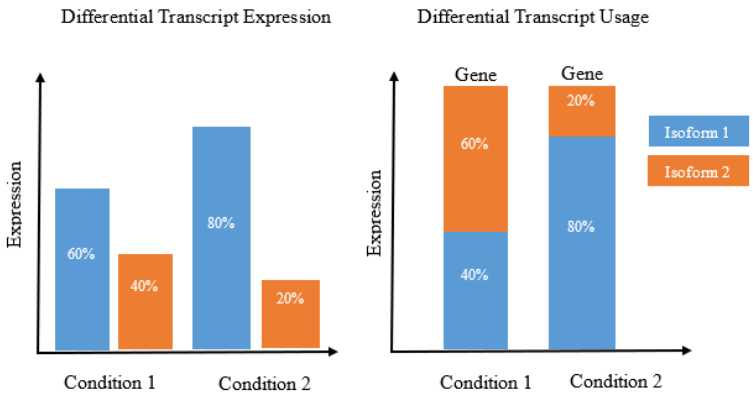

Figure 1-4:(a) Differential Transcript Expression (DTE): Isoform expression changes between two conditions irrespective of a gene (b) Differential Transcript Usage (DTU): shows the change in abundance of isoforms compared to other isoforms of the same gene

1.3 Importance of Isoform Switching

Differential transcript usage plays a significant role in regulating various biological processes, including development, homeostasis, pluripotency, and apoptosis. Furthermore, transcript isoforms are mostly tissue specific that might change the function, cellular localisation and stability of mRNA or protein. The change in the differential usage of isoforms DTU is often referred to as isoform switching Figure ***1-5***. It can have a significant biological impact due to the difference in functional potential of both isoforms. These Isoform switches are involved in different diseases and are particularly prominent in cancer [12].

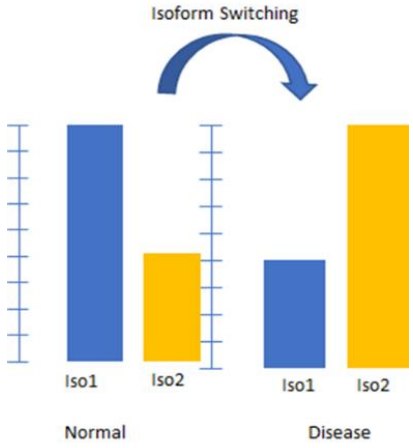

Figure 1-5: Relative abundance of isoform1 and isoform2 of the same gene is reversed when compared to normal (Complete shift of dominance of isoform1 in disease condition to the other alternatively spliced isoform2)

1.3.1 Examples of Switched Genes

One of the prominent examples of isoform switch from literature is the Anaplastic lymphoma kinase ALK gene. Due to the differential usage of aTSS, this switch produces a truncated protein without the extracellular domain, which causes aberrant cell proliferation and drives tumorigenesis in vitro [12]. Another example is the Myeloid Cell Leukaemia 1 (Mcl-1) gene responsible for induced myeloid leukaemia cell differentiation have two distinct alternatively spliced transcript isoforms. Longer transcript isoform Mcl-1L inhibits apoptosis to promote cell survival. In contrast, the shorter transcript isoform MCL-1S regulates apoptosis [13]. Many other studies have reported genes with identified switches that are primarily involved in all eight cancer hallmarks. Thus, targeting both specific splicing and general events, including splicing catalysis, splicing regulatory proteins are essential for improved therapeutic purposes [14]. In addition, various RNA-seq studies have been carried out to identify alterations in alternative splicing events involved in several diseases [15].

1.4 Tools for DTU

Tools for the identification of DTU are broadly classified into three major groups shown in Figure ***1-6***

Isoform based/Assembly Based:

1) Assembly Based methods reconstruct and quantify the absolute expression of transcripts. Old tuxedo suite (TopHat [16], Cufflinks, cuffdiff [17]) and new tuxedo suite (Hisat2 [18], StringTie, Ballgown [19]) are widely used for DTE. For example, Cuffdiff quantifies the DTU by measuring the similarity between two probability distributions.

Event-Based Methods:

2) The second group primarily focuses on identifying Alternative Splicing events (like Alternative 5′ splice-site, Alternative 3′ splice-site, Cassette-exon inclusion or

skipping and Intron retention) and the number of reads showing the presence/absence of different splicing events (by rMATS [20], SpliceR [21]). rMATS is used as for statistical tool for the differential analysis of alternative splicing events.

Exon Based Methods:

3) The third group does not quantify transcript expression directly, instead uses differential exon usage to infer relative transcript abundance. The genome is typically divided into counting bins, and the number reads overlapping each bin is counted. These methods use generalised linear models to infer differential exon(bins) between different conditions. A widely used method is the DEXseq R package, but various alternatives are present like diffSplice [22].

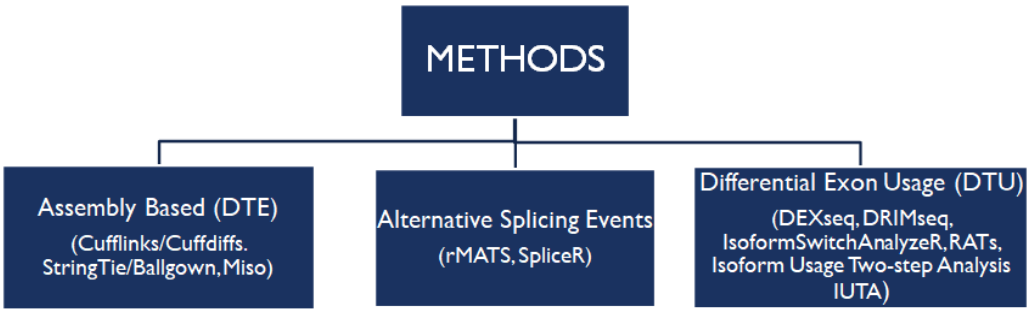

Figure 1-6: Methods for the identification of DTU: Assembly based. / Identification of DTU methods based on assembly. Alternative splicing events and differential exon usage.

Isoform Usage Two-step Analysis (IUTA) and IsoformSwitchAnalyzeR are the integrated pipelines for the differential transcript usage analysis. IUTA is implemented in the R package, designed to test each gene in the genome for differential isoform usage between two groups. IUTA also estimated isoform usage for each gene in each sample and averaged across samples within each group. IUTA tested the differential usage based on Aitchison geometry and outperformed the cuffdiff2 to detect significant genes under the same FDR. However, these methods failed to control the type 1 error because the p-value justifies many samples but not the small number of replicates [23]. Another integrated pipeline IsoformSwitchAnalyzeR identifies isoform switches based on calculating differential isoform usage following the visualisation of identified isoform switches with predicted potential function consequences [24].

1.5 IsoformSwitchAnalyzeR Framework

Due to the recent advancements in Bioinformatics, it is now possible to reconstruct and quantify the whole transcriptome from RNA-seq data using cufflinks, StringTie and Kallisto tools. These full-length transcripts make it easy to detect DTU. Therefore, many tools have been developed to identify DTU one such framework designed is IsoformSwitchAnalyzeR, an R Bioconductor package used to identify and visualise isoform switches [12]. However, due to several problems, RNA-seq data is not fully used to its potential:

- Lack of tools for isoform switch identification
- Theirs no integrated framework to analyse results from different tools.
- Isoform switch visualisation

This framework overcomes all these problems by enabling the import of transcript quantification files into R.

Full length quantified transcripts from different tools (cufflink, StringTie, Kallisto) are analysed for isoform usage analysis using IsoformSwitchAnalyzeR. Different tools are integrated along with isoforms usage, such as open reading frames ORF for annotations, PFAM for protein domains, SignalP for peptide signals, IDR, intrinsically unstructured regions of proteins based on estimated energy content (IUPred) for intrinsically disordered regions and coding potential calculator (CPAT/CPC2) for coding potential. IsoformSwitchAnalyzeR identifies the isoforms switches and their annotation to predict the potential functional consequences of that switch, such as loss of protein domain or removing a signal peptide. IsoformSwitchAnalyzeR performs five high-level tasks that are:

- Statistical identification of isoform switches using the DEXseq tool
- Integration and identification of predicted annotations for isoforms involved in isoform switching
- Visualisation of identified isoforms switches for the gene of interest along with its predicted consequence

- Genome-wide pattern analysis in switch consequence and alternative splicing events

1.6 Alternative Splicing in Cancers

As mentioned earlier in this section, alternative splicing (AS) generates various structurally and functionally different protein isoforms. This pre-mRNA splicing plays a vital role in gene regulation as it controls cellular proliferation, differentiation, and cell survival processes. In contrast, aberrant alterations in these processes/pathways have been implicated in cancers [25]. Thus, AS plays an important role in cancers by triggering hallmarks of cancer from the progression of primary tumour cells(tumorigenesis) to metastasis of secondary tumour cells to distant organs [26]. Many transcriptome studies have illustrated the splicing profiles for both normal and tumour cells showing significant variation due to aberrant splicing, influenced indirectly by mutations in splicing factors, transcription factors and chromatin modifications. These cancer-related AS events ultimately affect protein domains and disrupt protein-protein interaction in cancer-related pathways [27].

1.6.1 Oesophageal Cancer

Oesophageal cancer (EC) is one of the deadliest and least studied cancers worldwide because of its aggressive nature and low mortality rate. Nevertheless, it remains a public health concern worldwide [28]. EC ranks sixth in mortality rate because of high fatality rate and eighth in most common cancer incidents globally, despite the advancements in the treatment [29]. Among all other malignancies' cancer, ESC shows unique epidemiological features, emphasising that multiple etiologies are responsible [30]. These dramatic changes vary across two histological types of EC based on the site of origin, Adenocarcinoma cancer (AC), Squamous cell carcinoma (SCC), furthermore over race, gender, and region [28]. ESC Incidence rates vary regionally by 16-fold. Countries in Southern and Eastern Africa and Eastern Asia show the highest rate, Western and Middle Africa and Central America shows the lowest rate in males and females [31]. The major risk factor in these regions are not well studied, but some of them are illustrated in Table ***1-1***:

Table II.1: Common risk factors of EC **(Layke and Lopez, 2006)**

Risk Factors	**Squamous Cell Carcinoma (SCC)**	**Adenocarcinoma (AC)**
Age	60-70 years	50-60 years
Condition	Achalasia, Lye Ingestion, Plummer-Vison syndrome, history of head-neck SCC, radiotherapy, excessive use of tobacco and smoking, Alcoholic and diet with high starch without fruit and vegetables	Barrett's oesophagus, gastroesophageal reflux disease (GERD), hiatal hernia
Race	Black	White
Gender	Male	Male

Squamous cell carcinoma and adenocarcinoma are epithelial tumours of the oesophagus responsible for more than 95% of oesophageal carcinoma. In contrast, non-epithelial tumours of the oesophagus (lymphomas, sarcomas and metastatic tumours) are rare [30].

1.6.1.1 Squamous Cell Carcinoma (SCC)

Squamous Cell Carcinoma (SCC) is one of the most common subtypes of oesophageal cancer in the regions outside the united states [32]. Typically found in the upper middle (2/3rd) section of the oesophagus associated with smoking and alcohol [33].

1.6.1.2 Adenocarcinoma (AC)

AC is the most predominant subtype of oesophageal cancer. Typically found in the lower(1/3rd) section of the oesophagus, Gastroesophageal reflux disease (GERD) and Barrett's oesophagus are associated. Untreated GERD leads to the Barrette's, where the squamous epithelium is replaced with columnar epithelium. The chronic backflow to bile and gastric acid causes great damage to the oesophagus and has been implicated in Barrette's metaplasia. Recent studies have shown that oesophageal metaplasia is one of the contributing factors of adenocarcinoma. The most extended the oesophagus region is affected higher the probability of adenocarcinoma [34].

Oesophageal cancer is the deadliest because of its aggressive behaviour. It may conquer regional, local and distant areas by different metastasis pathways, including lymphatic

spread and hematogenous spread of tumour cells [35]. The most common metastatic pattern for ES is lymph nodes, liver, lung, brain, bone and adrenal glands [31].

1.7 A Disintegrin metalloproteases ADAMs Overview

Recent studies have shown several well-defined processes and genes implicated in tumorigenesis; one such gene family that received comparatively less attention is the ADAM family of proteins involved in different tumorigenesis processes, including cancer initiation, progression and cancer-specific therapies [36]. A Disintegrin metalloproteases ADAMs (also known as MDCs: metalloproteinase/disintegrin/cysteine-rich) are multi-domain proteins comprising transmembrane and secreted proteins [37] that are primarily found in eukaryotes. ADAM proteins are primarily involved in proteolysis and adhesion processes enabling them to perform cell adhesion, migration, and ectodomain shedding of membrane proteins to trigger cell signalling processes [38].

1.7.1 ADAM Superfamily

ADAMs belongs to the metzincin superfamily of matrix metalloproteinase. Together with the ADAMs containing thrombospondins sequences (ADAMTS) and snake venom metalloproteinases, they represent the adamlysin subfamily shown in Figure ***1-7***. According to phylogenetic analysis, ADAMs are classified based on their catalytic action and site of expression. Thus, 20 gene members are reported for the ADAM family, of which half are proteolytically actives and are globally expressed [39].

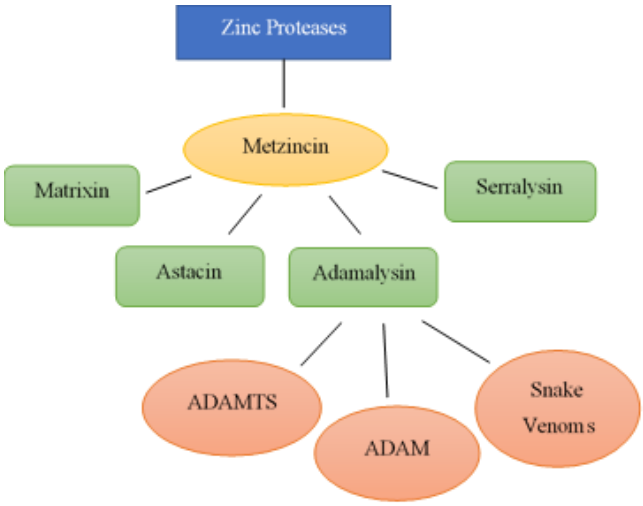

Figure 1-7: Phylogenetic classification of zinc protease superfamily

ADAMs protein comprises different combinations of the domains that are involved in the regulation variety of roles. For example, ADAMs behave as sheddases to shed the ectodomain of a membrane protein to initiate cell signalling processes using their metalloprotease domain. However, not all ADAMs have an active metalloprotease domain-containing HEXGHXXGXXHD motif (HEX motif), indicating their role in other cell adhesion and protein interaction processes [40]. These catalytically active ADAMs functions are proteolytic cleavage. The release of membrane attached factors is involved in different cell adhesion and proliferation processes illustrated in Table ***1-2*** [41]. In contrast, catalytically inactive ADAMs lacks the essential HEX motif on their metalloprotease domain for proteolysis and are involved in the cell adhesion process with the interaction with integrin proteins.

Table II.2: Studies elaborating the role of ADAMs in various cancers.

ADAM members	**Common Name**	**Potential Function**	**Localisation**
ADAM8	MS2, CD156	Adhesion, angiogenesis, inflammation	Plasma Membrane
ADAM9	Meltrin-gamma, MDC9	Adhesion, angiogenesis, sheddase, cell migration, proliferation	Plasma Membrane, Extracellular region.

			ER (Endoplasmic Reticulum)
ADAM10	Kuz, MADM, SUP-17	Adhesion, angiogenesis, sheddase, cell survival, inflammation, invasion, migration	Plasma membrane, Nucleoplasm
ADAM12	Meltrin-alpha	Angiogenesis, migration, proliferation, sheddase,	Plasma Membrane, Extracellular region
ADAM15	Metargidin, MDC15	Cell/cell binding	Trans-Golgi network, Plasma Membrane, Extracellular region
ADAM17	TACE	adhesion, angiogenesis, sheddase, cell survival, inflammation, invasion, migration	The plasma membrane, cytosol
ADAM19		Angiogenesis, adhesion, inflammation, invasion	
ADAM28	MDC-L	Immune surveillance, proliferation	The plasma membrane, mitochondria
ADAM33		Angiogenesis, genetically linked to asthma	Plasma membrane

1.8 Problem Statement

ADAMs family of transmembrane proteins has been implicated in shedding growth factors, cell migration and other processes. Among these ADAMs, several are expressed in multiple splice forms and perform different functions. One such member is ADAM9. ADAM9 has two experimentally validated transcript isoforms (S and L forms). Their gene expression is studied in different cancers; however, there is no study on the isoform expression level in oesophageal cancers. Therefore, this study explores the isoform switching of ADAM(S) and ADAM(L) in oesophageal cancer and its potential functional

consequences. Therefore, the aim is to identify ADAM9 isoform switching in primary oesophageal cancer using the IsoformSwitchAnalyzeR pipeline.

1.9 Aims and Objectives

The direct aims of this dissertation are to achieve the objectives listed below:

- Identify ADAM9 isoforms expression in oesophageal cancer via transcript reconstruction and quantification
- Investigate whether ADAM9 isoform switching exists between normal and tumour samples of primary oesophageal cancer by estimating the relative abundance of isoform usage
- Predict potential functional consequences of identified isoform switches
- Identify ADAM9 interacting partners, pathways, and their role in cancer through gene set enrichment analysis

Chapter 2

Literature Review

These ADAMs are multi-domain proteins comprising transmembrane and secreted proteins [37] primarily found in eukaryotes. ADAMs are type I transmembrane proteins containing seven domains with signal peptide sequence: pro-domain, metalloprotease domain, a disintegrin domain, a cysteine-rich domain, EGF-like motif, transmembrane and cytoplasmic domain shown in Figure ***2-1:***

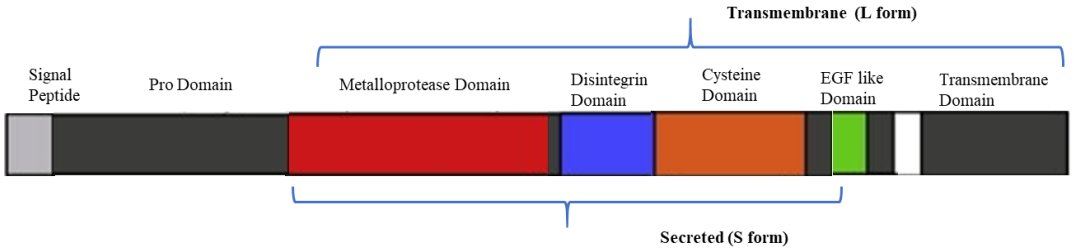

Figure 2-1: Multi-domain of ADAM gene

Since ADAMs transmembrane proteins, they are synthesised and transported via secretory pathways. After removing the signal peptide and properly folding the protein in the endoplasmic reticulum (ER), ADAMs are transported to the Golgi apparatus. The mature form is transported to cell surface Figure ***2-2***. Like many other proteases, ADAMs are synthesized inactive to attain proteolytic activity; they are processed to their mature form in the Golgi network by cleaving off the pro-domain, having an autoinhibitory effect over metalloprotease domain by furin convertase. The metalloprotease domain is responsible for the proteolytic activity of ADAMs; it can be active and inactive based on the presence of the HEXGHXXGXXHD motif [42]. The cysteine domain and disintegrin domain are involved in cell adhesion and protein-protein interaction processes. The role of the EGF-like domain is less clear, and the cytoplasmic domain has shown to be involved in signalling pathways due to the presence of an interaction motif [43].

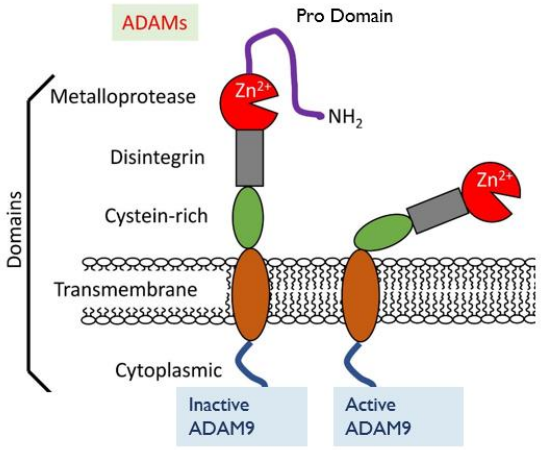

Figure 2-2: Conversion of inactive ADAM into the active mature form in Golgi network by cleaving off the pro-domain, having an autoinhibitory effect over metalloprotease domain by furin convertase.

2.1 Expression Patterns of ADAMs

In vertebrates, ADAMs are classified based on their catalytic activity and site of expression. A large number ADAMs are expressed in testis and hematopoietic cells. However, most of the ADAMs are expressed globally [36]. The classification of 20 ADAMs are shown in Figure ***2-3***. The expression pattern of each ADAM provides insights into their biological roles exclusively [40].

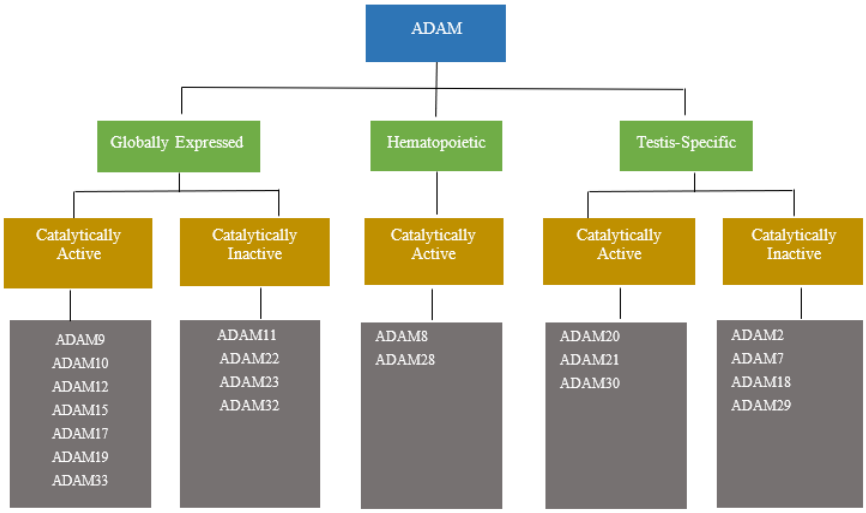

Figure 2-3: Classification of ADAMs based on their catalytic activity and site of expression.

Most of the catalytically active ADAMs are expressed globally among these two widely studied ADAMs are ADAM10 and ADAM17. Both are involved in development processes; the knock-out study of mice describes embryonic lethality [44]. These globally

expressed members include the ADAM9 gene, which relatively received less attention in multiple tumorigenesis processes.

Previously most of the studies for ADAMs were carried out to check their role in regions with high expression and gene deletions; however, scientists are curious to identify the role of these proteins in various diseases based on the transcriptomic data. Therefore, it is essential to consider that their activity might be related to the specific domain or the combination due to the multi-domain framework. Proteolytically active ADAMs are involved in ectodomain shedding, Notch, Tumour necrosis factor-alpha (TNF-α) and Epidermal growth factor receptor (EGFR) signalling, and several other proteolytic functions through the metalloprotease domain.

2.1.1 Shedding Membrane Proteins

One of the most studied proteolytic functions of ADAMs is their activity as sheddases [45]. As sheddases, they are involved in enzymatic cleavage of the extracellular portion of membrane proteins to shed soluble ectodomain shown in Figures ***2-4***. Further, these proteins activate and deactivate the signalling pathways [46]; these shed proteins can be growth factors, ligands, or membrane receptors.

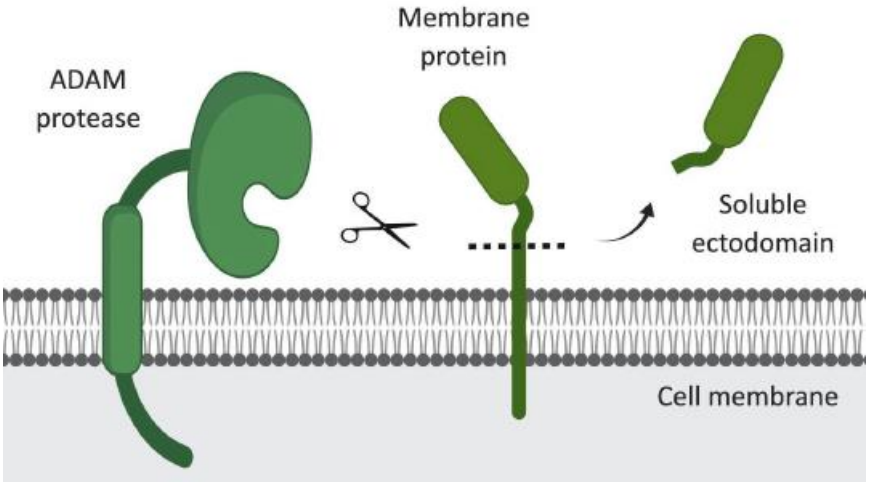

Figure 0-4: ADAMs functions as sheddase by cleaving the membrane proteins to shed ectodomains.

2.1.2 Epidermal Growth Factor Receptors (EGFR) Signalling

EGFR is a transmembrane protein that is a receptor to many EGF ligands. EGFR is involved in regulating different processes, including proliferation, differentiation, migration, and apoptosis. There are 11 EGFR ligands: epiregulin, amphiregulin, epigen, betacellulin, EGF, heparin-binding epidermal growth factor (HB-EGF), neuregulin, transforming growth factor-alpha (Kataoka, 2009). However, these ligands are membrane-tethered, and ADAMs helps cleave these EGFR pro-ligands to convert them into biologically active proteins during proliferation. Redundancy among ADAMs exists for cleaving the EGF pro-ligands, as the same EGF-ligand are activated by various ADAMs (ADAM9, ADAM10, ADAM12, ADAM15, ADAM17, ADAM19) shown in Table ***2-1*** [47].

Table 0.1: Membrane tethered pro-ligands are activated by various ADAMs.

Proteases	Transmembrane Substrates-Ligands
ADAM8	Pro TNF-α - TNF-α
ADAM9	Pro HB-EGF - EGF
ADAM10	Pro HB-EGF - EGF
ADAM12	Pro HB-EGF – EGF Pro epiregulin- epiregulin
ADAM15	Pro HB-EGF – EGF Pro epiregulin- epiregulin Pro amphiregulin- amphiregulin
ADAM17	Pro TNF-α - TNF-α Pro HB-EGF – EGF Pro epiregulin- epiregulin Pro amphiregulin- amphiregulin
ADAM19	Pro neuregulin - neuregulin

Besides their sheddase activity to release ectodomain of previously mentioned signalling processes, ADAMs are involved in various other proteolytic activities, including the shedding of cell-cell interacting and adhesion proteins (N cadherin, E cadherin, L selectins, Vascular endothelial (VE)-cadherin, Neural Cell adhesion molecule (N-CAM) [48]. These cleavage processes weaken the cell-cell linkages and have been inferred to play a crucial role in cancer metastasis [49]. Moreover, the shedding of E-cadherin and cell-cell interaction proteins enables the aberrant proliferation and extravasation of tumour cells, significant for cancer metastasis [49]. All the above-mentioned proteolytic processes might get deregulated under the aberrant expression of ADAMs that leads to cancer development.

Based on their ability to activate pro-ligands that instigate the proliferation and metastasis processes, it is noted that some ADAMs would be involved in malignancies. To date, ADAM17 and ADAM10 are well-studied members of the protease family, whereas ADAM9 has received relatively less attention despite being involved in multiple tumorigenesis processes.

2.1.3 Degradation of ECM Membrane

ECM is a specialised network that regulates various cellular functions. There is a continuous degradation or remodelling of ECM by different matrix proteases such as ADAMs in wound healing and tumorigenesis [50].

2.2 ADAM9

Initially, ADAM9 was known for myoblast fusion proteins, also known as Metalloprotease-Disintegrin-Cysteine domains (MDC9), identified from mouse lung complementary DNA *(cDNA*). Which revealed ADAM9 highly expressed canonical transmembrane form [51]. Many studies have identified their proteolytic role in processing EGFR ligands as they are known to activate pro-ligands. One cell-based research has shown that ADAM9 can shed EGF if both protease and ligands are expressed in the same cell other than EGFR ligands: HB-EGF, TNF-α, betacellulin, epiregulin or amphiregulin [52]. However, the cleaved form of soluble EGF by ADAM9 differs from the endogenous EGF form cleaved by other members due to the difference in cleavage site [53].

Furthermore, ADAM9 promotes the ectodomain shedding of other cell-cell interacting proteins: VE-cadherin, CD40, VCAM-1 and EphB4 when both protease and ligands are expressed in the same cells, leads to the possibility that ADAM9 might affect these molecules signalling when overexpressed *in vivo* [54]. In addition to shedding ectodomain, recent studies have shown that ADAM9 can also cleave other membrane proteases like; ADAM10 *in vitro,* resulting in the soluble proteinase and proteolysis transmembrane C terminal intramembrane [55]. Due to the domain-specific functionality, the ADAM9 disintegrin domain is involved in migration and invasion through interactions with

integrins. Mainly ADAM9 is involved in regulating cell adhesion to fibroblast cells by binding to integrins alpha 6/beta 1 (6ß1), increasing cell migration [56].

ADAM9 binding to integrins is also crucial to epithelial to mesenchymal transition (EMT), a continuous developmental process in our body regenerating new cells shown in Figure ***2-5***. However, in cancer, EMT is linked with tumour progression, metastasis, survival and stemness. Many extracellular proteases are involved in EMT among the metalloprotease family; ADAM9 regulates EMT by EGFR pathway Figure ***2-6***. Also, by IL6 via JNK-signalling, this EMT is blocked on the knock-out of ADAM9 in hepatocellular carcinoma [57]. Furthermore, ADAM9 alternatively spliced and secreted(short) form is involved in tumour invasion in carcinoma cell lines through binding with integrins alpha 6/beta 4 (6ß4) and alpha 2/beta 1 (2ß1).

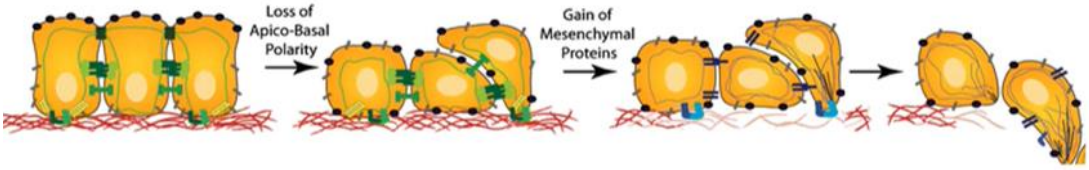

Figure 0-5: Loss of cells apical polarity during epithelial to mesenchymal transition

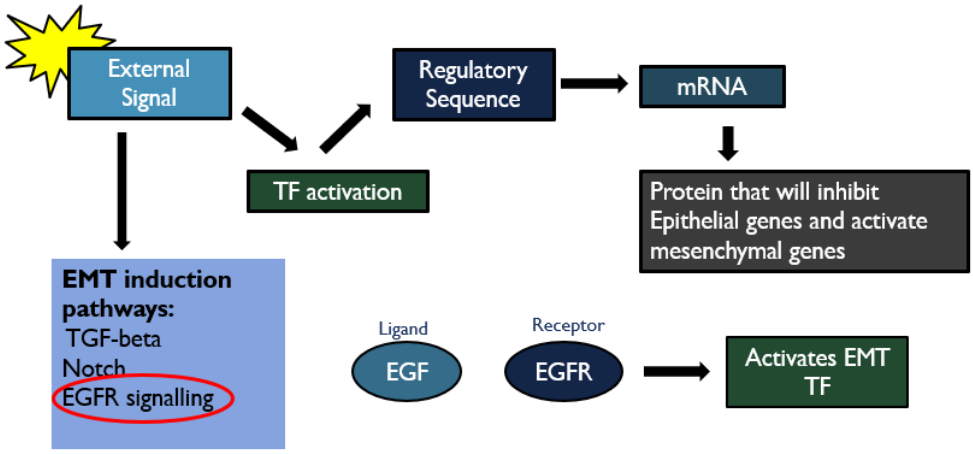

Figure 0-6: Onset of EMT after receiving the external signal activates the transcription factors to code mesenchymal genes and inhibit epithelial gens. EGFR is one of the EMT induction pathways that ADAM9 mediates.

2.2.1 ADAM9 In Cancers

ADAM9 is involved in regulating various cancer processes. In addition to invasion and metastasis, ADAM9 also plays an important role in cancer proliferation and angiogenesis. Table ***2.2*** shows the role of ADMA9 in various cancer along with the clinical outcomes and procedure.

Table 0.2: Role of ADAM9 gene in various cancers

Cancer Type	Role of ADAM9 and Outcome	Reference
Lung Cancer	**Clinical Outcomes:** overexpressed in cancer, negatively correlated with overall survival	[58][59][60][61]
	Procedure: 1. ADAM9 is involved in lung cancer metastasis by tPA mediated cleavage of CDCP1. 2. ADAM9 KO study - ADAM9 is a regulator of VEGFA and ANGPT involved in metastasis and angiogenesis.	[58][59][60][61]
Prostate Cancer	**Clinical Outcomes:** Overexpressed in cancer, negatively correlation with relapse-free survival	[62][63]
	Procedure 1. siRNA mediated KO of ADAM9 in PC3 cell line reduced the migration. ADAM9 mediates the Beta1 integrin degradation. 2. KO of Naal0p decreased the invasiveness- (Naa10p oncogene in prostate cancer form complex with ADAM9 and has a metastatic potential)	[62] [64]
Liver Cancer	**Clinical Outcomes:** Negatively correlated with immunotherapy feedback	[65] [66]
	Procedure: 1. siRNA-KO of ADAM9- to check ADAM9 mediated shedding of MICA (MHC class protein present in tumour cells) • Regorafenib and sorafenib drug inhibit the shedding of MICA by downregulating ADAM9.	[66] [67]

	2. KO of ADAM9 inhibits interleukin six mediated epithelial to mesenchymal transition ((EMT)	[65]
Breast Cancer	**Clinical Outcomes:** Overexpression in cancer, positively correlated with progression	[68]
	Procedure: NSD2 regulates ADAM9 and EGFR expression in triple-negative breast cancer (TNBC) – mediates invasion process	[69]
Pancreatic Cancer	**Clinical Outcomes:** Overexpression in cancer. Positively correlated with progression and negatively correlated with overall survival.	[70] [71] [72]
	Procedure KO of ADAM9 suppresses KRAS and MEK-ERK signalling Circ-ADAM9 reduces tumours growth in vivo	[70] [73]
Brain Cancer	**Clinical Outcomes:** Overexpression in cancer, negatively correlated with progression-free survival and overall survival	[74] [75]
	Procedure: TNBC **(**Tenascin C) treated glioblastoma cells	[75]
Oesophageal Cancer	**Clinical Outcomes:** Overexpression of ADAM9 in oesophageal adenocarcinoma	[76]
	Procedure: RT-PCR and western blotting were performed to check ADAM expression (9,10,12,17,19) in three oesophageal cell lines (OE19, Het1A, OE33).	[76]

Table 2-2: Role of ADAM9 gene in different cancers and their clinical outcomes and procedure used to achieve the study.

2.2.2 ADAM9 Isoforms in Cancers

Besides the classification into catalytic activity and expression site, alternative splicing produces even more transcript variants or isoforms of ADAMs with changed functionalities. According to GENCODE, 20,000 protein-coding genes generates approximately 144,000 transcripts. The current annotation estimates seven transcript isoforms per protein-coding gene; however, the annotation is far from complete. According to the University of California Santa Cruz (UCSC) genome Browser and ENSEMBL, catalytically active ADAM9 has two validated transcripts, transmembrane and secreted forms (ENST00000487273 and ENST00000379917) and five computational mapped transcripts, with several of them able to code for alternative forms of the proteins. For example, expression levels of the two alternatively spliced transcripts of ADAM9 have an opposing role in breast cancer where S-form is involved in cancer invasion. On the other hand, L-form suppresses invasion, illustrating the influence of different splice variants in cancer development [77].

Hatoda and his teams in 2012 were the pioneers to identify a soluble form of ADAM9(S-form) about its role as alpha-secretase in Amyloid precursor protein (APP) and its expression in several tissues [78]. Scientists have been curious about the role of alternatively spliced forms of ADMA9 in cancers since identifying the S-form. Mazzoca and his colleagues [79] explained the role of S-form in carcinoma invasion through tumour-stromal interaction in a knock-out study. According to this study, the S-form of ADAM9 secreted by special liver cells (hepatic stellate cells (HSC)) promotes tumour invasion using protease and disintegrin activities. This hypothesis was proved through Matrigel invasion assay and immunohistochemistry results, showed the binding of S-form with integrins ($\alpha6\beta4$ and $\alpha2\beta1$) on the tumour cells, and through its protease activity degraded the extracellular matrix that helped the invasion of tumour cells [79].

While on the contrary, Fry and Toker (2010) explained the opposing role of ADAM9 transcripts in breast cancer cell lines based on their metalloprotease domain. S-form is involved in the invasion of cancer cells requiring its metalloprotease domain. On the contrary, L-form inhibits invasion using the disintegrin domain, independent of its

proteolytic domain. L-form binds to the integrin protein via its disintegrin domain on the cell membrane, eventually altering the integrin-mediated signals and suppressing cell invasion and migration. In contrast, the S-form degrades the ECM via the metalloprotease domain and other substrates on the cell surface that are not cleaved by L-form due to its localization to the membrane. Thus, S-form is involved in cancer metastasis [77]. All these studies regarding the role of ADMA9 transcript recapitulate the previous research finding for other members ADAM11 and ADAM12, which have secreted and transmembrane forms and their role in cancers [80]. Another study demonstrated the role of ADAM9 in tumour invasion, where the knock-out of ADAM9 in MDA-MB-231 cells through *siRNA* inhibits the tumour cell invasion in-vitro in breast cancer [81]. This study shows the agreement with the Mazocca research group [79], which states that S-form promotes the colon carcinoma using both protease and disintegrin activities, whereas contradicts the Fry and Toker research group - which showed that S-form is involved in tumour invasion requiring metalloprotease domain. In contrast, the L-form suppresses the cell migration using the disintegrin domain [81]. Summarizing the literature findings suggests that ADAM9 is involved in invading tumour cells either by degrading ECM, activating other proteases, or binding with integrins and making ADAM9 the suitable target for clinical therapies against metastatic cancers.

Chapter 3

Methodology

This chapter describes the complete framework and steps used to achieve the study's aims and objectives mentioned in section 1.9. Broadly, the framework for this research is divided into three phases:

- Transcriptome Assembly and Reconstruction
- Isoform Switching
- Correlation Analysis

The generalized workflow for the first phase is illustrated in Figure ***3-1:***

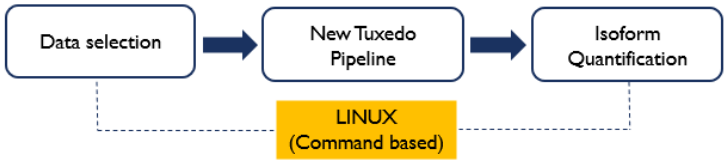

Figure 3-1: Major steps for Transcriptome Assembly and Reconstruction

3.1 Data Retrieval from GEO

The first research step was retrieving RNA-Seq datasets for oesophageal cancer from GEO (Gene Expression Omnibus) [82] and Array Express repositories [83]. Next, SRA (the Sequence Read Archive) files for RNA-Seq data were downloaded from the EMBL-EBI (European Bioinformatics Institute) [84]. FASTQ is the standard file format for sequenced data that contains both the sequenced reads and quality score of each base, known as PHRED (Public Health Research and Education Development) score and is encoded into ASCII (American Standard Code for Information Interchange) codes for human-readable form [85]. Specific criteria were considered while selecting RNA-Seq datasets to investigate the isoforms switching of the ADAM9 gene. For example, a dataset should be

extracted from *Homo Sapiens,* and it must include control and diseases samples, whereas cell lines should be avoided, as shown in Figure ***3-2***:

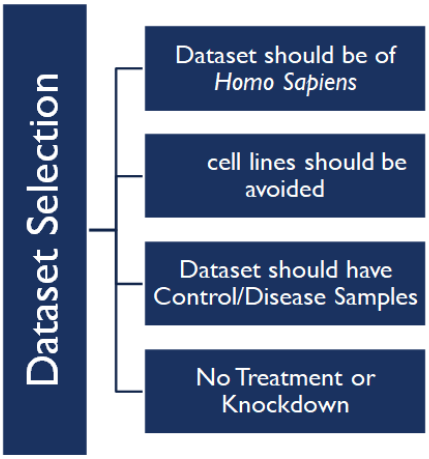

Figure 3-8: Criteria for selecting RNA-Seq datasets.

Based on the criteria mentioned above, three RNA-Seq datasets for oesophageal cancer were collected from GEO [82] and ArrayExpress [83] repository. The summary of datasets is presented in Table ***3-1:***

Table 0.1: Oesophageal cancer datasets and their accession numbers, sequencing platforms, and the number of samples from each dataset are shown.

Cancer	Accession no.	Platform	No. of Samples
Oesophageal Cancer	E-MTAB-4054	Illumina HiSeq 2000	27 (11 N, 16 T)
	GSE130078	Illumina HiSeq 2000	46 (23 N, 23 T)
	GSE111011	Illumina HiSeq 2500	14 (7 N, 7 T)
			Total= 87

Table 0.2*: Where N= Normal and T= Tumour samples.*

3.2 RNA-Seq Analysis

RNA-Seq experiment generates large amounts of complex raw data that needs accurate, fast and flexible software for comprehensive results. HISTA2 (Hierarchical indexing for spliced alignment of transcripts) [18], StringTie [86] and Ballgown [19] are freely available tools for RNA-Seq transcriptome analysis. Combinedly these tools make the new tuxedo

pipeline for RNA-Seq analysis. Steps involved in this pipeline are illustrated in Figure ***3-3***:

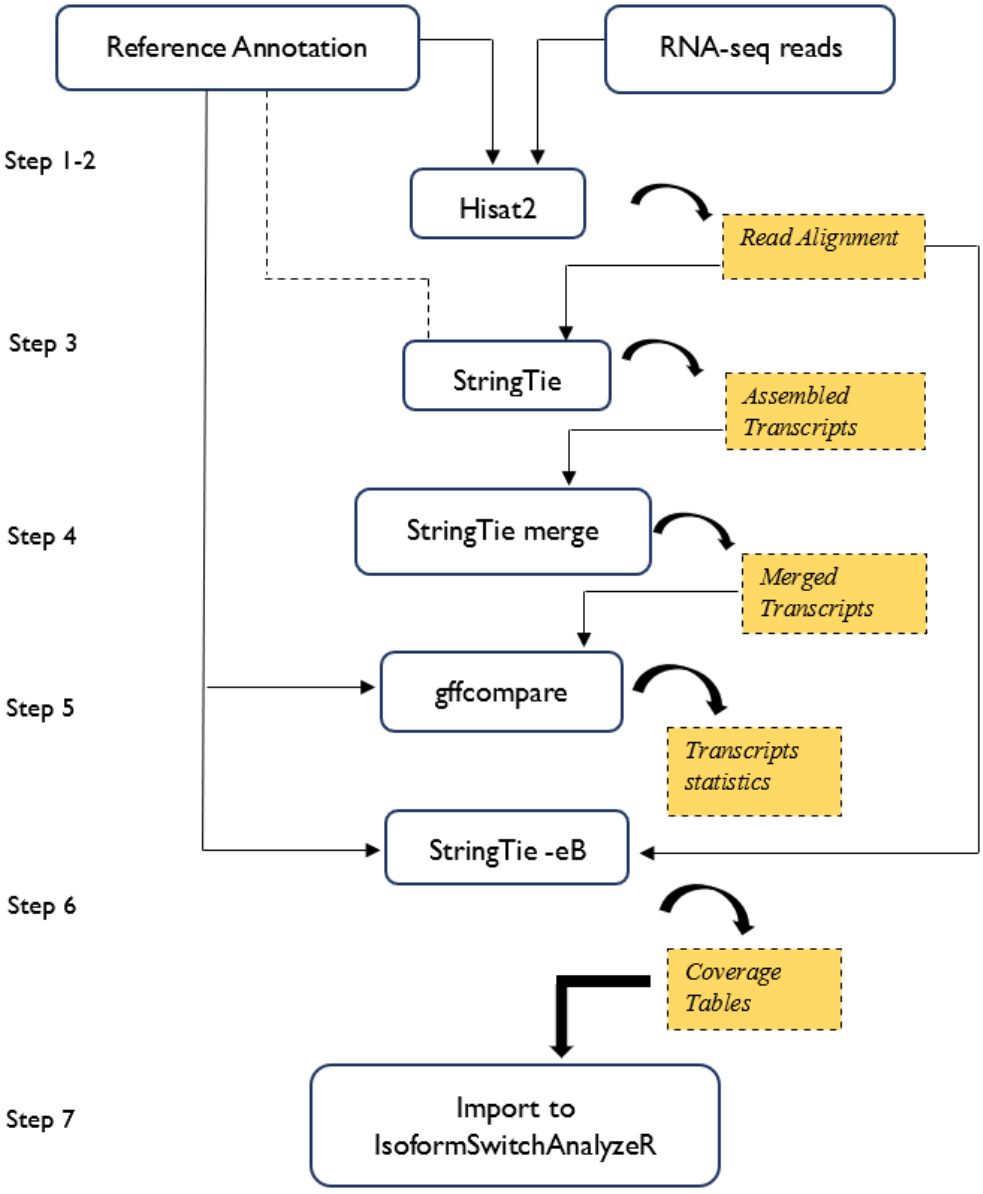

Figure 3-9: Detailed workflow for Transcriptome Assembly and Reconstruction using new Tuxedo pipeline.

3.2.1 Data Downloading

The first step after selecting data was to collectively download RNA-Seq samples from EMBL-EBI through FTP links using the Linux command:

```
wget -c -i file.txt
```

Where:

- wget= freely available non-interactive utility to download from the web
- -c = continue getting a partially downloaded file; This is mostly used for bulk download
- -i = input file parameter
- file.txt → file containing all the FTP links of samples

3.3 RNA-Seq Data Quality Control

The initial step before genomic data analysis was to evaluate the quality of raw FASTQ files. As the sequencers generate millions of reads in a single run, before processing those sequences to make biological conclusions, different quality control steps were performed to check the quality of raw data if it needed pre-processing. In addition, the FASTQC tool was used to identify technical biases in raw genomic data.

3.3.1 FASTQC

FASTQC is a powerful tool designed to identify the sequencing biases in high throughput sequencing data. It accepts the .fastq, .bam or .sam file formats as input to provide an overview of data quality. [87] Raw FASTQ files were uploaded to the FASTQC tool, which generated the quality check report to identify technical biases. Detail of quality checks is presented in Table ***3-2***. FASTQC output is the hypertext markup language (HTML) report that can be viewed on any web interface. The HTML report was generated using the Linux command where fastqc represents the tool name, and file.fastq.gz is raw sample files for paired-end:

```
fastqc file1.fastq.gz/file2.fastq.gz
```

Table 0.3: FASTQC quality check parameters

Quality Check Parameters	Explanation
Basic Statistics	It provides overall statistics of read length, GC content, the total number of reads, longest and shortest read length and filtered sequences.
Per Base Sequence Quality	It provides the quality values across all bases at each position in the FASTQ file
Per Sequence Quality Scores	It checks the quality of sequence by providing a quality score of each sequence, if it has low-quality bases in reads or not
Per Base Sequence Content	This plot displays the proportion of each base in a FASTQ file (for each of 4 DNA bases)
Per Sequence GC Content	It calculates the GC content of each sequence
Per Base N Content	This plot illustrates the % of base calls at each position with low confidence for which an N was called.
Sequence Length Distribution	Most of the sequences are of uniform length, but some are of varying length, so this plot depicts the distribution of sequence lengths.
Duplicate Sequences	It describes the level of duplication of each sequence. The level of duplication represents the coverage of sequence (high duplication indicates low coverage vice versa)
Overrepresented Sequences	It represents that either the sequence being reported is biologically significant or due to library contamination.
Adapter Content	This feature identifies the adapter content flanking on the end of sequences.

3.4 RNA-Seq Data Pre-processing

Pre-processing is one of the fundamental steps required in RNA-Seq analysis to remove biases from data. For example, raw reads may have adapter contamination along with low-quality bases and duplication. To resolve the sequencing biases generated in the FASTQC quality check report, pre-processing of fastq files was required before further analysis.

3.4.1 FASTP

FASTP (FASTQ Pre-processing) is a fast open-source tool used for pre-processing [88]. It can perform quality control (QC) and data filtering features, including adapter trimming and quality filtering of fastq files. This tool has the best features from previous pre-processing tools like Trimmomatic [89], FASTQC, After QC and Cutadapt [90] and some new features like UMI (Unique Molecular Identifier) and removal of Poly G tail [88]. Linux command used for FASTP is mentioned below. Adapter trimming was done by default but to disable, -A flag can be used:

```
fastp -i file1.fastq.gz -I file2.fastq.gz -o file1.fastq.gz
                        -O file2.fastq.gz
```

Where:

- -i= input file 1 (read on forward strand)
- -o= output file1 (read on forwards strand)
- -I= input File2 (read on reverse strand)
- -O= output file2 (read on reverse strand)
- file1 & file2= paired-end files (forward and reverse)

3.5 Alignment to Genome

After pre-processing, the next step was sequence alignment to the reference genome to detect genomic positions using hierarchical indexing for spliced alignment of transcripts (Hisat2) [91]. Reference genomes are available in different public databases. The human reference genome used in this study was downloaded from the Ensembl Genome Browser with Hg38.

3.5.1 HISAT2

Hisat2 is a splice aligner that aligns the splice junctions and the read alignments with referencing the genome. In terms of speed and efficiency, Hisat2 is much faster than its predecessors TopHat and TopHat2 [18].

Indexing is a significant step in alignment. Hisat2 builds both global (whole genome) and local (small chunks of a genome) indexing for alignment using the same BWT/FM (Burrows-Wheeler Transform/Ferragina Manzini) indexing as Bowtie2 and uses the chunk of indexing code of Bowtie[16]. Sequential steps for alignment are performed using Linux commands.

- Hisat2 alignment Linux command that was used is mentioned below:

```
hisat2 -p 8 --dta -x grch38_genome -1 file1.fastq.gz -2
                file2.fastq.gz -S file.sam
```

where:

- -p= provides multi-threads for processing, followed by the number of threads utilised for parallel processing.
- -dta= acronym for downstream transcriptome analysis.
- -x= directs towards the reference genome directory congaing indexing and the annotations file.
- -1/-2= indicator to input fastq files.
- -S= parameter to generate alignment output file which is of SAM format.

3.5.1.1 SAM to BAM File Conversion

The generic alignment output is Sequence Alignment Map (SAM) format, the text-based output that stores the read alignment results. For further analysis to make SAM files understandable and meaningful for computer programs, it is converted into binary (BAM format) using SAM tools. However, along with the conversion, the BAM file needs to be sorted because the alignments are generated randomly to their genomic positions to the reference genome; therefore, making alignments in genomic order sorting is necessary [92].

Both; SAM to BAM file conversion and sorting commands were executed jointly as well as separately using the following Linux Commands:

Joint Command

```
samtools sort -@ -8 -o file.bam file.sam
```

Individual commands

```
samtools view -S -b file.sam > file.bam

samtools sort file.bam -o file.sorted.bam
```

Where:

- view= to view sam or bam file
- -S= this Parameter specifies that input is in sam format
- -b= this flag indicates that output must be in bam format
- >= redirect operator to generate bam file
- sort= sort bam file to make alignments in genomic order

3.6 Transcriptome Assembly

One of the fundamental steps for transcriptome analysis after alignment is accurate assembly and reconstruction of all expressed isoforms and quantification of their relative abundances.

3.6.1 StringTie

In this study, StringTie was used for transcriptome assembly and reconstruction. Transcriptome data is in millions of short reads that have to be assembled. These short reads are used to map DNA transcripts; hence, it is a complex process requiring correct assembly and relative abundances. Compared to other assembly software like cufflinks, StringTie provides a complete and accurate reconstruction of genes and better estimates gene expressions [86].

- Assembly command using Linux is mentioned below:

```
stringtie -p 8 -G Homo_sapiens.GRCh38.84.gtf -l label -A
   file_ga.tab -C file_cr.gtf path/file.sorted.bam -o
                    file_at.gtf
```

Where:

- -p= provides multi-threads for processing, followed by the number of threads utilised for parallel processing.
- -G= reference genome
- -I= prefix for the output transcript name by default= "STRG"
- -A= generates the gene abundance file with the given name
- -C= generates the file with a given name including fully covered transcript reads present in reference; this Parameter was used with the -G flag (reference genome)
- -o= this Parameter outputs the assembled transcript file

3.6.1.1 Assembled Transcript Merging

After the initial assembly, merging assembled transcripts is crucial, as some of the transcripts are partially covered in some samples but entirely covered in others. Thus, to generate uniform sets of transcripts merging was done.

A text file containing assembled transcript files (file_at.gtf) from all samples was made using the StringTie command:

```
stringtie -G GRCH38.gtf --merge merged_list.txt -o
                    merged.gtf
```

Where:

- -G= reference genome annotation file in gtf format
- -merge= this parameter specifies the merge functionality
- merged_list.txt= file containing the path to the assembled transcript files generated in the previous step (file_at.gtf)
- -o= this parameter directs to the output file

- Merged.gtf= output file after merging

3.6.2 Gene/Transcript comparison to Reference Annotation

The merged gtf file generated in the previous step was compared with the reference annotation gtf file to check how accurately the transcript matched the reference annotation.

3.6.2.1 GFF Compare Utility

GFF Compare utility was downloaded, and environment variables were set. The command used for comparison is as follows:

```
gffcompare –r GRCh38.gtf –G Merged.gtf -o file_output
```

Where:

- -r= used to specify reference genome annotation
- -r= used to direct for comparison of merged.gtf file with reference genome gtf file
- -o= this flag is used to name output file, by default file name starts with "gffcmp", following the type of data in the file is (e.g., stats file contains precision and accuracy results for all the features).

3.7 Transcriptome Quantification

The ballgown tables were created using -B parameters in the StringTie command previously for assembled transcript quantification. Isoform and gene relative abundances were estimated in quantification steps. Command used to create ballgown quantification tables is as follows:

```
stringtie -p 8 -e -B -G Merged.gtf -o sample_quant.gtf
                    file.sorted.bam
```

Where:

- -p= provides multi-threads for processing, followed by the number of threads (8) utilised for parallel processing.
- -e= limit the processing to output/estimate only assembled transcript that matches the reference transcripts -G
- -G= reference annotation gtf file to guide the assembly.
- -o= generates the quantification file along with six other files that are mentioned below

This command generates six files with specific names that contain coverage data for all transcripts, namely:

- Exon data (e_data)
- Exon to Transcript data (e2t)
- Intron data (i_data)
- Intron to Transcript data (i2t)
- Quantification file (sample_quant)
- Transcript data (t_data)

Separate directories for each sample are made automatically to distinguish files for each sample, as the names of files are the same.

After the transcript quantification, the second phase of this study was to detect isoform switching in the ADAM9 gene; the IsoformSwitchAnalyzeR Bioconductor package was used. IsoformSwitchAnalyzeR is an easy and efficient R package for advanced post-analysis of transcript quantification [24]. The overall workflow of IsoformSwitchAnalyzeR is divided into two parts shown in Figure ***3-4,*** whereas sequential series of steps is illustrated in Figure *3-**5***:

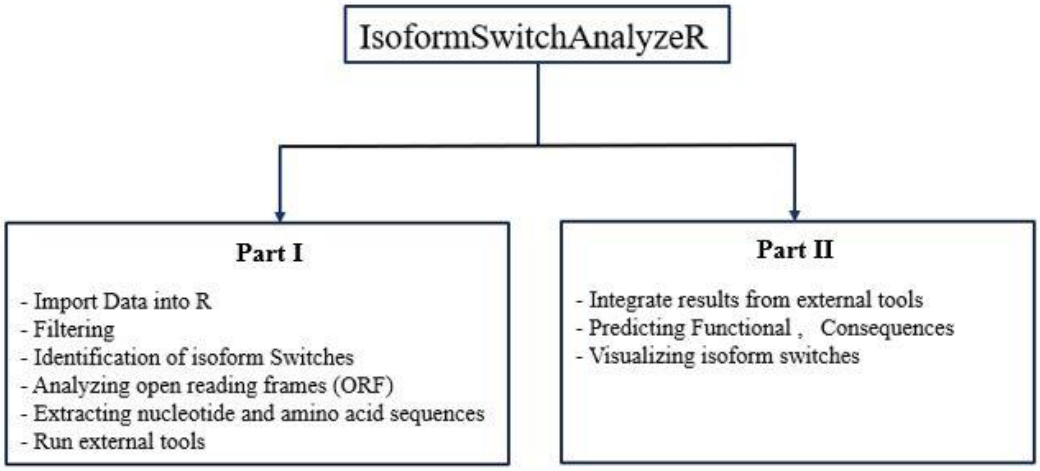

Figure 3-10: R Bioconductor package IsoformSwitchAnalyzeR is divided into two parts based on isoform switch calculation and visualisation.

Figure 3-11: Steps involved in isoform switch pipeline.

3.8 IsoformSwitchAnalyzeR-Part I

In this section, data files were imported into R for isoform switch identification.

3.8.1 Importing Data into R - Preparing Files:

- Four sets of data were required to import to IsoformSwitchAnalyzeR:
- Isoform quantification files (TPM/FPKM/RPKM) and coverage data generated in section 3.7.
- An experiment- design matrix file was made, specifying which sample belongs to which condition.
- The merged transcriptome gtf file section 3.6.1.1 for annotation to help specify which isoform belongs to which gene.
- Nucleotide sequences of quantified isoforms were extracted using gffread utility as explained below in section 3.8.1.1.

3.8.1.1 GFFread Utility

GFFread utility was used to generate the FASTA file for all the transcripts; the FASTA file was provided as a reference [93]. Linux command to extract FASTA sequences of transcripts are mentioned below:

```
gffread -w transcripts.fa -g /path to reference FASTA file/
                        transcripts.gtf
```

Where:

- -w= this indicates to write file for extracted DNA sequences, by default file name is a transcript. fa.
- -g= this Parameter specifies the path to the reference FASTA file

After the preparation, all four data files were imported to IsoformSwitchAnalyzeR.

3.8.2 Importing Quantification Files

The four data files are imported into the IsoformSwitchAnalyzeR as an object. The next step includes abundance estimation using isoform counts to incorporate bias correction in switch identification. Bias correction is done by inter-library normalisation via EdgeR of abundance estimates. By default, the trimmed mean of M-values (TMM) methods is used for normalisation, whereas relative log expression (RLE) and upper-quartile are also present. EdgeR normalises the library size by finding the scaling factor to minimise log fold change between samples [11]. TMM method is widely used for RNA seq experiments, which assumes that most genes are not differentially expressed. TMM normalises the total RNA from samples rather than considering gene length or library size used by other FPKM/RPKM and TPM [94].

After importing the data, the IsoformSwitchAnlayzeR performs the following functions:

- It sums up all isoforms belonging to genes and gets gene expression.
- For each isoform/gene in each condition, it calculates standard error and mean expression.
- For each pairwise comparison of condition, log2FC and isoform fractions (IF) values were calculated using mean gene expression and isoforms expression values.

3.8.3 Filtering

Not all identified isoforms are biologically relevant; hence pre-filtering was done on the following bases:

- Multi Isoform genes - genes with single isoforms were discarded
- Gene Expression - relatively low expressed genes were removed. Therefore, the average expression in both conditions should be greater than 1 (geneCutoff >1).
- Isoform Expression - Unused or relatively low expressed isoforms were removed (isoformCutoff=0.5).
- Isoforms Fraction (Isoform Usage) - Isoforms that were contributing little towards parent gene were removed (IF $\geq$ 0.05)

3.8.4 Identification of Isoform Switches

After the pre-filtering step, a statistical test was performed to identify ADAM9 isoform switching using DEXseq. Isoform switching was identified by two parameters:

- **Statistical Measure:** Switch Q-value cutoff - FDR corrected p-value to check the significance of the identified switch.
- **Effect Size**: dIF cut off - which is a minimum absolute change in differential usage (dIF), where dIF is calculated from isoform fractions (IF1 and IF2) in both normal and tumour conditions, respectively:

$$dIF = IF2 - IF1$$

$$IF = \text{isoforms exp/gene exp}$$

These two parameters work in combination to check the effect size and the statistical significance of isoform switches.

3.8.5 Analysing Open Reading Frames

After identifying isoform switches, the next step was to annotate isoforms by extracting open reading frames from transcript nucleotide sequences using `AnalyzeORF()` function [21]. This function utilises four different methods for ORF prediction for different circumstances:

- **The longest method**= identifies the longest ORF based on a canonical start and stop codon
- **mostUpstream method** = identifies the most upstream ORF based on a canonical start and stop codon
- **longestAnnotated**= identifies the longest ORF downstream the start site
- **mostUpstreamAnnotated** = identifies ORF downstream of the most upstream overlapping annotated stat site.

3.8.5.1 Extracting Nucleotide and Amino acid Sequences

The next step after annotating ORF was to extract their amino acid sequences by translating nucleotide sequences into amino acids, to perform sequence analysis using different internal and external tools like CPC2, Pfam, IUPRED2A and SignalP.

3.8.6 Running External Sequence Analysis Tool

Output files from previous steps were used as input for external tools CPC2, Pfam, IUPRED and SignalP for sequence analysis. Before running these tools, nucleotide and amino acid files were prepared for input by splitting the large files into small chunks to extract transcripts for ADAM9. These transcript sequences were extracted using Seqkit Tool on Linux, an ultra-fast toolkit for manipulating FASTA and FASTQ files, including splitting, filtering, searching, and shuffling [95]. Linux commands for nucleotide, and amino acid sequence extraction are mentioned below:

```
grep -A 1 -wFf < (sed -r 's/^/ENST: /' AminoAcid_list.txt)
   isoformSwitchAnalyzeR_isoform_AA.FASTA > output.FASTA

grep -A 1 -wFf < (sed -r 's/^/ENST: /' Nucleotide_list.txt)
   isoformSwitchAnalyzeR_isoform_nt.FASTA > output.FASTA
```

After extracting transcript sequences, external tools for sequence analysis were run on web servers.

3.8.6.1 CPC2

This tool was used for the assessment of isoforms coding potential. CPC2 can be used as a standalone downloadable package and available on a web server [96]. CPC2 uses a nucleotide file (_nt. FASTA) generated from the previous step with default parameters.

3.8.6.2 PFAM

In addition, PFAM was used to predict the domain for isoforms of interest. PFAM is a protein database that gives the functional overview of a protein family and domain [97]

- PFAM uses an amino acid file (_AA.FASTA) generated from the previous step with default parameters.

3.8.6.3 IUPRED2A

All proteins contain different domains, and some are intrinsically disordered regions (IDRs), polypeptide regions that lack hydrophobic amino acids to initiate folding. Thus, it lacks a fixed or ordered three-dimensional structure [98]. This tool was used to predict (IDRs); this tool can be run locally or via a web server [99].

- IUPred2A uses an amino acid file (_AA.FASTA) generated from the previous step with default parameters.

3.8.6.4 SignalP

This tool was used to predict signal peptides, small amino acid sequences in newly synthesised proteins that help proteins move across the membranes [100]. SignalP uses an amino acid file (_AA.FASTA) generated from the previous step with default parameters.

3.9 IsoformSwitchAnalyzeR-PART II

In this section, functional consequences for identified switches were predicted along with isoform switch visualization.

3.9.1 Predicting Functional Consequences of Switch

The list (SwitchListAnalyzed) contains all the objects from previous steps consisting of isoforms quantifications, isoforms switches and ORF annotation from external tools. The next step was to predict the functional consequence for identified isoforms switches only if there is a significant change in the isoform's contribution towards parent gene expression. `analyzeSwitchConsequences()` function extract isoforms that significantly change their isoform usage, calculated by α and dIF parameters in previous steps.

3.9.2 Visualising Isoform Switches

After the identification of isoform switches, visualisation was done. Individual isoform switch analysis (used in this study ADAM9 gene) and Genome-wide analysis of isoform switching are the two types of post-analysis visualisation supported by the IsoformSwitchAnalyzeR. `extractTopSwicthes()` function was used to extracts top switches from IsoformSwitchAnalyzeR based on smallest q-value and largest absolute dIF value (by default extract top 10 switches). The `switchplot()` function is used to plot the isoform switch and the predicted functional consequences such as protein domains, signal peptides, and intrinsically disordered regions.

3.10 Correlation Analysis - PART III

The last phase of this research was to perform a correlation analysis for identifying interacting partners for the ADAM9 gene. Before correlation analysis, differential expression analysis was performed using the Deseq2 tool.

3.10.1 Differential Expression Analysis using Deseq2

Deseq2 tool provides different methods to test differentially expressed genes. It uses a negative binomial model for differential expression analysis. Deseq2 analyses the RNA-seq count data in table format to identify differentially expressed genes between different conditions [101].

In previous steps, transcript quantification was performed using StringTie. For assembled transcript quantification, the ballgown tables were created using the -B parameter, which generate six files, including quantification file (sample_quant.gtf) for each sample that contains FPKM (Fragment per kilobase of transcript per million per read) and TPM (Transcript per million) values along with coverages.

3.10.1.1 StringTie with Deseq2

As mentioned earlier, Deseq2 takes the input of count tables for differential expression analysis. Python (version 2.6.6) script PrepDE.py was used to extract count information

from StringTie quantification files (sample_quant.gtf) created using the -B parameter. PrepDE.py script extracts the count data for each transcript from coverage values estimated by StringTie using the following formula:

ReadCounts_Per_Transcript =coverage * transcript_length/ read_length

Input provided to prepDE.py was a text file (sample_list.txt) containing quantification files (sample_quant.gtf) for each sample with their respective path. The following Linux command generates two CSV files: gene_count_matrix and transcript_count_matrix

```
python prepDE.py -i sample_list.txt
```

The gene_count_matrix file contains count data for normal and tumour samples, so it cannot be given as an input to Deseq2. To overcome this issue, the gene_count_matrix file was split into individual CSV files so that each file represents single sample counts. These count matrices CSV files were then uploaded into the Galaxy tool to use Deseq2 for further analysis. Galaxy is a publicly available web portal used for intensive genomic analysis [102]. Galaxy provides users with many tools needed for bioinformatics analysis and cloud storage to store results. In addition, it enables data integration from several sources, including users' computers, URLs and different online databases [102].Deseq2 outputs two files, one with normalised counts and the other with fold change values. For better insights into Deseq2, normalised counts were joined with the file containing p-values and fold-change values (logFC). After joining both files, gene names were mapped to extract their Entrez id using the **annotateMyIDs** tool.

3.10.2 Identification of Interacting Partners

After identifying DEGs (differentially expressed genes), the next step was to perform correlation analysis to find the interacting partners among identified DEGs for ADAM9.

3.10.2.1 Correlation Analysis in R

Correlation is a statistical measure that describes the strength of relationship and directionality between two variables, calculated by the correlation coefficient. There are

different statistical methods available for correlation analysis, such as Pearson, Spearman and Kendall. Pearson correlation shows the linear relationship between two variables, whereas Spearman and Kendall Correlation are non-parametric measures based on rank values of variables [103]. The most widely used method is the Pearson correlation coefficient.

Sequential steps performed for correlation analysis using R Bioconductor are mentioned below:

- The working directory was set.
- Count file for normal samples was read using read.csv.
- Data frame was made for count data.
- After loading the data frame, the excel sheet accession number of the ADAM9 gene was specified. `cor()` was used to compute correlation coefficient, and `cor.test()` was used to check the association between genes, which returns two values correlation coefficient and significance level (`cor_R` and `cor_P`, respectively).
- Filter the genes with correlation values ranging from -0.7 to +0.7 for further analysis.
- The same steps were performed for tumour count files.

3.10.3 Gene Set Enrichment Analysis

After getting interacting partners, the next step is the gene set enrichment analysis, abbreviated as GSEA. GSEA is a statistical method that describes the statistical significance of the genes to specific Gene Ontology (GO), KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways Reactome, Biocarta and many other pathway analyses. GSEA identifies the set of enriched genes in a particular dataset compared to control (In this study, it is between +ve and -ve correlated genes). Enrichment analysis is performed using the GSEA tool installed on PC. Steps performed are as follows:

- Correlation files for each dataset were prepared.RNK syntax.
- These files were then loaded in GSEA.

- Annotation file and gene set database parameters are specified.
- After loading the files and setting the parameters, GSEAPreranked is run.
- GSEA result analysis folder is saved containing different files and plots.

GSEA computes enrichment analysis based on Enrichment score (ES) and Nominal P-value. ES determines which gene set is over-represented in the top and bottom of the ranked correlated gene list. Positive enrichment score (ES) shows gene set enriched at the top of the list, whereas negative ES shows gene set enriched at the bottom. The nominal P-value evaluates the statistical significance of ES, showing the likelihood that this gene set is enriched in a pathway.

Chapter 4

Results and Discussion

Oesophageal cancer (ESC) is one of the deadliest and least studied cancer worldwide. Nevertheless, its aggressive nature and the low mortality rate remains a public health concern worldwide [28]. ADAM9 gene with two alternatively spliced transcripts (L and S forms) is involved in many cancers. However, differential usage of its transcripts in oesophageal cancer is not studied.

Therefore, the focus of the study was to understand the role of ADAM9 transcripts in oesophageal cancer, as stated earlier in chapter 2.2.2. The importance of differential transcript usage DTU due to its significant role in regulating various biological, including development, homeostasis, pluripotency, and apoptosis processes. The change in DTU is referred to as the isoform switch involved in different diseases and prominently in cancers. The functional impacts of these isoform switches lead to the gain or loss of functional domains and protein-binding activities (PPI) in cancer-related pathways [104].

4.1 Alignment to Genome

A splice aligner performed the alignment step using the Hisat2 tool [18]. Hisat2 maps RNA-seq reads using global indexing, which represents whole-genome and local indexing (small indexing). This collective usage of both indexings enables the effective alignment of reads as it spans multiple exons [91].

4.1.1 GSE130078

RNA-Seq paired tumour and non-tumour samples of 23 SCC patients (South Korea). Total 46 samples, out of which 23 were normal, and 23 were tumour samples. None of the samples had adapter content, with the overall alignment rate was more than 95% for all samples. Details are provided in Appendix A.

4.1.2 GSE111011

This dataset contained 14 samples in an equal ratio of RNA-Seq paired tumour and non-tumour samples. There was no adapter content similar to the previous one, and the alignment rate was more than 95%. Summary statistics of alignment scores for both normal and tumour samples are shown in Appendix A.

4.1.3 E-MTAB-4054

This dataset was different from previous ones as it contained 16 EAC patient samples, and 11 were normal. The samples were of good quality, with no adapter content and an overall alignment rate of 95% for all samples. Alignment scores are shown in Appendix A.

4.2 Transcriptome Assembly and Reconstruction

RNA-Seq produces millions of short reads, making assembly of these short reads into complete transcripts a complex task; as the mechanistic alternative splicing yields multiple isoforms sharing common exons; to identify, assemble and quantify the abundance of these isoforms is a challenging task. After the initial estimation, reconstruction of the transcriptome was performed by merging all gene structures found in any samples. This step is crucial as transcripts in some samples might be only partially covered by reads, resulting in only a partial version of the transcript is assembled. Merging creates the reconstructed transcripts consistent in all samples with re-estimated abundances, as the reads are reallocated for transcripts whose structure is changed due to merging [91].

In this study, StringTie was used as it can work on two main approaches: reference-based and *de-novo* assembly for transcriptome assembly and reconstruction using mapped reads. A reference-based approach was used in this study based on the following reasons:

- Reference genome for humans is publicly available.
- The presence of multicopy genes, multiple isoforms for the same gene and large variation in expression levels makes assembly difficult without any reference genome.

Reference-based approach clusters mapped reads and build splice graphs that represent all possible transcripts isoforms for each gene. After identifying transcripts, network flow for each transcript was created using a maximum flow algorithm to simultaneously assemble and estimate the expression levels of transcripts in ballgown readable format. After the initial estimation, transcripts were subjected to the merging step, all gene structures found in any samples were merged. Reconstruction of the transcriptome is crucial as transcripts in some samples might be only partially covered by reads, resulting in only a partial version of the transcript is assembled. Merging creates the reconstructed transcripts that are consistent in all samples. Furthermore, re-estimation of transcripts abundances is done as the reads are reallocated for transcripts whose structure is changed due to merging [105].

The primary output file generated by StringTie was a quantification file in GTF format containing all the details of assembled transcripts. Descriptors for GTF file are Seq name, source, feature, start, end, score, strand, frame, attribute.

4.3 Isoform Switch Analysis

After the initial quantification steps, Dexseq was used for the statistical identification of isoform switches by calculating the differential usage of L and S-forms of ADAM9 in both normal and tumour condition in all datasets. Table ***4-1*** shows the data generated by DexSeq for ADAM9 isoforms in all datasets based on the criteria described in the previous section 3.8.4. Where gene value represents the collective expression of all transcripts contributing towards gene expression in normal and tumour samples, similarly isoform value represents the individual isoform expression (FPKM) in a tumour, and normal samples, gene Fold Change (FC) and isoform FC represents the overall change in the expression of both gene and isoform between normal to tumour samples. In addition, differential isoform fraction (dIF) shows the shift in isoform fraction (IF) values from normal to tumour (negative dIF indicates the decreased usage of respective isoform in tumour samples, whereas positive dIF shows the increased isoform usage in normal samples). Finally, Q-value represents the significance level of identified isoform switch.

Table 0.1: Output Generated by Dexseq for ADAM9 Isoforms. Differential usage of L and S-forms along with the isoform fractions, isoforms/gene expressions and switch Q-values in all datasets: GSE130078, GSE111011 and E-MTAB-4054

Isoform_Id	GeneValue_ Normal	GeneValue_ Tumour	Gene_ FC	Isoform_value _Normal	Isoform_value _Tumour	Isoform_FC	IF_overall	IF_Normal	IF2_Tumour	dIF	IsoformSwitch_ Q_value
130078											
ENST00000379917(S)	28.17	30.09	0.09	3.027	0.782	-1.93	0.065	0.10	0.02	-0.07	0.00503
ENST00000487273(L)	28.17	30.09	0.09	16.27	21.61	0.40	0.641	0.53	0.74	0.21	0.0005
111011											
ENST00000379917(S)	23.75	45.99	0.953	1.865	0.640	-1.526	0.046	0.07	0.01	-0.06	1.43E-16
ENST00000487273(L)	23.75	45.99	0.953	17.48	39.73	1.183	0.803	0.73	0.86	0.12	7.17E-10
4054											
ENST00000379917(S)	38.77	61.61	0.667	5.037	1.248	-2.00	0.069	0.13	0.02	-0.10	3.69E-07
ENST00000487273(L)	38.77	61.61	0.667	28.41	57.31	1.011	0.835	0.71	0.91	0.19	2.98E-06

Table 0.2**:** Where dIF=>=0.05, IsoformSwitch_Q_value<=0.05. Where, IF=Isoform Fraction, FC=Fold change, dIF=Differential Isoform Fraction

4.3.1 Switch Plot

Gene and isoform expression plots are formed using FPKM values, whereas isoform usage/switch plots are formed based on IFcutoff, using isoform fractions (IF) mentioned above in Table ***4-7***. Furthermore, the dIFcutoff parameter was used to check the trend of isoform usage (increase/decreased/unchanged).

4.3.1.1 GSE130078

Plots for gene expression, isoform expression and differential usage of ADAM9 isoforms in the GSE130078 dataset are shown in Figure ***4-1***, ***4-2***, ***4-3***, respectively. Dexseq identifies 13 isoforms initially; However, most of these isoforms were chunks of domains of the gene, and when IFcutoff was utilised, these isoforms were no longer shown as significant. Figure ***4-1*** shows an increase in ADAM9 gene expression, the collective expression of all isoforms in tumour samples. In contrast, isoform expression and usage plots show the individual isoforms (Figure ***4-2***,***4-3***). Their plots show L-form is highly expressed in primary oesophageal cancer samples, whereas S-form is expressed in normal samples shown in Figures ***4-2*** and ***4-3***.

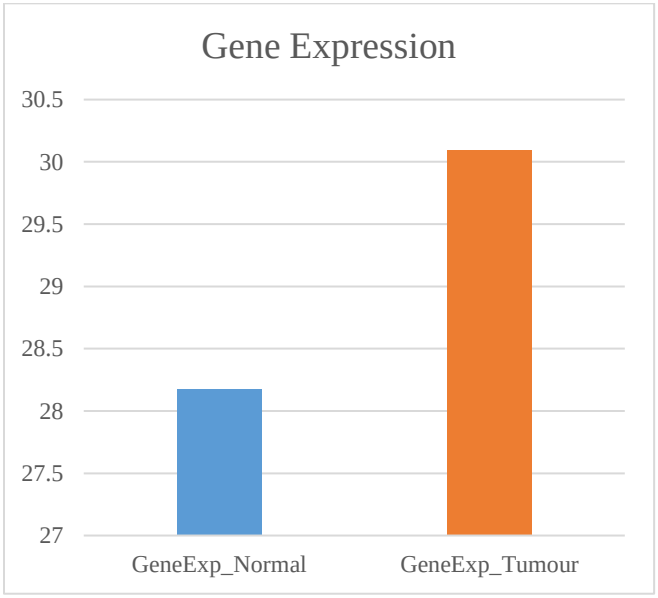

Figure 4-1: Show gene expression. Where: GeneExp_Normal =gene expression in normal, GeneExp_Tumour= gene expression in tumour.

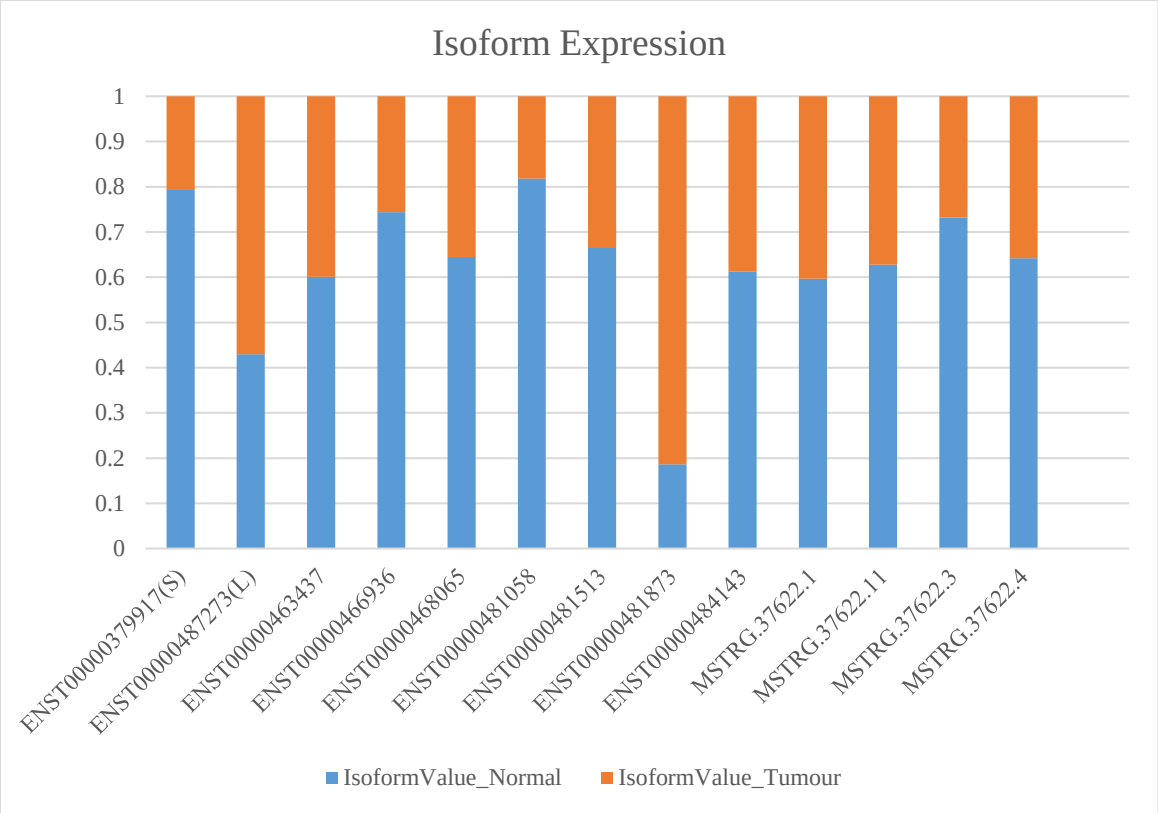

Figure 4-2: Shows Isoform Expression. Where: IsoformValue_Normal = isoform expression in normal, IsoformValue_Tumour= isoform expression in tumour.

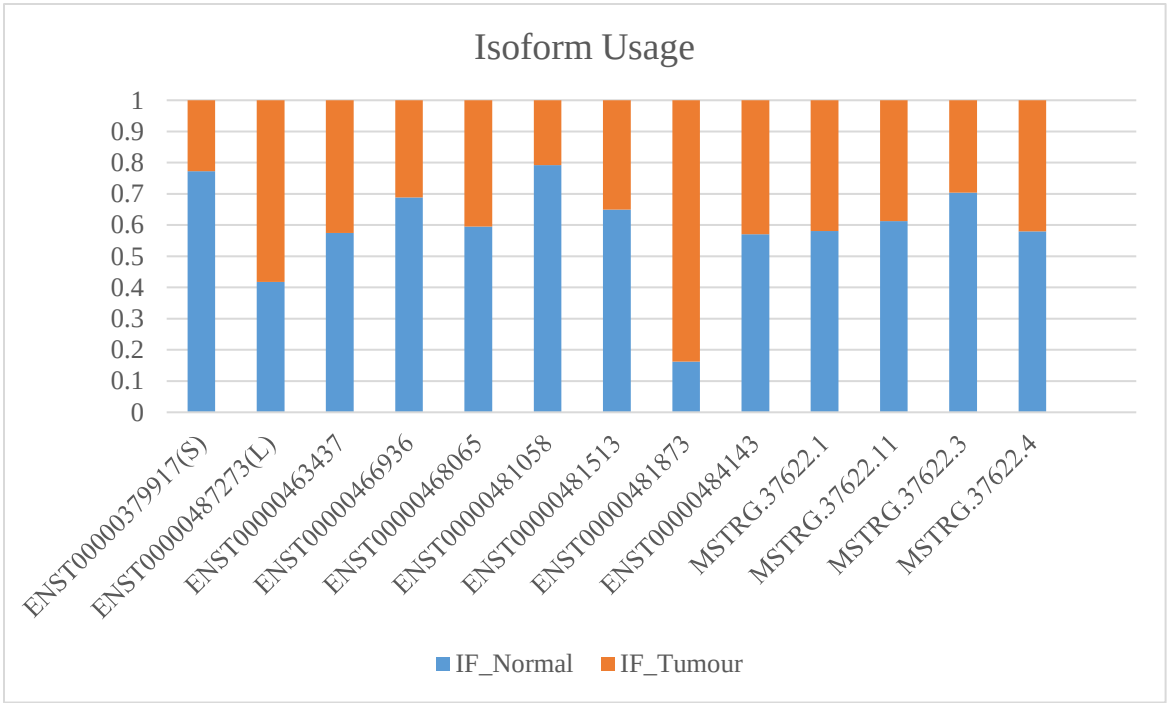

Figure 4-3: Shows Isoform Usage. Where: IF_Normal= isoform usage/isoform fraction in normal, IF_Tumour= isoform usage/isoform fraction in tumour samples.

4.3.1.2 GSE111011

Dexseq initially identifies six isoforms in GSE111011; however, these transcripts were not significantly expressed and thus were dropped after IFcutoff was utilised. The analysis showed that ADAM9 was expressed more in tumours than in normal samples shown in Figure ***4-4***. Gene expression is calculated by adding all the individual isoform expressions; this includes the isoforms considered insignificant by the IFcutoff. Figure ***4-5*** and Figure ***4-6*** represent isoform expression and usage plots of individual isoforms. It can be seen from the plots that the L-form is highly expressed and is the main source of overall gene expression in tumours, and the S-form primarily decreases in expression and usage.

Figure 4-4: Shows gene expression. Where: GeneExp_Normal =gene expression in normal, GeneExp_Tumour= gene expression in tumour

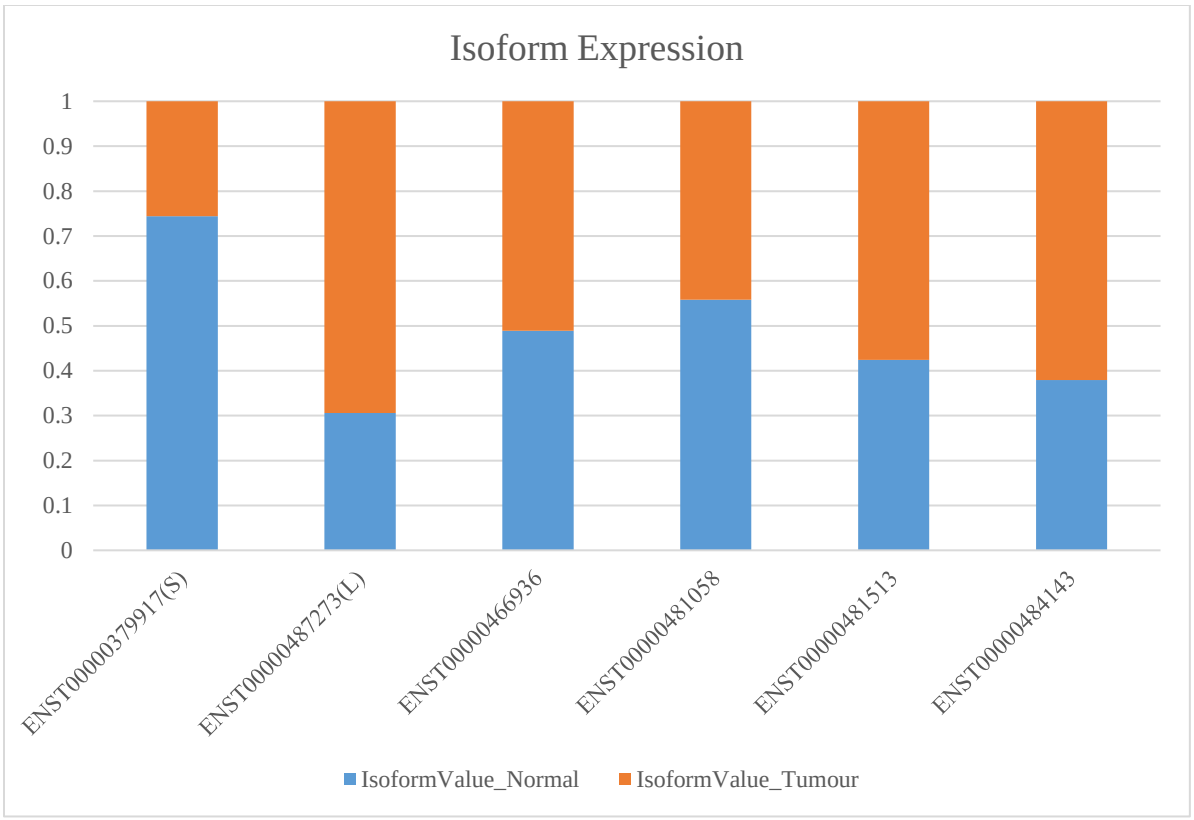

Figure 4-5: Shows isoform expression: Where: IsoformValue_Normal = isoform expression in normal, IsoformValue_Tumour= isoform expression in tumour

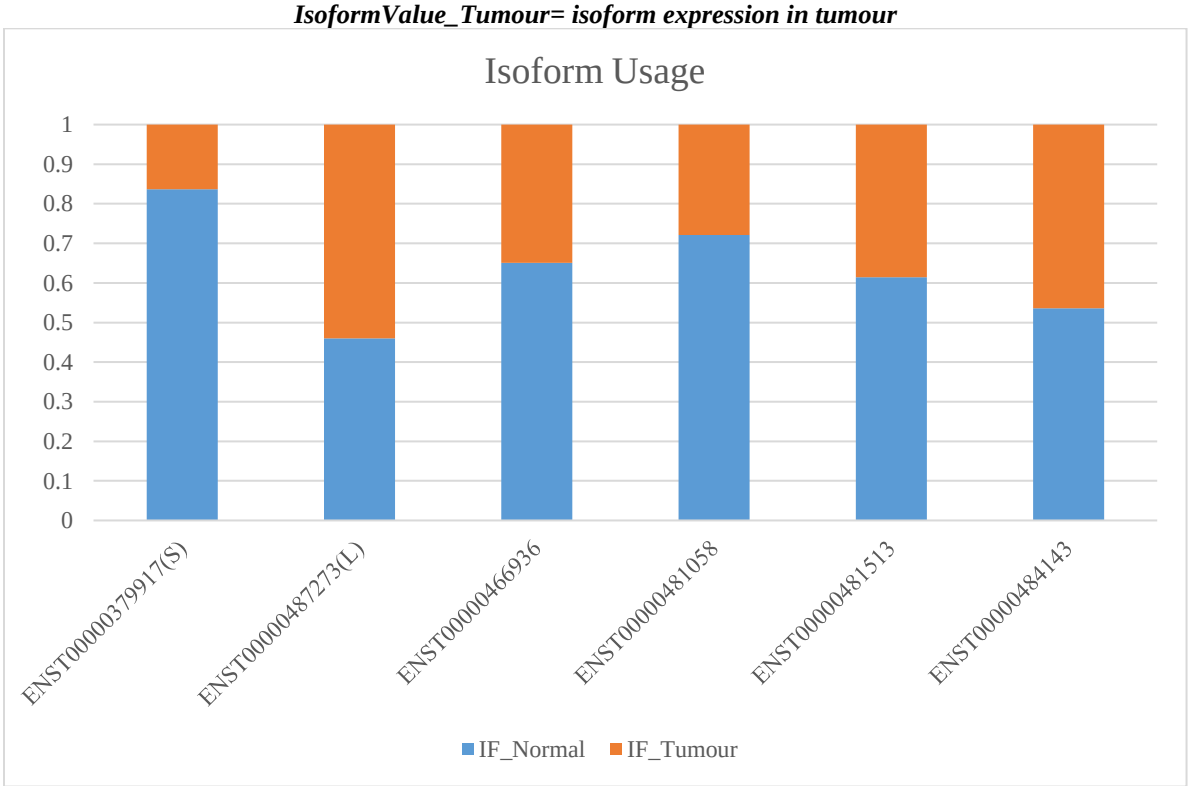

Figure 4-6: Shows Isoform Usage. Where: IF_Normal= isoform usage/isoform fraction in normal, IF_Tumour= isoform usage/isoform fraction in tumour samples.

4.3.1.3 E-MTAB-4054

Gene expression in Figure ***4-7*** shows that ADAM9 is highly expressed in tumour samples compared to normal, similar to the previous one GSE111011. Dexseq identifies ten isoforms, but on 0.05 IFcutoff, only two (L and S) isoforms are expressed significantly. Isoform expression and differential usage plots show L-form is highly expressed in primary oesophageal cancer samples, whereas S-form is expressed in normal samples shown in Figures ***4-8*** and ***4-9***.

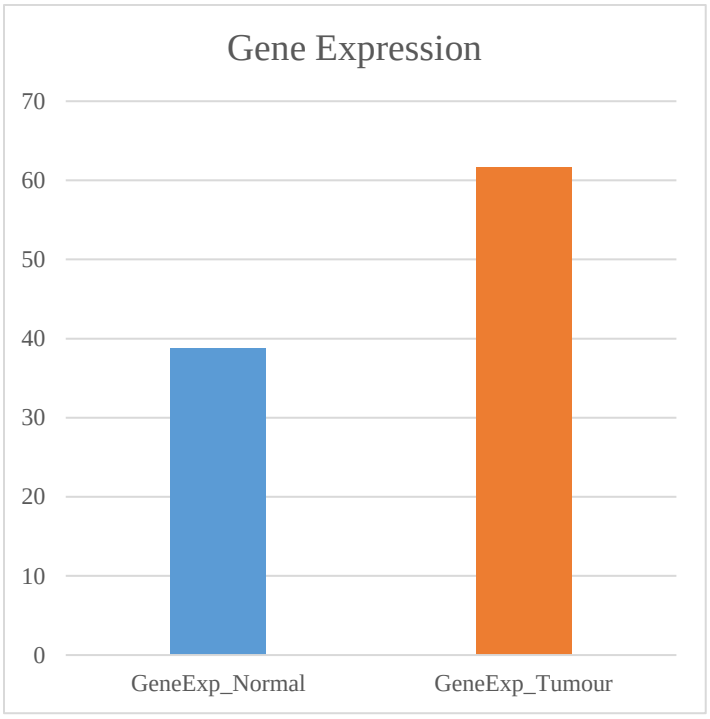

Figure 4-7: Shows Gene Expression. Where: GeneExp_Normal =gene expression in normal, GeneExp_Tumour= gene expression in tumour

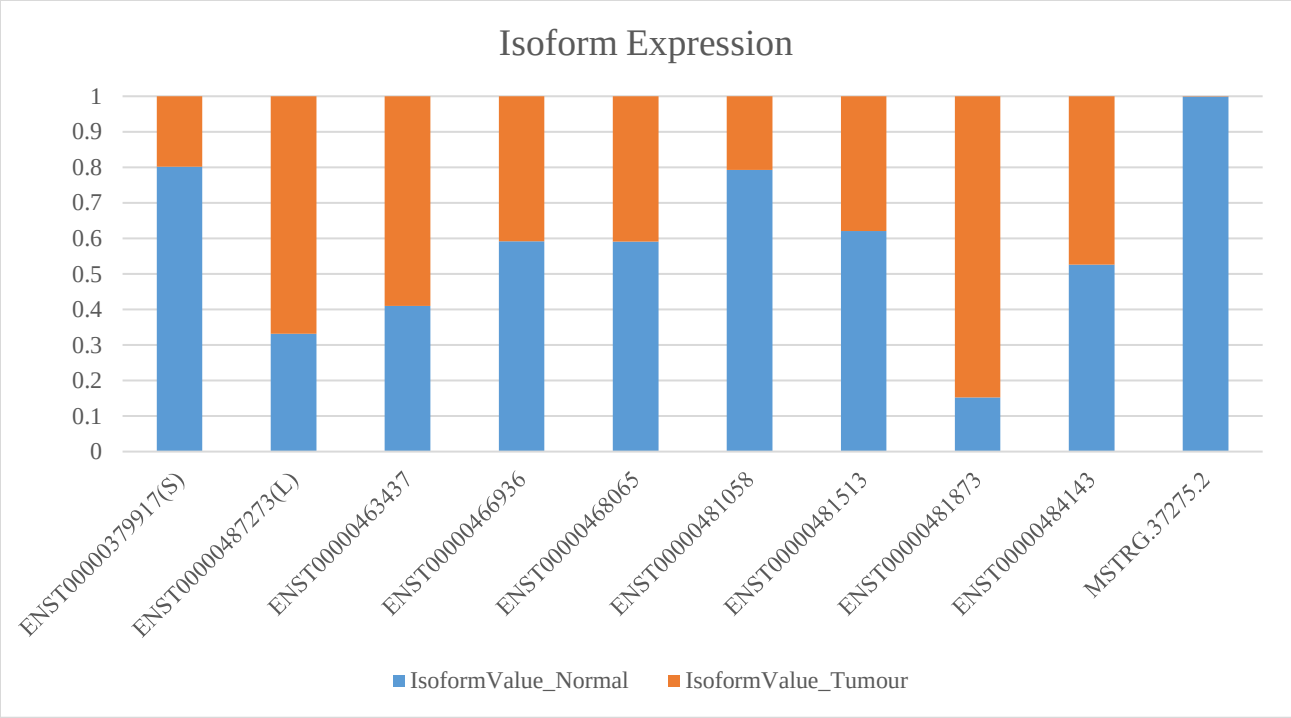

Figure 4-8: Shows Isoform Expression. Where: IsoformValue_Normal = isoform expression in normal, IsoformValue_Tumour= isoform expression in tumour

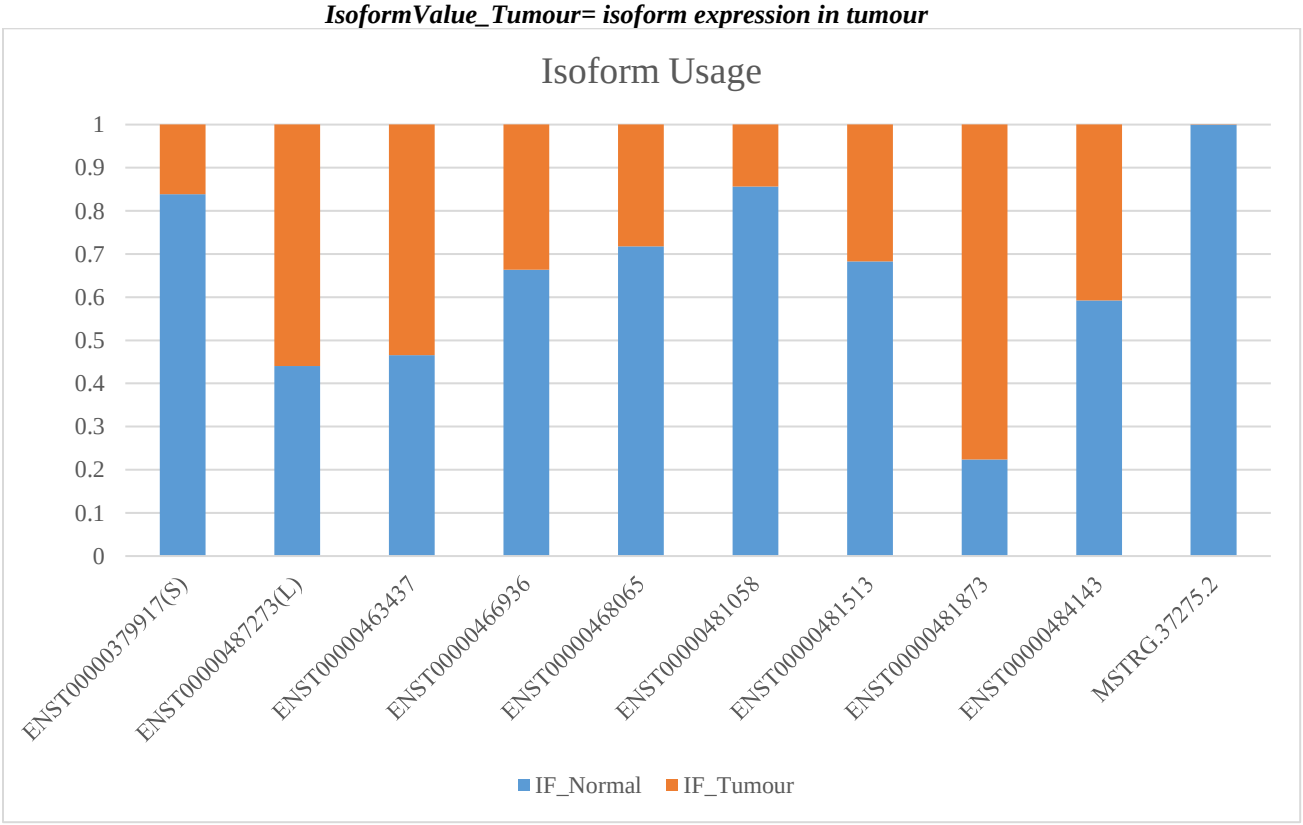

Figure 4-9: Shows Isoform Usage. Where: IF_Normal= isoform usage/isoform fraction in normal, IF_Tumour= isoform usage/isoform fraction in tumour samples

The overall trend seen in the above plots represents significantly increased gene expression in oesophageal tumour samples as compared to normal samples, which can be due to significant changes in the isoform usage of L and S-forms across both conditions shown in bottom plot Figure ***4-3***, ***4-6*** and ***4-9***. Furthermore, isoform expression and usage plots illustrate that L isoform contributes to the ADAM9 gene expression in primary tumour samples compared to S-form.

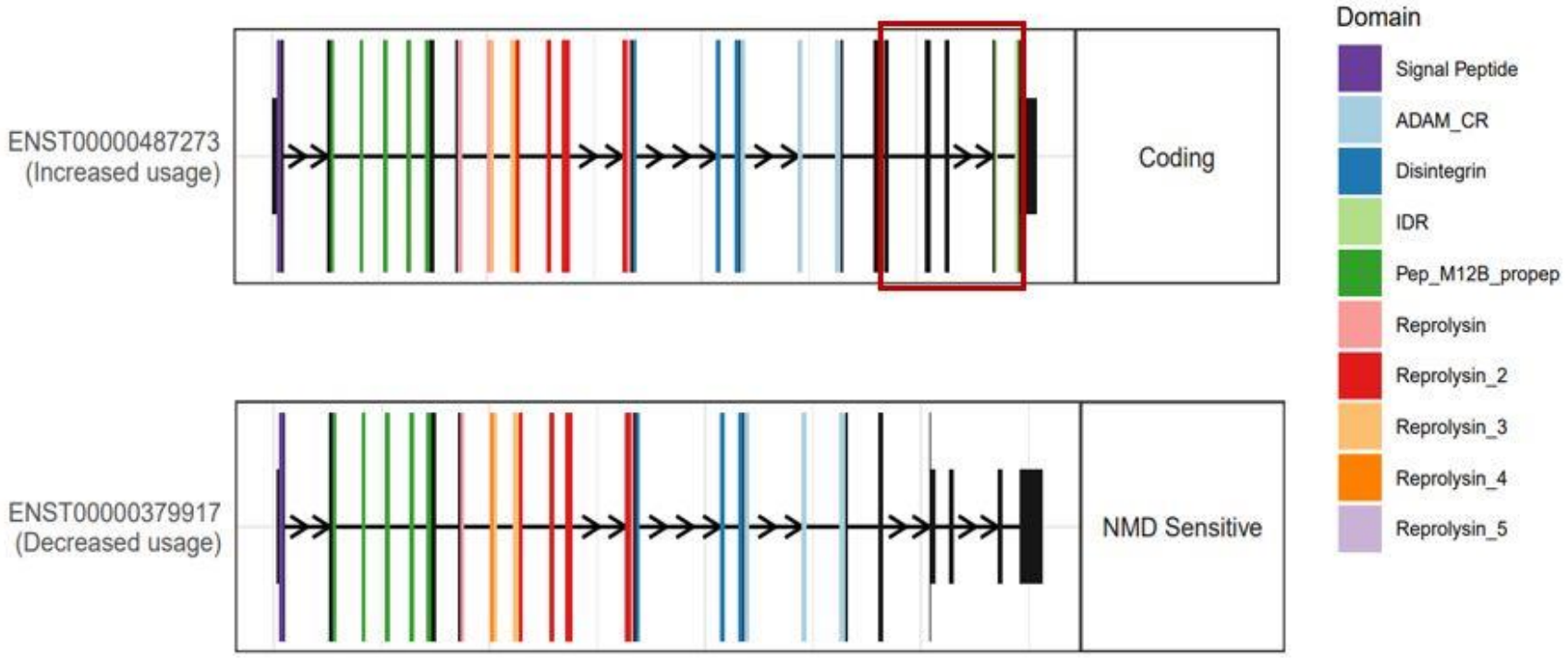

Figure 4-10: Isoform structure, showing the trend of ADAM9 isoforms (L and S forms)

By comparing the isoform usage plots to the isoform structure (showing the isoform trend) shown in Figure ***4-10***, it is inferred that in primary tumour samples, L-form is used. As discussed earlier in section 2.2.2, ADAM9 is involved in different types of cancers. Its isoform switch is well studied in breast cancer. The L-form was upregulated during primary cancer stages in breast cancer studies, and the S-form mainly indicates metastasis. In this study, secondary data was processed using the IsoformSwitchAnalyzeR pipeline. The samples were mainly from the primary tumour site of EC, and thus the increased expression of L-form is similar to what is observed in breast cancer primary tumour site. As discussed in the section, the L-form is involved in the initial stages of cancer and affects cell proliferation and adhesion. ADAM9(L) mediates tumour cell proliferation by hyper activating EGFR signalling; further triggering the downstream signalling of Ras/mitogen-activated protein kinase MAPK and PI3K/AKT and transduces the signals for activation of cell growth genes to promote tumour cell proliferation [106], shown in Figure ***4-11.*** EGFR is a receptor to many ligands and regulates different processes, including proliferation and differentiation. The heparin-binding-EGF((HB-EGF) and transforming growth factor-alpha((TGF-alpha) ligands (Kataoka, 2009) are cleaved by ADAM9, making them biologically active proteins to initiate proliferation. Jeong Min Kim and colleagues have validated this phenomenon who performed the knockout study to explain the role of ADAM9 in tumour growth/progression. They found that in the absence of ADAM9, EGFR phosphorylation was blocked, which decreased the EGFR activation and downstream signalling of Ras/MAPK [107].

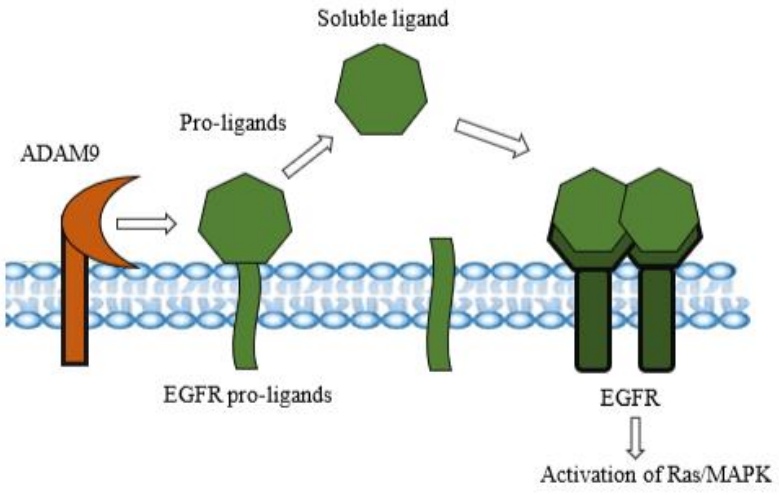

Figure 4-11: ADAM9 activates EGFR signalling by cleaving pro-ligands to soluble forms (HB-EGF and TNF-alpha), which further activates the Ras/MAPK signalling.

Furthermore, the increased expression of L-form in primary tumour samples also might be due to the presence of the transmembrane domain highlighted in Figure ***4-10,*** which keeps L-form membrane-tethered.

In contrast, S-form lacks this transmembrane domain and is secreted outside the cell. As mentioned earlier in section 2.2.2, the L-form suppresses the cancer cell invasion. It remains membrane-tethered by binding integrin protein via its disintegrin domain, altering the integrin-mediated signals for ECM degradation. The secreted Golgi vesicle help S-form to be secreted outside the cell and degrades ECM through the metalloprotease domain, eventually metastasising the tumour cells [77]. Another study conducted by Fry and Toker (2010) saw that the switch from L to S happens when the primary tumour has progressed to a stage where metastasis is imminent; this ties in with the fact that S is secreted form of ADAM9 and its metalloprotease domains are needed for the degradation of ECM so that it is easier for cancerous cells to break free and colonise other organs [77]. This study recapitulates with the previous research findings, where L-form is involved in the primary stage of oesophageal cancer. However, the cancer is localised to oesophageal only and is not distantly metastasised to other regions.

4.4 Identification of Interacting Partners for ADAM9

For a better understanding of molecular functions and pathways, identifying closely correlated gene partners is essential [108]. Through Pearson correlation, interacting partners for ADAM9 were identified. The total number of correlated genes for dataset GSE130078, GSE111011 and E-MTAB-4054 were 105, 5021 and 565, respectively, based on correlation value >=+-0.7.

4.4.1 Gene Enrichment Analysis

Correlated genes obtained from the previous analysis were subjected to the GSEA to identify enriched biological pathways in Oesophageal cancer. GSEA identifies the set of genes that are enriched in a particular biological pathway. Unlike Gene Ontology, it does not only consider a subset of a gene with a significant change in gene expression. Instead, it considers the change in gene expression of all genes in that dataset. Molecular functions and pathways were identified through various pathway databases (Kegg, Reactome, Biocarta and MsigDB) integrated into GSEA by changing parameters e-g number of minimum genes in the gene set, while other

parameters were on default setting. Subsequently, Kegg and Reactome pathways were selected as they share common pathways in all datasets. Table ***4-2*** shows the common pathways seen in all three datasets in Kegg and Reactome databases.

Table 0.2: Common pathways in KEGG and Reactome databases in all three datasets

Database	Pathway
KEGG	Focal Adhesion
Reactome	Vesicle Mediated Transport
	Membrane Trafficking

Table 4-2: KEGG: Kyoto Encyclopaedia of Genes and Genomes

Figures ***4-12, 4-13, 4-14*** show the enrichment plot for focal adhesion in GSE130078, GSE111011 and E-MTAB-4054, respectively. The top portion of each plot shows the ES for gene set as the analysis walks down the ranked list. For example, in Figures ***4-12***, ***4-13*** and ***4-14***, the score at the peak of the plot (the score furthest from 0.0 shown y-axis) is the ES of the gene set, which is 0.26, -0.55, and -0.16, respectively. Thus, positive ES indicates that the gene set enriched at the top of the ranked gene list is involved in the focal adhesion pathway, whereas negative ES indicates that the gene set enriched at the bottom of the ranked gene list is involved. The middle portion of the plot shows where the gene set members appear in the ranked list of genes. Finally, the bottom portion of the plot shows the ranking metric as moving down the ranked list, which shows the gene correlation with the phenotype (positively or negatively correlated)—the value of the ranking metric ranges from positive to negative as moving down the ranked list. Thus, a positive value indicates correlation with the positively correlated genes, and a negative value indicates correlation with the values of the negatively correlated ranked genes.

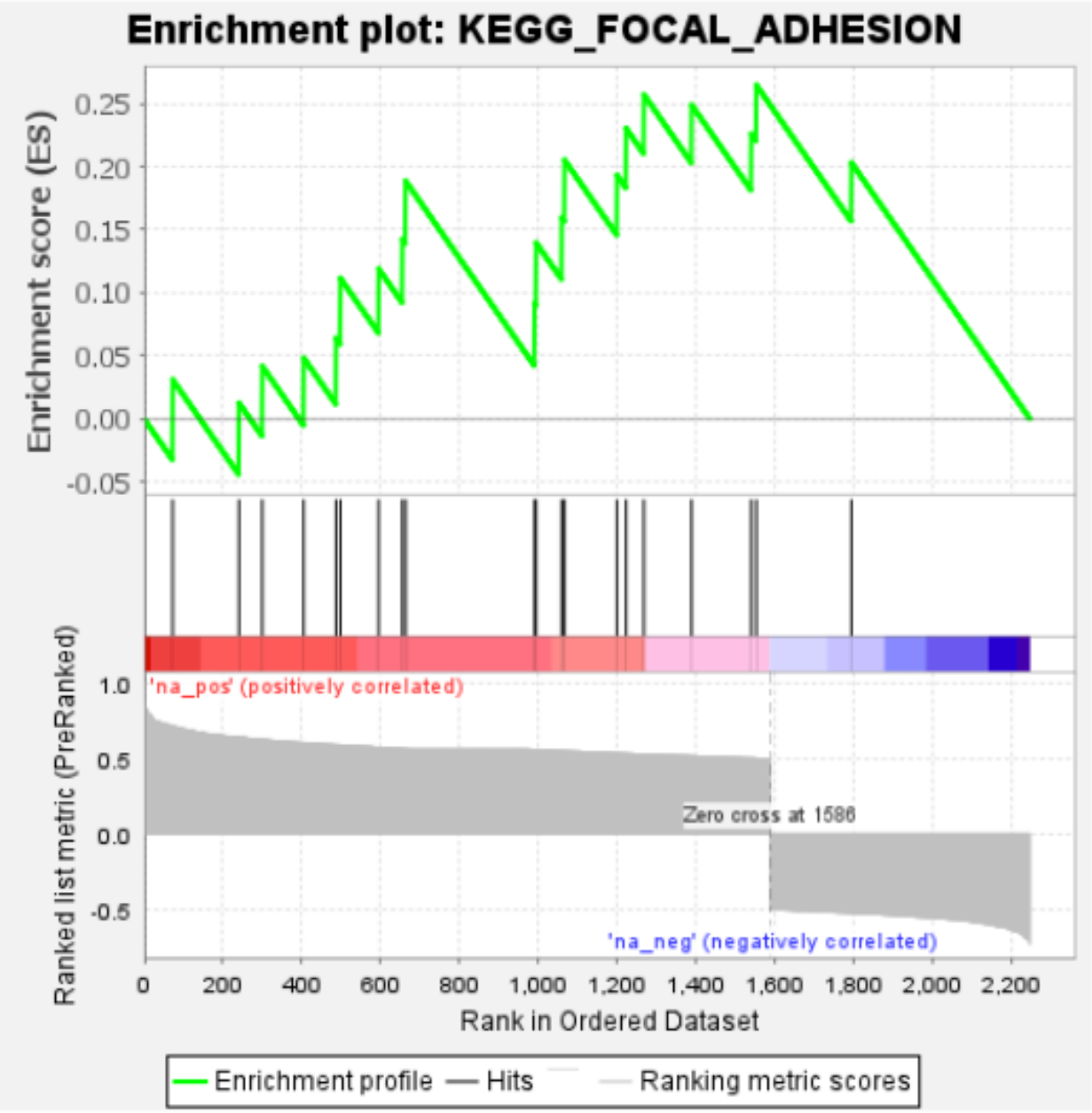

Figure 4-12: GSE130078 Enrichment plot: Focal Adhesion, showing the profile of the running ES score and positions of gene set members on the ranked ordered list.

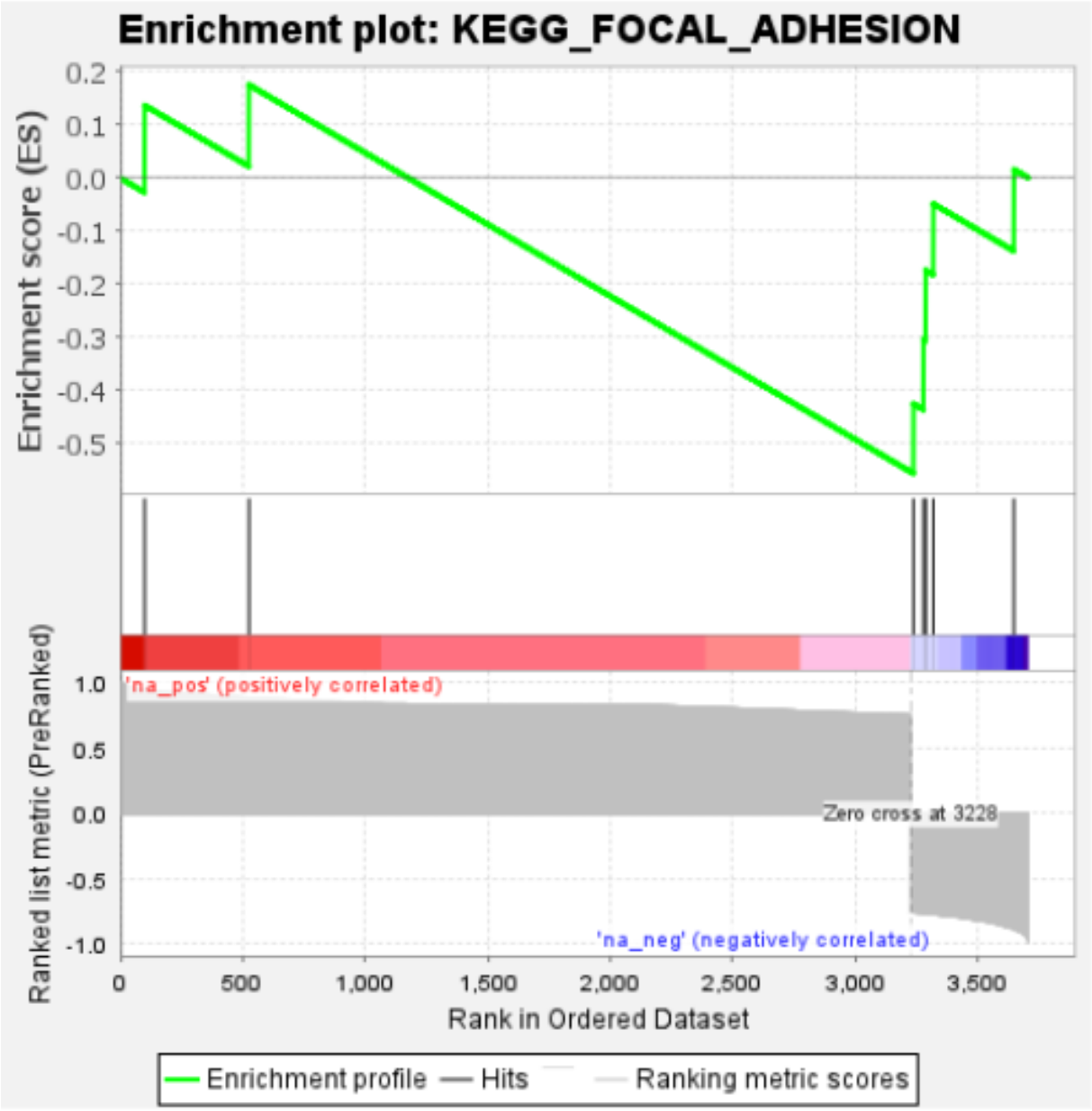

Figure 4-13: GSE111011 Enrichment plot: Focal Adhesion, showing the profile of the running ES score and positions of gene set members on the ranked ordered list.

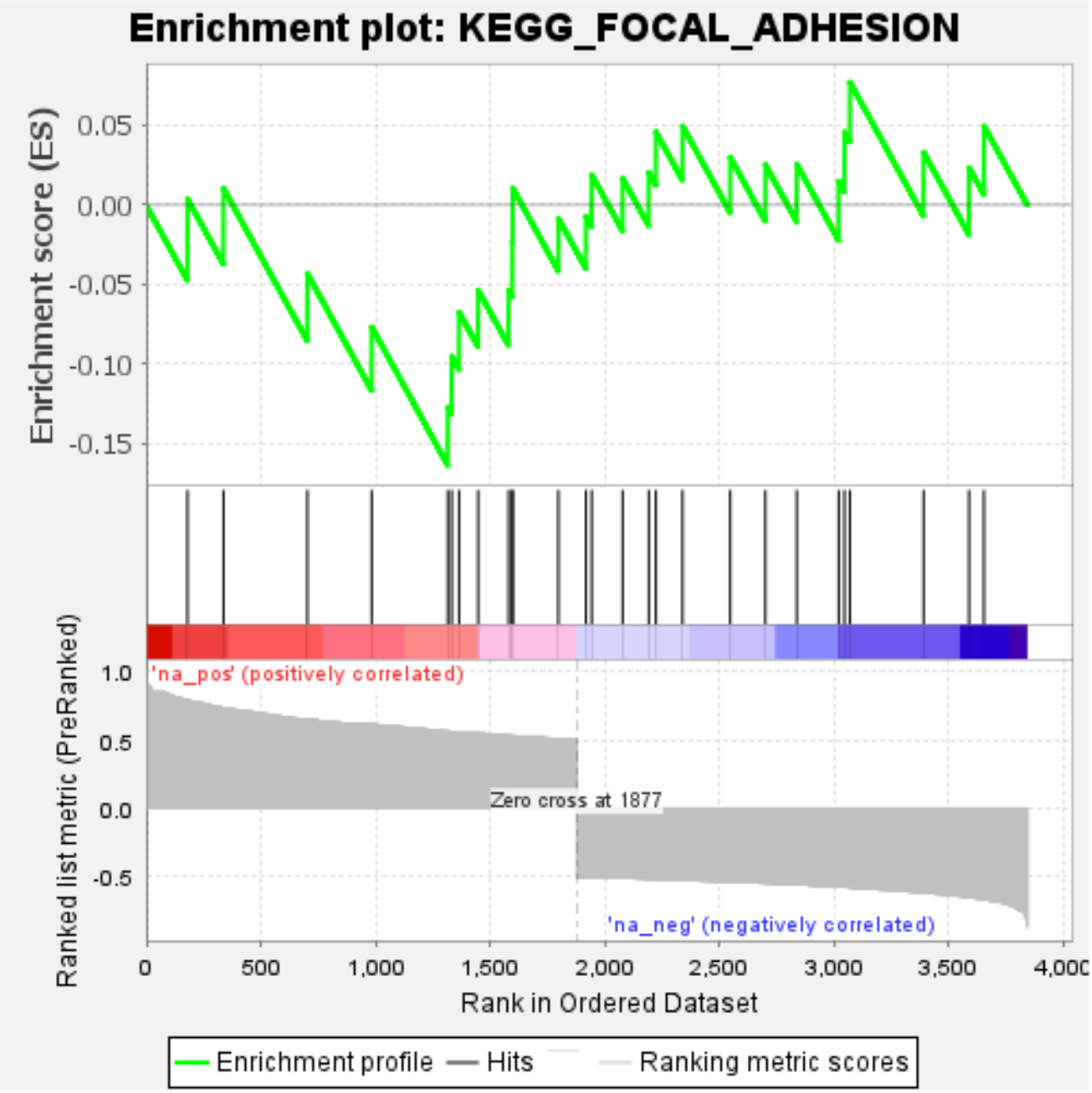

Figure 4-14: E-MTAB-4054 Enrichment plot: Focal Adhesion, showing the profile of the running ES score and positions of gene set members on the ranked ordered list.

The flow of information in cells occurs through the ECM interactions, and these interactions occur at specialised zones of the cell surface that are focal adhesions. These focal adhesions are rich in integrin adhesion receptors which play an essential role in bi-directional transmembrane communication by connecting cell cytoskeletons to the extracellular membrane matrix. These focal adhesions sense extracellular matrix and different physiological and mechanistic stress

signals. In response to these focal adhesion signalling, the cell initiates diverse processes, including cell growth or death, cell motility and cytoskeleton reorganisation [109]. These specialised focal adhesions are seen altered in cancer [110]. ADAM9 plays a crucial role in tumorigenesis and is upregulated in many cancers. ADAM9 is reported to be multifunctional due to its multi-domain structure. One of the prominent proteolytic roles is the shedding of membrane-anchored substrates e-g growth factors.

Moreover, non-proteolytically it is involved in cancer progression and migration by interacting with cell surface receptor protein-Integrins presumably via its disintegrin domain. These cell surface integrin proteins consist of α and β subunits that work as adhesive receptors for ECM and transmit signals into the cell. Studies have shown various subtypes of integrins that bind to ADAM9: ITGA2, ITGA3, ITGA6, ITGA6:ITGB4, ITGA6:ITGB4, ITGAV [62], which are also reported in our GSEA results for focal adhesion shown in Table ***4-3***. These integrin-mediated cell-ECM interactions at the cell surface are responsible for cell migration adhesion processes, ultimately triggering the signalling and remodelling of ECM components, including forming the cytoskeleton and focal adhesions [110].

Table 0.3: Ranked genes mapped on GSEA focal adhesion gene sets.

Dataset	**Genes Mapped**	**Gene**
GSE130078	20	ITGB1, ITGB6, ILK, CAV2, PDGFC, ITGA3, ACTB, ITGB4, BAD, PXN, DIAPH1, EGFR, VEGFC, LAMC2, BCAR1, VEGFB, MYL7, ITGA6, LAMA4, BRAF
GSE111011	7	DIAPH1, BCL2, ITGA5, PTEN, RAC2, PGF, CAV2
E-MTAB-4054	27	PARVB, CAPN2, FLNC, ITGA2B, PIK3R2, MYLPF, MAPK3, CRK, RAC2, LAMB3, PIK3R5, MAP2K1, LAMC1, LAMA4, RAPGEF1, LAMA2, PDGFRB, LAMB1, KDR, HGF, ITGA1, BRAF, ITGA5, SHC3, CAV1, FLT1, LAMC3

Table 0.3: ITGB: Integrin Subunit Beta (1,5,6), ITGA: Integrin Subunit Alpha (A3, A5, A6)

As Integrin-mediated cell adhesion to ECM is crucial, according to De Franceschi and his colleagues (2015), their fate is continually internalised, after which their fate is to either get degraded or recycled back to the plasma membrane. This integrin trafficking plays an essential role in regulating tumour cell migrations and adhesions [111]. To gain better insight on how ADAM9 regulates integrin function, Mygind and his colleagues (2018) performed a knockout

study that illustrated the role of ADAM9 in regulating integrin function and alteration of focal adhesion in cancer. Knock out of ADAM9 revealed the increased ITGB1/ β 1integrin levels and decreased adhesion and cell migration; this loss was conferred due to disrupted ADAM9-integrin interaction, which ultimately decreased the internalisation and degradation of ITGB1 and focal adhesion. Thus, ADAM9 is required for optimised ITGB1 endocytosis without accumulating at the cell surface and alters the focal adhesion by slowing down its maturation, leading to the perturbed cell adhesion and migration of tumour cells [62].

5 Conclusion

Recently many studies have proven the strong relationship between cancer and AS and its involvement in tumorigenesis and escaping cell death. Thus, AS plays an important role in cancers by triggering hallmarks of cancer from a progression of primary tumour cells(tumorigenesis) to metastasis of secondary tumour cells to distant organs [26].

Knowing the already available technologies to study gene and transcript expression, differential transcript usage technique (DTU) was preferred to identify ADAM9 isoform switch using the IsoformSwitchAnalyzeR package.

To identify isoform switching, the first transcriptome assembly was performed using a new tuxedo pipeline. Hisat 2 was used to align oesophageal samples (normal and tumour) to reference the human genome. Transcriptome assembly and reconstruction were performed using StringTie to identify transcripts and their expression levels accurately, and quantification files were generated using the Ballgown -e parameter. After the quantification, Dexseq was used for the statistical identification of isoform switches by calculating the differential usage (dIF) of L and S-forms of ADAM9 in all three datasets: for GSE130078 dIF for ENST00000379917(S) is -0.07, ENST00000487273(L) is 0.21, for GSE111011 dIF for ENST00000379917(S) is -0.06, ENST00000487273(L) is 0.12 and for E-MTAB-4054 dIF for ENST00000379917(S) is -0.17, ENST00000487273(L) is 0.19. Thus, the overall trend in the above plots represents significantly increased gene expression in oesophageal tumour samples compared to normal samples, whereas isoform expression and usage show that L isoform is majorly contributing to the ADAM9 gene expression in primary tumour samples compared to S-form. Afterwards, correlation analysis was performed to identify interacting partners for ADAM9. Gene set enrichment analysis was performed using GSEA. Focal adhesion was enriched in all three datasets. As non-proteolytically ADAM9 is involved in cancer progression and migration by interacting with cell surface receptor protein-Integrins presumably via its disintegrin domain, these integrins are present in these focal adhesion sites. They are involved in bi-directional transmembrane communication by connecting cell cytoskeletons to the extracellular

membrane matrix. Thus, these integrin-mediated cell-ECM interactions at the cell surface are responsible for cell migration adhesion processes, ultimately triggering the signalling and remodelling of ECM components, including forming the cytoskeleton and focal adhesions [110].

6 References

[1] T. A. Brown, “Understanding a Genome Sequence,” 2002, Accessed: Mar. 03, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21136/.

[2] A. Reyes and W. Huber, “Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues,” *Nucleic Acids Res.*, vol. 46, no. 2, pp. 582–592, Jan. 2018, doi: 10.1093/nar/gkx1165.

[3] I. Dunham *et al.*, “An integrated encyclopedia of DNA elements in the human genome,” *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[4] T. W. Nilsen and B. R. Graveley, “Expansion of the eukaryotic proteome by alternative splicing,” *Nature*, vol. 463, no. 7280. Nature, pp. 457–463, Jan. 28, 2010, doi: 10.1038/nature08909.

[5] D. Brett, H. Pospisil, J. Valcárcel, J. Reich, and P. Bork, “Alternative splicing and genome complexity,” *Nat. Genet.*, vol. 30, no. 1, pp. 29–30, 2002, doi: 10.1038/ng803.

[6] S. Larochelle, “Systems biology: Protein isoforms: More than meets the eye,” *Nature Methods*, vol. 13, no. 4. Nature Publishing Group, p. 291, Mar. 30, 2016, doi: 10.1038/nmeth.3828.

[7] E. T. Wang *et al.*, “Alternative isoform regulation in human tissue transcriptomes,” *Nature*, vol. 456, no. 7221, pp. 470–476, Nov. 2008, doi: 10.1038/nature07509.

[8] M. Dapas, M. Kandpal, Y. Bi, and R. V. Davuluri, “Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms,” *Brief. Bioinform.*, vol. 18, no. 2, p. bbw016, Mar. 2016, doi: 10.1093/bib/bbw016.

[9] C. Zhang, B. Zhang, M. S. Vincent, and S. Zhao, “Bioinformatics Tools for RNA-seq Gene and Isoform Quantification,” *J. Next Gener. Seq. Appl.*, vol. 03, no. 03, 2016, doi: 10.4172/2469-9853.1000140.

[10] K. Froussios, K. Mourão, G. Simpson, G. Barton, and N. Schurch, “Relative abundance of transcripts (RATs): Identifying differential isoform abundance from RNA-seq [version 1; referees: 1 approved, 2 approved with reservations],” *F1000Research*, vol. 8, p. 213, Feb. 2019, doi: 10.12688/f1000research.17916.1.

[11] C. Soneson, M. I. Love, and M. D. Robinson, “Differential analyses for RNA-seq:

Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]," *F1000Research*, vol. 4, p. 1521, Feb. 2016, doi: 10.12688/F1000RESEARCH.7563.2.

[12] K. Vitting-Seerup and A. Sandelin, "The landscape of isoform switches in human cancers," *Mol. Cancer Res.*, vol. 15, no. 9, pp. 1206–1220, Sep. 2017, doi: 10.1158/1541-7786.MCR-16-0459.

[13] V. C. Palve and T. R. Teni, "Association of anti-apoptotic Mcl-1L isoform expression with radioresistance of oral squamous carcinoma cells," *Radiat. Oncol.*, vol. 7, no. 1, Aug. 2012, doi: 10.1186/1748-717X-7-135.

[14] S. C.-W. Lee and O. Abdel-Wahab, "Therapeutic targeting of splicing in cancer," *Nat. Med. 2016 229*, vol. 22, no. 9, pp. 976–986, Sep. 2016, doi: 10.1038/nm.4165.

[15] G. A. Merino, A. Conesa, and E. A. Fernández, "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies," *Brief. Bioinform.*, vol. 20, no. 2, pp. 471–481, 2019, doi: 10.1093/bib/bbx122.

[16] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biol.*, vol. 14, no. 4, p. R36, Apr. 2013, doi: 10.1186/gb-2013-14-4-r36.

[17] C. Trapnell *et al.*, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat. Biotechnol. 2010 285*, vol. 28, no. 5, pp. 511–515, May 2010, doi: 10.1038/nbt.1621.

[18] D. Kim, B. Langmead, and S. L. Salzberg, "HISAT: A fast spliced aligner with low memory requirements," *Nat. Methods*, vol. 12, no. 4, pp. 357–360, Mar. 2015, doi: 10.1038/nmeth.3317.

[19] A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek, "Flexible isoform-level differential expression analysis with Ballgown," *bioRxiv*, p. 003665, Mar. 2014, doi: 10.1101/003665.

[20] S. Shen *et al.*, "rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data," *Proc. Natl. Acad. Sci.*, vol. 111, no. 51, pp. E5593–E5601, Dec. 2014, doi: 10.1073/PNAS.1419161111.

[21] K. Vitting-Seerup, B. T. Porse, A. Sandelin, and J. Waage, "SpliceR: An R package for classification of alternative splicing and prediction of coding potential from RNA-seq data," *BMC Bioinformatics*, vol. 15, no. 1, p. 81, Mar. 2014, doi: 10.1186/1471-2105-15-81.

[22] C. Soneson, K. L. Matthes, M. Nowicka, C. W. Law, and M. D. Robinson, “Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage,” *Genome Biol.*, vol. 17, no. 1, p. 12, Jan. 2016, doi: 10.1186/s13059-015-0862-3.

[23] L. Niu, W. Huang, D. M. Umbach, and L. Li, “IUTA: A tool for effectively detecting differential isoform usage from RNA-Seq data,” *BMC Genomics*, vol. 15, no. 1, Oct. 2014, doi: 10.1186/1471-2164-15-862.

[24] K. Vitting-Seerup and A. Sandelin, “IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences,” *Bioinformatics*, vol. 35, no. 21, pp. 4469–4471, Nov. 2019, doi: 10.1093/bioinformatics/btz247.

[25] E. Wang and I. Aifantis, “RNA Splicing and Cancer,” *Trends in Cancer*, vol. 6, no. 8. Cell Press, pp. 631–644, Aug. 01, 2020, doi: 10.1016/j.trecan.2020.04.011.

[26] F. Qi, Y. Li, X. Yang, Y. P. Wu, L. J. Lin, and X. M. Liu, “Significance of alternative splicing in cancer cells,” *Chinese medical journal*, vol. 133, no. 2. NLM (Medline), pp. 221–228, Jan. 20, 2020, doi: 10.1097/CM9.0000000000000542.

[27] R. Sciarrillo *et al.*, “The role of alternative splicing in cancer: From oncogenesis to drug resistance,” *Drug Resistance Updates*, vol. 53. Churchill Livingstone, p. 100728, Dec. 01, 2020, doi: 10.1016/j.drup.2020.100728.

[28] R. S. Holmes and T. L. Vaughan, “Epidemiology and Pathogenesis of Esophageal Cancer,” *Semin. Radiat. Oncol.*, vol. 17, no. 1, pp. 2–9, Jan. 2007, doi: 10.1016/j.semradonc.2006.09.003.

[29] F. Kamangar, W. H. Chow, C. C. Abnet, and S. M. Dawsey, “Environmental Causes of Esophageal Cancer,” *Gastroenterol. Clin. North Am.*, vol. 38, no. 1, pp. 27–57, 2009, doi: 10.1016/j.gtc.2009.01.004.

[30] J. C. Layke and P. P. Lopez, “Esophageal Cancer: A Review and Update,” Jun. 2006. Accessed: Feb. 03, 2021. [Online]. Available: www.aafp.org/afp.

[31] O. Shaheen, A. Ghibour, and B. Alsaid, “Esophageal Cancer Metastases to Unexpected Sites: A Systematic Review,” *Gastroenterology Research and Practice*, vol. 2017. Hindawi Limited, 2017, doi: 10.1155/2017/1657310.

[32] C. Lepage, B. Rachet, V. Jooste, J. Faivre, and M. P. Coleman, “Continuing rapid increase in esophageal adenocarcinoma in England and Wales,” *Am. J. Gastroenterol.*, vol. 103, no. 11, pp. 2694–2699, Nov. 2008, doi: 10.1111/j.1572-0241.2008.02191.x.

[33] O. N. Alema and B. Iva, “Cancer of the esophagus; histopathological sub-types in

northern Uganda," *Afr. Health Sci.*, vol. 14, no. 1, p. 17, 2014, doi: 10.4314/ahs.v14i1.4.

[34] K. J. Napier, "Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities," *World J. Gastrointest. Oncol.*, vol. 6, no. 5, p. 112, 2014, doi: 10.4251/wjgo.v6.i5.112.

[35] T. J. Kim, H. Y. Kim, K. W. Lee, and M. S. Kim, "Multimodality assessment of esophageal cancer: Preoperative staging and monitoring of response to therapy," *Radiographics*, vol. 29, no. 2, pp. 403–421, Mar. 2009, doi: 10.1148/rg.292085106.

[36] N. Giebeler and P. Zigrino, "A disintegrin and metalloprotease (ADAM): Historical overview of their functions," *Toxins*, vol. 8, no. 4. MDPI AG, Apr. 23, 2016, doi: 10.3390/toxins8040122.

[37] M. J. Duffy, E. McKiernan, N. O'Donovan, and P. M. McGowan, "Role of ADAMs in cancer formation and progression," *Clin. Cancer Res.*, vol. 15, no. 4, pp. 1140–1144, Feb. 2009, doi: 10.1158/1078-0432.CCR-08-1585.

[38] M. Mullooly, P. M. McGowan, J. Crown, and M. J. Duffy, "The ADAMs family of proteases as targets for the treatment of cancer," *Cancer Biology and Therapy*, vol. 17, no. 8. Taylor and Francis Inc., pp. 870–880, Aug. 02, 2016, doi: 10.1080/15384047.2016.1177684.

[39] J. S. M. Souza *et al.*, "The evolution of ADAM gene family in eukaryotes," *Genomics*, vol. 112, no. 5, pp. 3108–3116, Sep. 2020, doi: 10.1016/j.ygeno.2020.05.010.

[40] J. S. M. Souza *et al.*, "The evolution of ADAM gene family in eukaryotes," *Genomics*, vol. 112, no. 5, pp. 3108–3116, Sep. 2020, doi: 10.1016/j.ygeno.2020.05.010.

[41] D. F. Seals and S. A. Courtneidge, "The ADAMs family of metalloproteases: Multidomain proteins with multiple functions," *Genes and Development*, vol. 17, no. 1. Cold Spring Harbor Laboratory Press, pp. 7–30, Jan. 01, 2003, doi: 10.1101/gad.1039703.

[42] F. X. Gomiz-Rüth, "Catalytic domain architecture of metzincin metalloproteases," *Journal of Biological Chemistry*, vol. 284, no. 23. J Biol Chem, pp. 15353–15357, Jun. 05, 2009, doi: 10.1074/jbc.R800069200.

[43] D. R. Edwards, M. M. Handsley, and C. J. Pennington, "The ADAM metalloproteinases," *Molecular Aspects of Medicine*, vol. 29, no. 5. Mol Aspects Med, pp. 258–289, Oct. 2009, doi: 10.1016/j.mam.2008.08.001.

[44] L. Tian, X. Wu, C. Chi, M. Han, T. Xu, and Y. Zhuang, "ADAM10 is essential for

proteolytic activation of Notch during thymocyte development," *Int. Immunol.*, vol. 20, no. 9, pp. 1181–1187, Sep. 2008, doi: 10.1093/INTIMM/DXN076.

[45] K. Reiss and P. Saftig, "The 'A Disintegrin And Metalloprotease' (ADAM) family of sheddases: Physiological and cellular functions," *Seminars in Cell and Developmental Biology*, vol. 20, no. 2. Elsevier Ltd, pp. 126–137, 2009, doi: 10.1016/j.semcdb.2008.11.002.

[46] R. A. Black and J. M. White, "ADAMs: Focus on the protease domain," *Curr. Opin. Cell Biol.*, vol. 10, no. 5, pp. 654–659, Oct. 1998, doi: 10.1016/S0955-0674(98)80042-2.

[47] G. Murphy, "The ADAMs: Signalling scissors in the tumour microenvironment," *Nature Reviews Cancer*, vol. 8, no. 12. Nat Rev Cancer, pp. 929–941, Dec. 2008, doi: 10.1038/nrc2459.

[48] L. Przemyslaw, H. A. Oguslaw, S. Elzbieta, and S. M. Malgorzata, "ADAM and ADAMTS family proteins and their role in the colorectal cancer etiopathogenesis," *BMB Reports*, vol. 46, no. 3. Korean Society for Biochemistry and Molecular Biology, pp. 139–150, Mar. 2013, doi: 10.5483/BMBRep.2013.46.3.176.

[49] M. J. Duffy *et al.*, "The ADAMs family of proteases: New biomarkers and therapeutic targets for cancer?," *Clinical Proteomics*, vol. 8, no. 1. BioMed Central, p. 9, Dec. 09, 2011, doi: 10.1186/1559-0275-8-9.

[50] D. L. Kusindarta and H. Wihadmadyatami, "The Role of Extracellular Matrix in Tissue Regeneration," *Tissue Regen.*, 2018, doi: 10.5772/intechopen.75728.

[51] L. Peduto, "ADAM9 as a Potential Target Molecule in Cancer," *Curr. Pharm. Des.*, vol. 15, no. 20, pp. 2282–2287, 2009, doi: 10.2174/138161209788682415.

[52] L. Peduto, V. E. Reuter, D. R. Shaffer, H. I. Scher, and C. P. Blobel, "Critical function for ADAM9 in mouse prostate cancer," *Cancer Res.*, vol. 65, no. 20, pp. 9312–9319, Oct. 2005, doi: 10.1158/0008-5472.CAN-05-1063.

[53] C. P. Blobel, "ADAMs: Key components in egfr signalling and development," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 1. Nat Rev Mol Cell Biol, pp. 32–43, Jan. 2005, doi: 10.1038/nrm1548.

[54] V. Guaiquil, S. Swendeman, T. Yoshida, S. Chavala, P. A. Campochiaro, and C. P. Blobel, "ADAM9 Is Involved in Pathological Retinal Neovascularization," *Mol. Cell. Biol.*, vol. 29, no. 10, pp. 2694–2703, May 2009, doi: 10.1128/mcb.01460-08.

[55] M. A. Cissé, C. Sunyach, S. Lefranc-Jullien, R. Postina, B. Vincent, and F. Checler, "The disintegrin ADAM9 indirectly contributes to the physiological processing of cellular prion by modulating ADAM10 activity," *J. Biol. Chem.*, vol. 280, no. 49,

pp. 40624–40631, Dec. 2005, doi: 10.1074/jbc.M506069200.

[56] D. Nath, P. M. Slocombe, A. Webster, P. E. Stephens, A. J. P. Docherty, and G. Murphy, "Meltrin γ (ADAM-9) mediates cellular adhesion through α6β1 integrin, leading to a marked induction of fibroblast cell motility," *J. Cell Sci.*, vol. 113, no. 12, pp. 2319–2328, 2000.

[57] J. Mitschke, U. C. Burk, and T. Reinheckel, "The role of proteases in epithelial-to-mesenchymal cell transitions in cancer," doi: 10.1007/s10555-019-09808-2.

[58] C. Y. Lin *et al.*, "ADAM9 promotes lung cancer metastases to brain by a plasminogen activator-based pathway," *Cancer Res.*, vol. 74, no. 18, pp. 5229–5243, Jul. 2014, doi: 10.1158/0008-5472.CAN-13-2995.

[59] C. Y. Lin *et al.*, "ADAM9 promotes lung cancer progression through vascular remodeling by VEGFA, ANGPT2, and PLAT," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, Dec. 2017, doi: 10.1038/s41598-017-15159-1.

[60] C. M. Kossmann *et al.*, "ADAM9 expression promotes an aggressive lung adenocarcinoma phenotype," *Tumor Biol.*, vol. 39, no. 7, pp. 1–11, Jul. 2017, doi: 10.1177/1010428317716077.

[61] R. Liu *et al.*, "MicroRNA-425 promotes the development of lung adenocarcinoma via targeting A disintegrin and metalloproteinases 9 (ADAM9)," *Onco. Targets. Ther.*, vol. Volume 11, pp. 4065–4073, Jul. 2018, doi: 10.2147/OTT.S160871.

[62] K. J. Mygind, J. Schwarz, P. Sahgal, J. Ivaska, and M. Kveiborg, "Loss of ADAM9 expression impairs β1 integrin endocytosis, focal adhesion formation and cancer cell migration," *J. Cell Sci.*, vol. 131, no. 1, Jan. 2018, doi: 10.1242/jcs.205393.

[63] F. R. Fritzsche *et al.*, "ADAM9 Expression is a Significant and Independent Prognostic Marker of PSA Relapse in Prostate Cancer," *Eur. Urol.*, vol. 54, no. 5, pp. 1097–1108, Nov. 2008, doi: 10.1016/j.eururo.2007.11.034.

[64] Y. W. Lin *et al.*, "Stabilization of ADAM9 by N-α-acetyltransferase 10 protein contributes to promoting progression of androgen-independent prostate cancer," *Cell Death Dis.*, vol. 11, no. 7, pp. 1–16, Jul. 2020, doi: 10.1038/s41419-020-02786-2.

[65] Y. Dong *et al.*, "ADAM9 mediates the interleukin-6-induced Epithelial–Mesenchymal transition and metastasis through ROS production in hepatoma cells," *Cancer Lett.*, vol. 421, pp. 1–14, May 2018, doi: 10.1016/j.canlet.2018.02.010.

[66] K. Kohga *et al.*, "Sorafenib inhibits the shedding of major histocompatibility complex class I-related chain A on hepatocellular carcinoma cells by down-regulating a disintegrin and metalloproteinase 9," *Hepatology*, vol. 51, no. 4, pp.

1264–1273, Apr. 2010, doi: 10.1002/hep.23456.

[67] S. Oh *et al.*, “A Disintegrin and Metalloproteinase 9 (ADAM9) in Advanced Hepatocellular Carcinoma and Their Role as a Biomarker During Hepatocellular Carcinoma Immunotherapy,” *Cancers (Basel).*, vol. 12, no. 3, p. 745, Mar. 2020, doi: 10.3390/cancers12030745.

[68] C. O’Shea *et al.*, “Expression of ADAM-9 mRNA and protein in human breast cancer,” *Int. J. Cancer*, vol. 105, no. 6, pp. 754–761, Jul. 2003, doi: 10.1002/ijc.11161.

[69] J. J. Wang *et al.*, “Histone methyltransferase NSD2 mediates the survival and invasion of triple-negative breast cancer cells via stimulating ADAM9-EGFR-AKT signaling,” *Acta Pharmacol. Sin.*, vol. 40, no. 8, pp. 1067–1075, Aug. 2019, doi: 10.1038/s41401-018-0199-z.

[70] R. Grützmann *et al.*, “ADAM9 expression in pancreatic cancer is associated with tumour type and is a prognostic factor in ductal adenocarcinoma,” *Br. J. Cancer*, vol. 90, no. 5, pp. 1053–1058, Mar. 2004, doi: 10.1038/sj.bjc.6601645.

[71] J. G. M. Van Kampen *et al.*, “MiRNA-520f reverses epithelial-to-mesenchymal transition by targeting ADAM9 and TGFBR2,” *Cancer Res.*, vol. 77, no. 8, pp. 2008–2017, Apr. 2017, doi: 10.1158/0008-5472.CAN-16-2609.

[72] V. O. Oria *et al.*, “<scp>ADAM</scp> 9 contributes to vascular invasion in pancreatic ductal adenocarcinoma,” *Mol. Oncol.*, vol. 13, no. 2, pp. 456–479, Feb. 2019, doi: 10.1002/1878-0261.12426.

[73] C. Xing *et al.*, “Circular RNA ADAM9 facilitates the malignant behaviours of pancreatic cancer by sponging miR-217 and upregulating PRSS3 expression,” *Artif. Cells, Nanomedicine, Biotechnol.*, vol. 47, no. 1, pp. 3920–3928, Dec. 2019, doi: 10.1080/21691401.2019.1671856.

[74] X. Fan *et al.*, “ADAM9 Expression Is Associate with Glioma Tumor Grade and Histological Type, and Acts as a Prognostic Factor in Lower-Grade Gliomas,” *Int. J. Mol. Sci.*, vol. 17, no. 9, p. 1276, Aug. 2016, doi: 10.3390/ijms17091276.

[75] S. Sarkar, F. J. Zemp, D. Senger, S. M. Robbins, and V. W. Yong, “ADAM-9 is a novel mediator of tenascin-C-stimulated invasiveness of brain tumor–initiating cells,” *Neuro. Oncol.*, vol. 17, no. 8, pp. 1095–1105, Aug. 2015, doi: 10.1093/neuonc/nou362.

[76] T. Kauttu *et al.*, “Disintegrin and metalloproteinases (ADAMs) expression in gastroesophageal reflux disease and in esophageal adenocarcinoma,” *Clin. Transl. Oncol.*, vol. 19, 2094, doi: 10.1007/s12094-016-1503-3.

[77] J. L. Fry and A. Toker, “Secreted and membrane-bound isoforms of protease ADAM9 have opposing effects on breast cancer cell migration,” *Cancer Res.*, vol. 70, no. 20, pp. 8187–8198, 2010, doi: 10.1158/0008-5472.CAN-09-4231.

[78] N. Hotoda, H. Koike, N. Sasagawa, and S. Ishiura, “A secreted form of human ADAM9 has an α-secretase activity for APP,” *Biochem. Biophys. Res. Commun.*, vol. 293, no. 2, pp. 800–805, 2002, doi: 10.1016/S0006-291X(02)00302-9.

[79] A. Mazzocca *et al.*, “A secreted form of ADAM9 promotes carcinoma invasion through tumor-stromal interactions,” *Cancer Res.*, vol. 65, no. 11, pp. 4728–4738, 2005, doi: 10.1158/0008-5472.CAN-04-4449.

[80] R. Roy, S. Rodig, D. Bielenberg, D. Zurakowski, and M. A. Moses, “ADAM12 transmembrane and secreted isoforms promote breast tumor growth: A distinct role for ADAM12-S protein in tumor metastasis,” *J. Biol. Chem.*, vol. 286, no. 23, pp. 20758–20768, 2011, doi: 10.1074/jbc.M110.216036.

[81] K. C. Micocci *et al.*, “ADAM9 silencing inhibits breast tumor cell invasion in vitro,” *Biochimie*, vol. 95, no. 7, pp. 1371–1378, 2013, doi: 10.1016/j.biochi.2013.03.001.

[82] E. R, D. M, and L. AE, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002, doi: 10.1093/NAR/30.1.207.

[83] A. A *et al.*, “ArrayExpress update - from bulk to single-cell expression data.,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D711–D715, Jan. 2019, doi: 10.1093/NAR/GKY964.

[84] C. Kanz *et al.*, “The EMBL Nucleotide Sequence Database,” *Nucleic Acids Res.*, vol. 33, no. Database Issue, p. D29, Jan. 2005, doi: 10.1093/NAR/GKI098.

[85] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Res.*, vol. 38, no. 6, pp. 1767–1771, Dec. 2009, doi: 10.1093/nar/gkp1137.

[86] M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg, “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads,” *Nat. Biotechnol.*, vol. 33, no. 3, pp. 290–295, 2015, doi: 10.1038/nbt.3122.

[87] “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.” http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Jan. 31, 2021).

[88] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “Fastp: An ultra-fast all-in-one FASTQ

preprocessor," in *Bioinformatics*, Sep. 2018, vol. 34, no. 17, pp. i884–i890, doi: 10.1093/bioinformatics/bty560.

[89] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/BIOINFORMATICS/BTU170.

[90] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, May 2011, doi: 10.14806/EJ.17.1.200.

[91] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown," *Nat. Protoc.*, vol. 11, no. 9, pp. 1650–1667, Sep. 2016, doi: 10.1038/nprot.2016.095.

[92] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[93] M. Pertea and G. Pertea, "GFF Utilities: GffRead and GffCompare," *F1000Research*, vol. 9, 2020, doi: 10.12688/f1000research.23297.1.

[94] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biol.*, vol. 11, no. 3, p. R25, Mar. 2010, doi: 10.1186/gb-2010-11-3-r25.

[95] W. Shen, S. Le, Y. Li, and F. Hu, "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation," *PLoS One*, vol. 11, no. 10, p. e0163962, Oct. 2016, doi: 10.1371/journal.pone.0163962.

[96] Y. J. Kang *et al.*, "CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W12–W16, Jul. 2017, doi: 10.1093/nar/gkx428.

[97] M. Punta *et al.*, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, Jan. 2012, doi: 10.1093/nar/gkr1065.

[98] M. M. Babu, "The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease," *Biochemical Society Transactions*, vol. 44, no. 5. Portland Press Ltd, pp. 1185–1200, Oct. 15, 2016, doi: 10.1042/BST20160172.

[99] B. Mészáros, G. Erdös, and Z. Dosztányi, "IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W329–W337, Jul. 2018, doi: 10.1093/nar/gky384.

[100] J. J. Almagro Armenteros *et al.*, “SignalP 5.0 improves signal peptide predictions using deep neural networks,” *Nat. Biotechnol.*, vol. 37, no. 4, pp. 420–423, Apr. 2019, doi: 10.1038/s41587-019-0036-z.

[101] M. I. Love, S. Anders, V. Kim, and W. Huber, “RNA-Seq workflow: gene-level exploratory analysis and differential expression,” *F1000Research*, vol. 4, p. 1070, Oct. 2015, doi: 10.12688/f1000research.7035.1.

[102] B. Giardine *et al.*, “Galaxy: A platform for interactive large-scale genome analysis,” *Genome Res.*, vol. 15, no. 10, pp. 1451–1455, Oct. 2005, doi: 10.1101/gr.4086505.

[103] C. M. Koch *et al.*, “A beginner’s guide to analysis of RNA sequencing data,” *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 2. American Thoracic Society, pp. 145–157, Aug. 01, 2018, doi: 10.1165/rcmb.2017-0430TR.

[104] H. Climente-González, E. Porta-Pardo, A. Godzik, and E. Eyras, “The Functional Impact of Alternative Splicing in Cancer,” *Cell Rep.*, vol. 20, no. 9, pp. 2215–2226, 2017, doi: 10.1016/j.celrep.2017.08.012.

[105] M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, and S. L. Salzberg, “RNA-seq experiments with HISAT , StringTie and Ballgown,” *Nat. Protoc.*, vol. 11, no. 9, pp. 1650–1667, 2016, doi: 10.1038/nprot.2016-095.

[106] J. M. W. Gee and J. M. Knowlden, “ADAM metalloproteases and EGFR signalling,” *Breast Cancer Res.*, vol. 5, no. 5, pp. 223–224, 2003, doi: 10.1186/bcr637.

[107] J. A. R. Jonathan Posner and Bradley S. Peterson, “基因的改变NIH Public Access,” *Bone*, vol. 23, no. 1, pp. 1–7, 2008, doi: 10.1158/1535-7163.MCT-13-1001.The.

[108] I. Michalopoulos *et al.*, “Human gene correlation analysis (HGCA): A tool for the identification of transcriptionally co-expressed genes,” *BMC Res. Notes 2012 51*, vol. 5, no. 1, pp. 1–11, Jun. 2012, doi: 10.1186/1756-0500-5-265.

[109] A. Huttenlocher and A. R. Horwitz, “Integrins in cell migration,” *Cold Spring Harb. Perspect. Biol.*, vol. 3, no. 9, pp. 1–16, 2011, doi: 10.1101/cshperspect.a005074.

[110] M. Maziveyi and S. K. Alahari, “Cell matrix adhesions in cancer: The proteins that form the glue,” *Oncotarget*, vol. 8, no. 29, pp. 48471–48487, 2017, doi: 10.18632/oncotarget.17265.

[111] N. de Franceschi, H. Hamidi, J. Alanko, P. Sahgal, and J. Ivaska, “Integrin traffic-the update,” *J. Cell Sci.*, vol. 128, no. 5, pp. 839–852, 2015, doi: 10.1242/jcs.161653.

7 Appendix

7.1 Appendix A – Alignment tables

7.1.1 GS130078

Sample (Normal)	Alignment score %
SRR8931987	96%
SRR8931989	95.74%
SRR8931991	94.42%
SRR8931993	95.18%
SRR8931995	95.27%
SRR8931997	95.42%
SRR8931999	95.50%
SRR8932001	95.03%
SRR8932003	95.01%
SRR8932005	95.85%
SRR8932007	95.00%
SRR8932009	94.31%
SRR8932011	95.36%
SRR10173245	98.16%
SRR10173247	98.30%
SRR10173249	97.56%
SRR10173251	97.77%
SRR10173253	97.41%
SRR10173255	95.43%
SRR10173257	96.40%
SRR10173259	96.92%
SRR10173261	95.97%
SRR10173263	97.09%

Sample (Tumour)	**Alignment score %**
SRR8931988	95.25%
SRR8931990	95.11%
SRR8931992	93.81%
SRR8931994	94.95%
SRR8931996	95.22%
SRR8931998	93.00%
SRR8932000	94.17%
SRR8932002	93.58%
SRR8932004	95.53%
SRR8932006	94.88%
SRR8932008	93.70%
SRR8932010	95.77%
SRR8932012	95.94%
SRR10173246	96.56%
SRR10173248	94.24%
SRR10173250	94.65%
SRR10173252	94.29%
SRR10173254	97.84%
SRR10173256	96.63%
SRR10173258	97.71%
SRR10173260	95.09%
SRR10173262	98.03%
SRR10173264	95.18%

7.1.2 GS111011

Sample (Normal)	Alignment score %
SRR6762723	94.32%
SRR6762724	95.99%
SRR6762725	95.25%
SRR6762726	93.61%
SRR6762727	95.71%
SRR6762728	96.01%
SRR6762729	94.77%

Sample (Tumour)	Alignment score %
SRR6762730	82.37%
SRR6762731	93.61%
SSRR6762732	94.63%
SRR6762733	95.47%
SRR6762734	95.74%
SRR6762735	95.31%
SRR6762736	94.80%

7.1.3 E-MTAB4054

Sample (Normal)	Alignment score %
ERR1141723	96.69%
ERR1141724	96.92%
ERR1141725	93.37%
ERR1141726	98.94%
ERR1141727	94.06%
ERR1141728	93.81%

ERR1141729	93.78%
ERR1141730	98.63%
ERR1141731	96.66%
ERR1141732	96.58%
ERR1141733	97.02%

Sample (Tumour)	**Alignment score %**
ERR1141704	98.31%
ERR1141705	94.03%
ERR1141706	95.70%
ERR1141707	97.08%
ERR1141708	92.70%
ERR1141709	98.40%
ERR1141710	98.23%
ERR1141711	93.60%
ERR1141712	97.37%
ERR1141713	95.63%
ERR1141714	97.86%
ERR1141715	93.06%
ERR1141716	96.57%
ERR1141717	98.32%
ERR1141718	93.34%
ERR1141719	95.67%