Exploring the Transcriptional Landscape of Epithelial Tumors in order to Decipher the Role of Nonsense-mediated mRNA Decay in Cancer

By

Bareera Bibi

(Spring-2020-MSBI-5-00000327795)

School of Interdisciplinary Engineering and Sciences

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Exploring the Transcriptional Landscape of Epithelial Tumors in order to Decipher the Role of Nonsense-mediated mRNA Decay in Cancer

By

Bareera Bibi

(Spring-2020-MSBI-5-00000327795)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Bioinformatics

Supervisor: Dr. Mehak Rafiq

School of Interdisciplinary Engineering and Sciences

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

Annex A to NUST Letter
0972/102/Exams/Thesis
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms BAREERA BIBI Registration No. 00000327795 of School of Interdisciplinary Engineering & Sciences (SINES) has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp:

DR. MEHAK RAFIQ
Assistant Professor
SINES, National University
of Science & Technology
H-12 Islamabad

Name of Supervisor: Dr Mehak Rafiq

Date: 24/6/24

Signature of HoD with stamp:

Dr. Fouzia Malik
HoD Sciences
Professor
SINES NUST Sector H-12
Islamabad

Date: 25-4-2024

Countersign by

Signature (Dean/Principal):

Dr. SYED IRTIZA ALI SHAH
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Date: 25 JUN 2024

Dedication

This thesis is dedicated to my late father for his enduring support and inspiration, and to my beloved mother for her boundless love and encouragement.

Acknowledgments

Praise be to Allah (S.W.T), the Creator and the Sustainer of the Universe, who blessed me with knowledge and guided me through the intricate paths of my educational journey.

I extend my heartfelt gratitude to my mentor, the compassionate Assistant Professor Dr. Mehak Rafiq, whose unwavering support and insightful criticism became the compass guiding my research odyssey.

Deep appreciation goes to my esteemed examination committee members, Dr. Rehan Zafar Paracha and Dr. Zamir Hussain, for their invaluable guidance, assistance, and encouragement during my research.

I am thankful for the forward-thinking leadership at NUST, from the esteemed Rector to the dedicated Principal of SINES and the HOD of SINES. Your commitment to academic excellence has shaped my educational experience.

To my family, and especially my beloved mother, Saeeda Bibi, your boundless love, tireless care, and both financial and moral support have been my pillars of strength. I owe my success to you.

And to my dear friends, thank you for being a part of this profound chapter of my life. May joy and success accompany you in all your future endeavours.

Bareera Bibi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NMD	Nonsense-mediated mRNA decay
AS-NMD	Alternative Splicing coupled Nonsense-mediated mRNA Decay
mRNA	Messenger Ribonucleic Acid
DTU	Differential transcript Usage
DTE	Differential Transcript Expression
DGE	Differential Gene Expression
GSEA	Gene Set Enrichment Analysis
IF	Isoform Fraction
dIF	Differential Isoform Usage
GO	Gene Oncology
ISAR	IsoformSwitchAnalyzeR
WGCNA	Weighted Gene Co-expression Network Analysis
TSG	Tumor Suppressor Gene

ABSTRACT

Nonsense-mediated mRNA decay (NMD) plays a crucial role in normal physiological processes by co-regulating the expression of a wide group of genes including cancer-related pathways. Although the role of NMD in cancer is integral, its exact contribution is often regarded as paradoxical with NMD playing both tumor-promoting and tumor-suppressing roles. The development of more effective cancer treatment strategies demands a better understanding of the impact of NMD in cancer. The current study explores the role of NMD in the context of cancer by analysing the functional outcomes of alternative splicing coupled NMD in 10 epithelial cancers. It makes use of isoform switching analysis and weighted gene co-expression network analysis to identify a set of pan-cancer-wide switching hub genes. The study reports 6 NMD-sensitive hub genes across multiple carcinomas, with 4 genes being directly or indirectly involved in tumorigenesis. Notably, the findings support an anti-tumorigenic role of NMD as evidenced by the downregulation of potentially oncogenic isoforms of TIMM17B and FAM136A genes as well as increased isoform usage of potentially tumor-suppressing ZNRD1 gene. However, the results also report the upregulation in the isoform usage of protein-coding isoforms of CDK20, a potentially oncogenic gene. It can be concluded that cancer genes possess the capacity to manipulate NMD processes, thereby contributing to tumor progression. Consistent with the prevailing view, this research substantiates both pro- and anti-tumorigenic functions of NMD in cancer. Nevertheless, it broadens our understanding of the nuanced interplay between NMD and cancer genes.

Keywords: Cancer, NMD, Isoform, gene.

CHAPTER 1: INTRODUCTION

1.1 Alternative Splicing

The central dogma of molecular biology has laid down the pivotal foundation of the flow of genetic information from DNA to RNA to a functional product called Protein. The DNA contains the necessary information required to make all the proteins. RNA is a messenger molecule that carries this information to cellular machinery through transcription and decodes it into a functional product called proteins via translation [1]. Post-transcriptional modifications following protein biosynthesis are biochemical changes that are crucial for its homology and function. Alternative splicing is an essential post-transcriptional phenomenon that allows a single gene to produce various transcripts, also known as isoforms, that may have distinct, similar, or opposing functions, thereby increasing proteome diversity [2]. In the cell, the RNA quality control mechanism has evolved at various levels of gene regulation to ensure the removal of mis-spliced or aberrant alternatively transcribed transcripts isoforms [3]. Notably, alternative splicing can degrade mRNA by a phenomenon called nonsense-mediated decay (NMD) possessing premature termination codons (PTCs), which are detrimental to cell survival [4].

1.1.1 Alternative Splicing Events

Different spliced events in eukaryotes are characterised by the rearrangement or skipping of exons and portions of introns (junctions) as shown in the Figure 1.1. Exon skipping or cassette exon is when the exon is spliced out of the nascent or primary transcript and is the most common type of event in mammals [5]. Kim (2007) compared expressed sequence tag (EST) data of eight eukaryotic species and revealed that the frequency of exon skipping phenomenon grows moderately from invertebrates to vertebrates [6]. Mutually exclusive exons are characterised

by splicing one out of the two exons while the other retains [7]. An alternative donor site is an event in which an alternative 5' splice junction (donor site) is used that changes the 3' boundary of the upstream exon. Whereas, in the Alternative acceptor site event, an alternative 3' splice junction (acceptor site) is used that changes the 5' boundary of the downstream exon. Intron retention is an important event mainly occurring in plants in which part of the intron is not spliced out correctly and remains in the final mRNA [8].

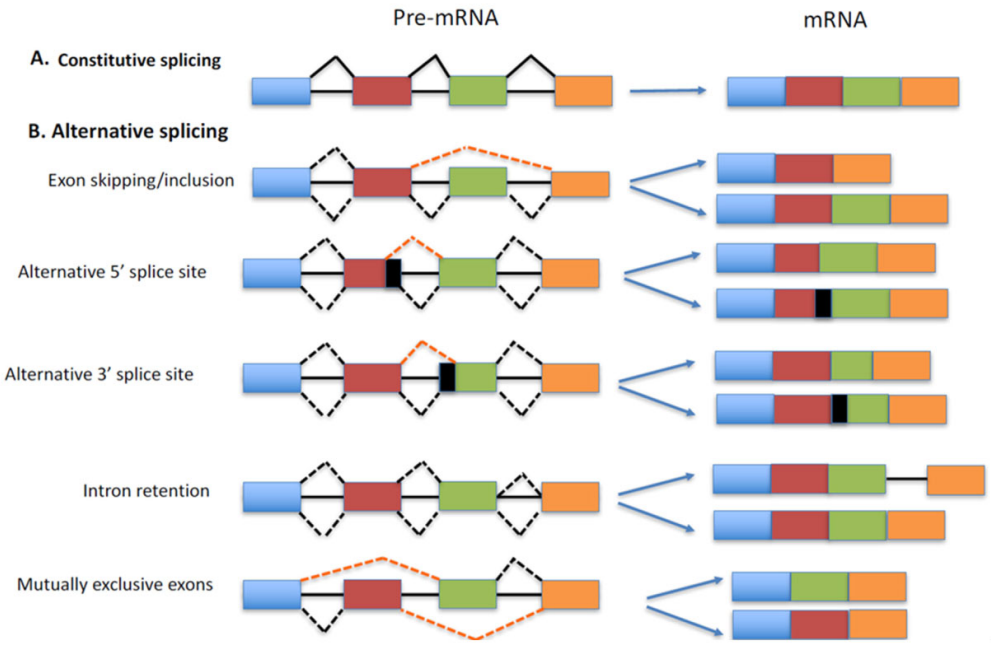

Figure 1.1: Schematic Representation of different types of splicing events [9]

These events give rise to transcripts that carry different biological functions that vary in their protein-protein interactions, sub-cellular localisation, or even catalytic behaviour. More than 90% of the eukaryotic genes in mammals generate multiple isoforms, and aberrant splicing has become the cause of many diseases in humans [10]. According to the latest statistics by GENCODE release 25, there are approximately 60,000 annotated genes, out of which 20,000 are protein-coding genes and 144,0000 transcripts. This statistic suggests seven transcripts per protein-coding gene.

1.1.2 Significance of Alternative Splicing Analysis

Alternative splicing significantly affects tissue specificity, developmental stages, disease conditions and cellular processes taking place in the human body. Dysregulation of alternative splicing can cause muscle dysfunction, neurodegenerative disorders, auto immune diseases including cancers; this emphasises the need to analyse the expression at isoform level [11]. To date, several studies have demonstrated gene-level expression at RNA sequencing (RNA-seq) and microarray platforms. However, less is known about isoform expression and its abundance estimation using these platforms. RNA seq, exon arrays, and reverse transcription-quantitative polymerase chain reaction (RT-qPCR) have been developed that adopt unique approaches to evaluate and estimate isoform expression. In the past, tiling arrays and exon junction arrays using oligonucleotide probe sets have been proposed to study splicing events [12]. Nevertheless, with the advent of high-throughput sequencing, RNA seq became the choice of researchers to study alternative splicing. An advanced form of exon microarrays, junction arrays, can work as an alternative to RNA-Seq for the detection of splice events with equal accuracy when it comes to well-described regions of trancriptome. Nevertheless, RNA-Seq is more flexible and stronger specifically for the detection of novel transcripts splicing events [13].

1.1.3 Computational Analysis of Alternative Splicing

Alternative splicing has been studied by characterising the analyses into two main categories: 1) when the subject of analysis is transcripts/isoforms abundance estimation/quantification and its relative or absolute expression (isoform and exon based tools are commonly used). 2) when the analysis is focused primarily on splicing events such as cassette exons, alternative 5‘ splice site, mutually exclusive exons, intron retention and alternative termination sites.

Various computational tools (as shown in Table 1.1) are used for differential splicing analysis [14].

Table 1.1: Overview of the Differential Splicing analysis approaches and tools [14]

Approach	**Tools**
Isoform-based	Cufflinks/cuffdiff2, DiffSplice, Ballgown
Exon-based	DEXSeq, EdgeR, Limma, JunctionSeq, DRIMSeq
Event-based	dSpliceType, MAJIQ, rMATS, SUPPA

1.2 Computational Analysis of Transcript Isoforms

The investigation of aberrant splicing events and isoform quantification has been shaped not only by the development of the techniques but also by the statistical approaches and algorithms being designed for their authentic biological interpretations. With the advent of the alternative splicing (AS) analysis tools, it became necessary to categorise how existing gene-level expression can be differentiated from transcript-level expression. Therefore, researchers termed gene-level expression analysis as a conventional 'union exon based' approach while transcript level expression analysis as a 'transcript-based approach' [15]. The transcript level dynamics [16] among isoforms may not be observed and detected with gene-level analysis because of the following reasons: Cancellation (the abundance of transcripts changing in the opposite direction cancels out upon conversion to gene abundance), Dominance (change in the abundance of the minor transcript can be masked by the abundance of the dominant transcript, also known as the canonical transcript) or, Collapsing (different transcripts of a gene with minute effect size or abundance in the same direction do not cause significant change when observed after summation for differential gene expression but their independent changes, relative or absolute, constitute substantial evidence for differential expression at isoform level).

These dynamic changes are necessary to study at the transcript level, crucial for gene regulation, and contribute to the physiological mechanism leading to any phenotypic change. The transcript-based approach has been widely studied for statistical analysis by differentiating between relative and absolute isoform expression analysis. Change in the absolute abundance of individual transcripts isoforms

across conditions is referred to as differential transcript expression (DTE). In contrast, change in the relative abundance of individual transcripts isoforms across conditions is known as differential transcript usage (DTU) [17]. The difference between differential gene expression (DGE), differential transcript expression (DTE) and differential transcript usage (DTU) has been illustrated in Figure 1.2.

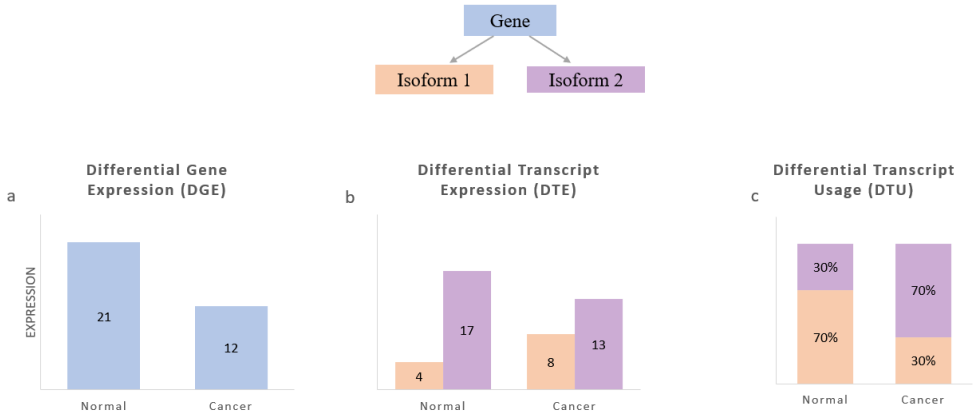

Figure 1.2: Graphical representation of the three types of differential expression analysis (DGE, DTE and DTU)

The Figure 1.2 represents differential biological expression across changing biological conditions. For instance, if a gene has two isoforms where isoform 1 is the canonical or primary isoform while isoform 2 is the secondary isoform then the sum of the expression of individual isoforms will be termed as gene expression. An overall change in the gene expression or sum of all individual changes to transcript isoforms of that gene is known as differential gene expression (DGE). As shown in Figure 1.2(a), expression of the gene is significantly higher in Normal than Cancer. In other words, the gene is down-regulated in cancer. In 1.2(b), the absolute abundance or expression of individual isoforms of the gene increases or decreases in the individual condition representing DTE. In 1.2(c) The contribution of the individual transcript to the overall gene expression across conditions changes. This change in the relative abundance of the individual transcript isoforms is known as differential transcript usage (DTU) and represents isoform switching events.

1.3 Isoform Switching in Cancer

The importance of analysing isoforms instead of genes has been highlighted by many examples showing functionally essential changes in the body [17]. Typically, coding genes have a transcript isoform expressed significantly higher than other alternatively spliced transcript isoforms, often known as canonical isoforms. Under unfavourable circumstances like disease states, the dominance completely shifts from canonical to the other alternative transcript isoforms. One splicing phenotype in cancers can be isoform switching [11]. Isoform switching is the relative abundance of different isoforms of the same gene that is significantly different in different cell types or when disease cells are compared to normal as shown in Figure 1.2(c).

Cancer is one of the most extensively reported diseases where isoform switching and modification of splicing patterns has been observed. Isoform switches extensively influence protein domains from protein families which are frequently found to be mutated in cancer. In addition to that, some of these switches often act as cancer drivers or modulate or modify the interaction network of cancer-driving proteins. Consequently, isoform switches could potentially contribute to tumor progression either independently or in conjunction with mutated cancer-driver genes [18].

1.4 Nonsense-mediated mRNA decay

Nonsense mutations or PTCs account for a significant(10-30%) fraction of genetic diseases including cancer, making them potential therapeutic targets for suppression therapies [19, 20]. Eukaryotic cell has developed some proofreading mechanisms which maintain the integrity of transcriptome by downregulating abnormal transcripts. Nonsense-mediated mRNA decay (NMD) is one such surveillance pathway that helps regulate mRNA quality by degrading aberrant transcripts that harbour NMD-stimulating factors, specifically premature termination codons (PTCs) [21]. Besides post-transcriptional surveillance, NMD has also shown to target a significant quantity of normal transcripts playing an active role in gene

expression regulation pertaining to cellular requirements [22]. This suggest that NMD controls not only quality but also quantity of transcriptome regulating a wide variety of genes.

1.4.1 NMD Substrates

Before initiating the degradation of transcript, NMD must distinguish between the potential NMD-targeted transcripts from other normal transcripts by identifying NMD-activating factors present on the transcript. As a surveillance pathway, NMD targets PTC-harbouring aberrant transcripts containing genetic mutations, aberrant splicing or inaccurate transcription. As a gene expression fine-tuning mechanism, NMD degrades normal transcripts commonly marked by the presence of NMD-stimulating factors (as shown in the Figure 1.4) including unusually long 3' untranslated region (UTR), upstream open reading frame (uORF) and presence of intron in the 3' UTR [22]. In addition to these factors, NMD can also be triggered by programmed ribosomal frameshifting in yeast and coupling mechanism of NMD and alternative splicing in mammals [23, 24].

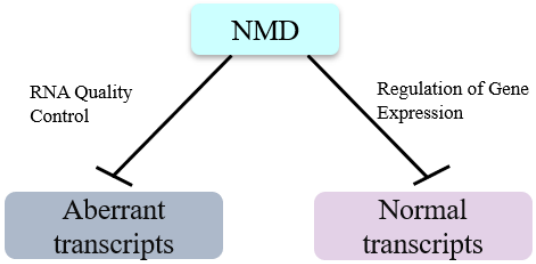

Figure 1.3: Role of NMD as Quality Control and Gene Expression Regulation Pathway

1.4.2 Mechanism of NMD Pathway

Nonsense-mediated mRNA decay works by identifying and flagging NMD targets leading to degradation of such transcripts by activating enzyme complexes. In order to explain the identification of PTC by NMD, two major models have been proposed: 3' UTR exon junction complex (EJC) dependent NMD and 3' UTR EJC-independent NMD. Among mammals, the EJC dependent model relies on the mammalian '50-55' nucleotides rule. Some NMD targets violate the 50-55 nts

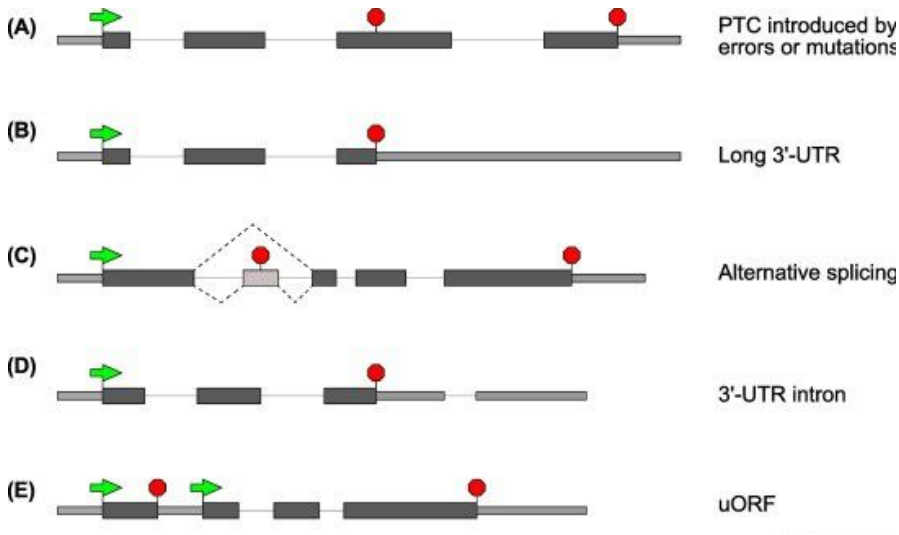

Figure 1.4: NMD factors: Light-shaded thinner rectangles are UTRs, dark-shaded thicker rectangles represent exons while black lines represent introns. Green arrows denote start codons, and red stop signs indicate stop codons. [25]

rule and are degraded by EJC-independent NMD. As the molecular mechanisms involved in EJC-independent NMD are not clear, the EJC-dependent model and the '50-55 nts' rule is used in bioinformatics to predict NMD-sensitivity of an mRNA transcript [26].

EJC-dependent NMD depends on the presence of EJC such that whenever a PTC is located more than 50-55 nucleotides upstream of at least one exon–exon junction in an mRNA, the transcript becomes a target of NMD. This rule, referred to as "50-55 nucleotide rule", is the dominant approach for NMD-target identification in mammals [27]. Under normal circumstances, EJC is displaced by ribosomes during translation and mRNA transcript is converted into a functional protein. On the contrary, PTC-harbouring transcripts retain EJC 50-55 bases downstream of stop codon resulting in the activation of NMD marked by the formation of SURF complex [28]. As illustrated in Figure 1.5, the NMD surveillance complex called SURF (SMG1–UPF1-eRF1–eRF3) complex is formed by the interaction of kinase protein SMG1, UPF1 helicase and eukaryotic polypeptide release factors eRF1 and eRF3 [29]. Eventually, SURF complex binds with EJC complex and interacts with UPF2 and UPF3b proteins forming so-called decay-inducing DECID complex which stimulates SMG1-mediated phosphorylation of UPF1 pro-

tein [28]. Activation of UPF1 protein results in the recruitment of SMG5, SMG6 and SMG7 proteins which stimulate decapping and deadenylation of the mRNA transcript exposing it to nucleases and other degrading factors [30, 31, 32, 33].

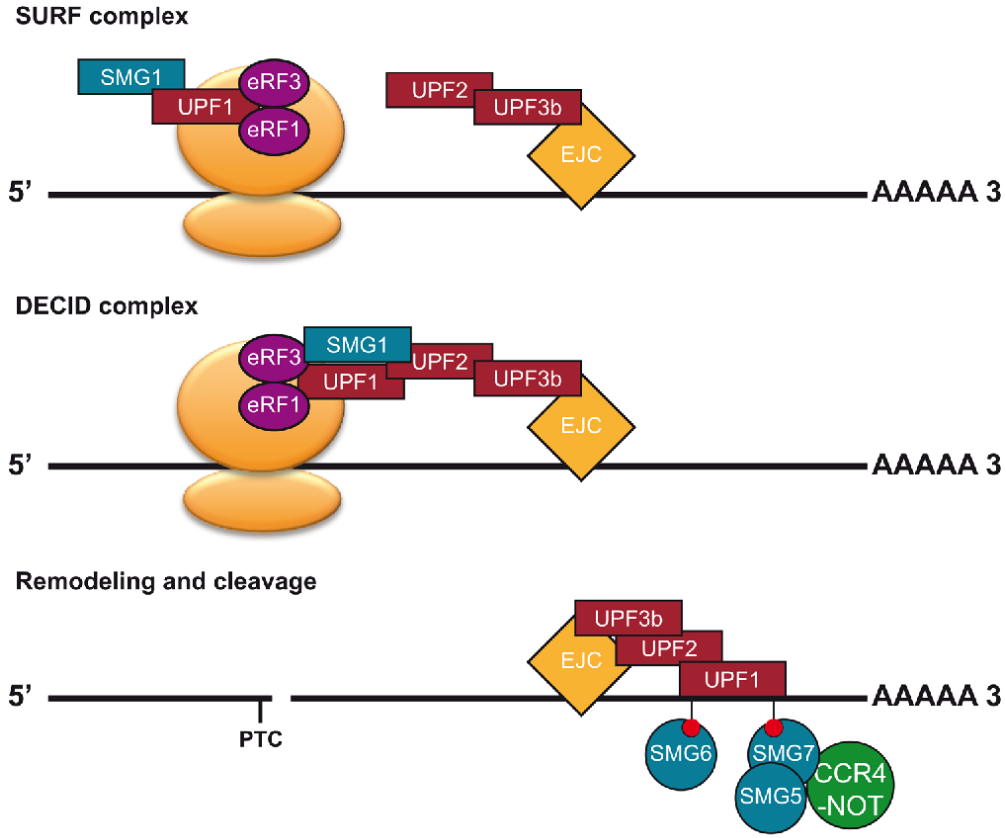

Figure 1.5: Mechanism of EJC dependent NMD pathway in mammalian cells [34]

Variations of this classical NMD pathway exist among different organisms where NMD has been perceived as a "branched network" among vertebrates involving a variety of factors that can contribute towards UPF1 activation. Thus, NMD is probably manifested in the form of multiple routes depending upon the properties of transcript, cell, tissue, or physiological condition. Moreover, inhibition of a specific branch or NMD route may potentially act as a therapeutic avenue towards treating the given disease [25, 35].

1.4.3 Regulation of NMD

Nonsense-mediated mRNA decay has been reported to be regulated in a cell and tissue-specific manner with different cell and tissue types representing different potentials for eliciting NMD. Similarly, NMD activity and NMD targets also vary

across different conditions hinting towards dependence of NMD efficiency on cell and tissue type. For instance, NMD is more efficient in HeLa cells as compared to MCF7 cells [36]. Likewise, differential NMD efficiency has been observed in mice across varying tissue types [37]. Moreover, NMD activity is also regulated developmental stage-wise in different eukaryotes. Nevertheless, reviewers also point toward the epigenetic basis of NMD regulation. A better understanding of the regulatory mechanism behind NMD and the study of NMD homeostasis is crucial for elucidating the role of NMD across different biological processes and conditions [35].

The molecular mechanism behind NMD regulation under such contexts is not yet clear, however, based on some pieces of evidence it has been speculated that NMD activity might be regulated by differential expression of NMD factors or varying translation rate [35]. Generally, NMD activity is influenced by genetic background and quantity of nonsense transcripts in the cell [38]. Evidence also suggests the presence of feedback regulation in the NMD pathway directed by different branches of NMD operating in a tissue-specific manner [39]. Autoregulatory unproductive splicing operates in a plethora of splicing factors and similar elements involved in splicing and post-transcriptional modifications. Splicing factor proteins, such as SC35 and polypyrimidine tract binding protein (PTB), autoregulate their own production by inhibiting splicing of their transcripts and creating a negative feedback loop [40, 41, 42]. In this way, such autoregulated splicing factors not only maintain their concentration but also regulate splicing of other pre-mRNAs [43].

1.5 Alternative Splicing and NMD

NMD-sensitive transcripts can be produced by various mechanisms in the cell with mRNA splicing being the most extensively studied cause. Alternative splicing (AS) is an efficient mechanism for achieving proteome-level diversity and complexity among eukaryotes. However, AS is also a lead source of aberrant mRNAs with around one-third of all AS events leading to PTCs which are then targeted by NMD [44]. Intron retention in 3' UTR, poison cassette exons, and exon skip-

ping are some of the most frequent AS events that generate NMD-triggering isoforms [45].

Although mechanisms related to the generation and degradation of PTC^+ transcripts have been outlined, the exact purpose of their abundant generation is still unclear. They might get produced as a result of splicing errors or "side effects" of a particular gene. However, most explanations are somehow connected with the coupling of mRNA splicing and NMD. AS-NMD coupling, known as Regulated Unproductive Splicing and Translation (RUST), acts as an additional regulatory mechanism that can down-regulate expression of a certain gene in order to meet requirements of the cell [45, 46].

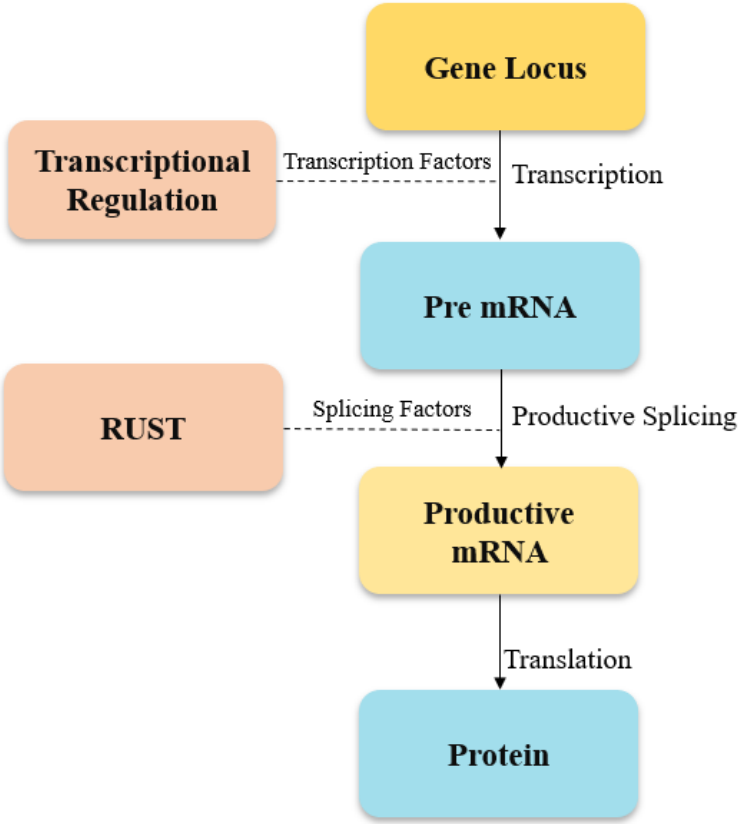

Figure 1.6: Regulated Unproductive Splicing and Translation (RUST)

1.5.1 Mechanism of action of AS-NMD

During transcription, transcriptional regulation regulates the amount of transcription products with the help of transcription factors. However, once the pre-mRNA has been produced, the transcriptional regulation is no longer enough and there

must be some kind of additional regulatory mechanism to maintain the quantity of the productive mRNA which is to be translated into protein. Gene regulation through the coupled action of alternative splicing and nonsense-mediated mRNA decay (AS-NMD), often referred to as RUST, comes in handy at this point. Splicing factors work analogous to transcription factors by altering the abundance of productive isoforms and eventually directing the overall gene expression. They regulate the production of functional or primary isoforms relative to NMD-targeted aberrant isoforms by choosing alternative splice sites. For instance, the inclusion or exclusion of an alternative exon can give rise to frameshift and PTC+ transcripts. Likewise, the exclusion of intron from 3' UTR can produce an NMD-sensitive transcript. Consequently, change in splicing patterns alters the ratio of productive to NMD-targeted isoforms and therefore the rate of protein production [43, 46]. Polypyrimidine tract binding protein (PTB) is one of the commonly studied genes regulated by RUST. Higher PTB concentration is balanced by NMD targeting of PTB transcripts. In other words, PTC is introduced in PTB protein coding transcripts leading to their degradation by NMD. On the contrary, decreased proportion of PTB are compensated by reducing the degradation levels [42, 43].

1.5.2 Significance of AS-NMD

Regulated unproductive splicing is a widespread regulatory mechanism found to be conserved between kingdoms, ranging from yeast and plants to higher animals like humans [47]. Coupled action of NMD and alternative splicing provides a cost-effective regulatory approach by preventing wasted translation. Although both mechanisms can play their role as regulatory pathways individually, their coupled action presents additional benefits. For instance, alternative splicing can effectively regulate gene expression by producing non-functional proteins with the help of splicing factors, however, NMD prevents accumulation of potentially harmful truncated proteins [43].

Generally, RUST acts as an additional fine-tuning mechanism in the cell. It regulates expression of a wide variety of genes including RNA-binding proteins while allowing the autoregulation of various splicing factors and temporal control

of larger genes [42, 43]. It also enables and directs tissue-specific and development-dependent production of proteins such as MID1 [48].

It has been widely proposed that perturbing RUST may lead to the predisposition of diseases [45]. One of the very recent reports suggests a strong connection between dysregulation of RUST and failure of pancreatic cells in diabetes [49]. Moreover, it is also noteworthy that one of the most common groups of genes regulated by AS-NMD is cancer-related genes (i.e., oncogenes and tumor suppressor genes) pointing towards the crucial role of AS-NMD in cancer.

1.6 Role of NMD in Cancer

Besides maintaining the quality of mRNA transcripts, NMD plays a crucial role in normal physiological processes by co-regulating the expression of a wide group of genes including cancer-related pathways [26]. Under normal circumstances, NMD has a protective role against cancer, but during cancer role of NMD has been reported as that of "bipolar" nature [50]. *On the one hand*, under-activity of NMD during cancer progression suggests a tumor-suppressing role of NMD. Reports show that NMD factors are often down-regulated in cancerous cells leading to under-activity of NMD. As NMD regulates the expression of genes involved in cell growth, differentiation, survival, apoptosis etc.; in order for tumors to proliferate, NMD must be inhibited or at least lowered in some manner. Inhibition of NMD results in the up-regulation of certain growth factors, immune infiltration and consequently metastasis. Such tumor-suppressing role of NMD has been seen in many cancers including hepatocellular carcinoma and pancreatic adenosquamous carcinoma [51, 52]. As a quality control mechanism, degrades many PTC-harbouring dominant-negative TSGs which can otherwise antagonise with the wild-type genes and compromise their role as tumor-suppressors [53, 54].

Contrary to its normal tumor-suppressing role, NMD has been extensively reported to be an active facilitator in cancer progression by inactivating TSGs. It can be due to NMD targeting the wild-type, PTC-harbouring TSGs which might preserve their function even when translated into truncated proteins [50].

Cancer cells exploit NMD-TSGs association for their own benefit in many ways (adding NMD-resistant mutations in dominant-negative alleles, biallelic inactivation, NMD-stimulating mutations in wild-type alleles), making NMD a tumor-promoting mechanism [55, 56]. Moreover, NMD has been reported to induce anti-tumor immunity suggesting that NMD inhibition can help increase the effectiveness of cancer immunotherapy [57, 58].

In conclusion, the role of NMD in cancer is complex and conflicting results have been reported, demanding for the requirements of more research in the area. Distinguishing whether NMD has a tumor-suppressing or tumor-promoting role in cancer is important for designing therapeutic targets against cancer.

1.7 Gaps in the Literature

Besides surveillance, NMD is extensively involved in the regulation of genes including cancer-related genes. Previous studies have reported both tumor-suppressing and -enhancing role of NMD. In conclusion, the role of NMD in cancer is complex, and conflicting results have been reported, demanding the requirements of more research in the area. Exploring the splicing and switching patterns of NMD-targeted isoforms with primary isoforms will enable us to better understand the contribution of MD towards cancer. Distinguishing whether NMD has a tumor-suppressing or tumor-promoting role in cancer is important for selecting therapeutic targets against cancer.

1.7.1 Limitations of using NMD Modulation for Cancer

Although NMD inhibition has been widely proposed as part of cancer therapy and a few NMD inhibitors have been clinically tried, such an approach has yet to overcome certain limitations. For instance, the application of global NMD inhibitors in cancer therapy can mute the tumor-suppressing function of NMD as well. Moreover, available NMD inhibitors are not sufficiently specific in choosing their substrates, leading to the deregulation of other cellular processes. However, the use of chemically modified antisense oligonucleotides (ASOs) mediated NMD

inhibition can potentially help overcome this problem [26]. Likewise, the use of NMD inhibition in cancer immunotherapy is also met with limitations such as a low response rate of NMD inhibition approach and absence of already existing immune response against tumor [59].

In conclusion, although having a paradoxical role in cancer, NMD presents a useful opportunity in cancer therapy. Keeping in view all the above-mentioned limitations of using NMD modulation for treating cancer, it can be easily deduced that more research is required in the field. The development of a successful treatment is not feasible without grasping the horizons of the NMD-cancer relationship. Better characterisation of NMD in cancer can help overcome the existing limitations as well as design personalised treatment strategies. This has become now more feasible with the advancement in full-length reads sequencing techniques and better computational tools.

1.8 Problem Statement

Nonsense-mediated mRNA decay (NMD) is one of the key features of the pathophysiology of cancer but the exact contribution of the pathway towards cancer is paradoxical as it can act both as a tumor suppressor and a tumor promoter pathway. It is crucial to explore the effects of NMD in the context of cancer to develop more effective cancer treatment strategies. Our study aims to understand the role of nonsense-mediated mRNA decay in epithelial cancers by investigating the transcriptional landscape of NMD-sensitive genes while exploring their coupling alternative splicing and NMD (AS-NMD) profiles.

1.9 Aims and Objectives

General Objective: To prepare a tentative pathway using Gene Set Enrichment Analysis (GSEA) and literature to understand the role of common Nonsense-mediated mRNA decay (NMD) sensitive genes in epithelial cancers.

- To elucidate whether common NMD-sensitive genes exist in various epithelial

cancers.

- To identify the NMD-targeted hub genes common in all cancers
- To understand the role of the identified genes in different biological pathways using Enrichment analysis

CHAPTER 2: LITERATURE REVIEW

2.1 Differential Splicing in Cancer

As oncogenes and tumor suppressor genes contain multiple promoters and alternatively spliced variants, differentially spliced genes are common among metastatic cancer cells. Aberrant alternative splicing mechanism results in alternative transcripts being more dominantly expressed than in normal tissues, often leading to tumor initiation, progression, and adaptation. Isoform switching is commonly implicated in cancer, resulting in an increased interest in isoform-specific and splicing event-targeting therapies against cancer [60].

Cancer-related pathways are frequently targeted by alternative splicing events. Alterations in the regulation and expression levels of splicing factors have been linked with metastatic colonization prominently in the case of prostate cancer cells and their derivatives. Some of these splicing proteins include ESRP1, ESRP2, and RBFOX2, NOVA1 and MBNL3. Along with their potential contribution to epithelial–mesenchymal transition, such differentially regulated SF proteins also influence a plethora of AS events as well as other key regulators such as CD44 and GRHL1, involved in regulating cell growth, motility, and signalling [61].

2.2 Isoform switching in Cancer

Advances in sequencing and transcriptome reconstruction at isoform resolution have led to the identification and analysis of isoform switching. For example, isoform switches in cancer have been analysed by a group of scientists [62] using nine cancer types from "The cancer genome atlas" (TCGA). They identified 244 genes with significant isoform switches, out of which 59 were found in multiple cancer types (pan-cancer switches) with FBLN2 being the most common at the pan-cancer level. However, this and other studies have only observed iso-

form switching and have not estimated the potential consequence of the identified switches.

Despite the transcriptome reconstruction and profiling at isoform resolution, isoform-based expression analysis is rare. Studies mostly identify individual isoform switches and analyse their frequency rather than characterising them or predicting their functional impact. In addition to that, reports also associate the presence of mutations with splice variants and isoform switching. Nevertheless, a few researchers have moved their focus from individual switches to the functional effects of isoform usage on the predisposition of diseases specifically cancer.

2.2.1 Pan-cancer Studies of Isoform Switches

Pan-cancer isoform switching studies investigate same and unique switching patterns between various cancer types. To date, three major bioinformatics analyses [63, 18, 64] have been documented which use genome-wide sequencing data from TCGA and PCAWG databases to analyse isoform switching patterns alongside their functional impact at the pan-cancer level.

Systematic analysis performed on samples from 11 TCGA cancers exploring functional consequences and impact of cancer-associated splicing events reported splicing changes as one of the key features of tumors. A subset of such splicing changes has been reported to potentially act as alternative splicing drivers. Although alternative splicing events are generally mapped with isoform switching events, more than 30% of the switches remained unmapped. Isoform switches were found to affect protein-protein interaction networks mostly by modifying protein domains. Similarly, switching between coding and NMD-sensitive or non-coding transcripts may impact regulatory networks. Moreover, the study proposed that a subset of alternative splicing events observed in tumor samples highlight independent oncogenic processes that could explain the impact of functional transformations in cancer and hence could be termed as alternative splicing drivers [18].

Another similar study determined isoform switching in 12 different TCGA cancers using RNA-seq data from more than 55000 cancer patients and reported

potential functional impact of pan-cancer switches. The results showed that 19% of the genes contained isoform switches with functional consequences, with 31 switches identified as powerful biomarkers across multiple cancers based on patient survival analysis. The most common consequences marked protein domain loss, ORF loss, or gain of signal peptides. Furthermore, PPI network revealed enrichment of cell signalling genes mainly AKT1 and VEGFA as hub genes with switching in the former appearing in 5 different cancers [63].

In another comprehensive study, 1209 cancer samples covering 27 cancer types and subtypes from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and normal samples from Genotype-Tissue Expression (GTEx) were retrieved. The idea of cancer-specific most dominant transcripts (cMDT) was introduced. For a cMDT, a transcript must be the most dominant transcript (MDT) first. Significantly expressed most dominant transcripts unique to PCAWG samples in contrast to GTEx samples were termed as cMDT. Novel isoform specific protein-protein interactions were performed to assess the functional and pathogenic consequences. According to the study, most cancer-specific MDTs were caused by genomic mutations. Correlation analyses were performed between coding and noncoding mutations and alternatively spliced cancer-specific MDTs. Interestingly, 22 transcripts from 19 genes having cMDT switch were found in 100% of the cancer samples. The highest number of cMDT switches were detected in breast, ovarian and female reproductive organ carcinomas. Head and neck carcinomas, glioma, melanomas, and beta-cell non-Hodgkin lymphomas showed the least number of the cMDT switches. The study showed that different cancers had different levels of switching and it can be cancer-specific [64].

Genome-wide studies use different statistical approaches and parameters to determine isoform switches. When focusing on cancer-specific isoform switches between dominant transcripts then comparing MDTs (as in [64]) is ideal. However, when considering the minor transcripts switches as well, then isoform usage is a more suitable parameter. This is achievable by comparing either the Isoform Fraction (IF) computed from FPKM values (as in [63]) or Proportion Spliced-In (PSI) score computed from TPM values (as in [18]) between two conditions. It

has been suggested that, based on the use of statistical tests, the Vitting-Seerup approach is more appropriate when comparing AS changes between matched or unmatched cohorts, while the other two approaches are to be used when the focus is to be retained on patient-specific differences [65].

2.3 Nonsense-mediated mRNA Decay in Cancer

Generally, studies attempt to predict the NMD-sensitivity of a transcript using the mammalian 50-55 nucleotides rule, however, there can be other underlying factors such as mutations and AS events which affect the NMD-sensitivity of a transcript as well. As mutations can cause PTC and elicit NMD, researchers often correlate the type and frequency of mutations with the NMD-sensitivity of genes. One such study [55] analysed more than a million mutations across TCGA cancer data and reported 73,855 NMD-causing mutations with STAD, BLCA, and UCEC carrying most of the burden among all 24 TCGA cancers. These NMD-elicit mutations predominantly targeted DNA repair, DNA or RNA binding, and chromatin remodelling pathways.

Mostly NMD targets are identified using knockdown experiments where one or more NMD factors are muted so that depleting NMD-targets can be stabilised and detected. One of the major challenges in studying characterising NMD targets is short-read sequencing. With the advent of long-read sequencing technologies, the identification of many novel NMD candidates has become feasible. Karousis et al. (2021) utilised NMD factors knockdown followed by nanopore sequencing to identify NMD-targeted mRNAs in human cell lines. The study reported several genes with multiple AS events leading to NMD. Additionally, a higher frequency of NMD-sensitive transcripts was marked by the inclusion of alternative exons rather than intron retention or alternative splice sites. Overall, the study reported that most NMD targets follow the mammalian NMD rule and can be both normal and aberrantly spliced transcripts with degradation of aberrant mRNA isoforms being the primary purpose of NMD. Nevertheless, innate variations in gene products and regulatory processes like NMD allow cells the flexibility to adapt in response to environmental alterations or developmental changes [66].

2.3.1 NMD-elicit Mutations

Association of mutation type with NMD stimulation showed that two-thirds of all TCGA mutations that elicit NMD were either frameshift indels or nonsense mutations. Although mutations predominantly associated with NMD stimulation are nonsense mutations (exhibited by TSGs), most frequent mutations in oncogenes are missense [67]. In cancer, NMD causes loss of function in genes by loss of expression. A subset of genes most affected by NMD is TSGs where NMD results in loss of function of gene by loss of expression. On the contrary, NMD only causes hypofunction in non-TSGs while occurring at low frequency [55].

2.3.2 AS-NMD Interplay

Widespread coupling of nonsense-mediated mRNA decay and alternative splicing has been reported extensively in the literature. For instance, Lewis et al. (2003) analysed RefSeq data and reported that 1106 genes undergo AS to produce 1989 NMD-sensitive mRNA isoforms. Such extensive AS-NMD coupling hints towards the crucial role of NMD in cellular processes. Another study utilising NMD inhibition found that muting of NMD contributes to the deregulation of AS, as NMD inhibition strongly impacts levels of splicing regulators. This presents another evidence for the intricate interplay between AS and NMD machinery [68].

2.4 Pathophysiology of NMD in Cancer

2.4.1 Stress, NMD, and Tumorigenesis

Tumor cells are exposed to various cellular stresses ranging from hypoxia and nutrient deprivation to oxidative stress and reduced pH. In order to survive and adapt to such a stressful micro-environment, cells must bring about some modifications in cellular processes and induce adaptive mechanisms. Integrated stress response (ISR) and unfolded protein response (UPR) are two such responsive pathways activated by the cell to re-establish homeostasis. NMD regulates a plethora of genes involved in ISR and UPR, hindering their activation under low stresses. How-

ever, prolonged stress activates ISR and eventually UPR, which in turn inhibits NMD by activating eukaryotic translation initiation factor 2 alpha (eIF2a). For instance, it has been observed that hypoxia suppresses NMD which causes the stabilisation of NMD-sensitive transcripts including stress-responsive mRNAs. In addition to that, phosphorylation of eIF2a induces global attenuation of protein synthesis and eventual suppression of NMD. It also induces the production of stress-responsive factors which help cells adapt to stresses and survive in the tumor micro-environment. On the same note, the generation of reactive oxygen species (ROS) activates eukaryotic translation initiation factor 2 alpha (eIF2a) which suppresses NMD, making the cell more tolerant to oxidative stress. Moreover, amino acid deprivation in tumor micro-environment suppresses NMD eventually upregulating amino acid transporters which help cell maintain homeostasis. In addition to its stress-responsive role, eIF2a is also required for stimulating autophagy in the cell, which contributes towards maintaining homeostasis under amino acid deprivation, cell survival, and tumorigenesis. High levels of prolonged stress can trigger UPR which helps cells adapt and shut down after stress has ceased [54, 69].

Concisely, ISR, UPR, and NMD are associated in a symbiotic regulation. Stress in tumor microenvironment causes activation of eIF2a which, via some unclear molecular mechanisms, leads to NMD suppression eventually enhancing ISR and UPR. Activation of ISR and UPR, in turn, promotes cell survival. A similar relationship exists between NMD and apoptosis. If stress persists even after ISR and UPR, the last resort for cells is to trigger apoptosis. Additionally, excessive stress induces caspases that cleave UPF1, which deregulates NMD activation leading to stabilisation of pro-apoptotic factors that are otherwise targeted by NMD. Normally, NMD inhibits apoptosis, however, induction of apoptosis leads to attenuation of NMD. All this evidence suggests that tumor micro-environment modulates NMD to benefit tumorigenesis.

2.4.2 Controversial role of NMD in Cancer

Mutations in UPF1 in various cancers including pancreatic adenosquamous carcinoma (ASC), liver hepatocellular carcinoma (HCC) and lung adenocarcinoma

(ADC) lead to suppression of NMD causing upregulation of different growth factors and growth pathways which eventually lead to metastasis, immune infiltration and enhanced malignancy of tumor. Likewise, oncoprotein interaction causes upregulation of UPF3A which naturally inhibits NMD, resulting in tumor aggressiveness. These reports show the anti-tumorigenic role of NMD. However, NMD also has shown a tumorigenic role specifically against tumor suppressor genes [50].

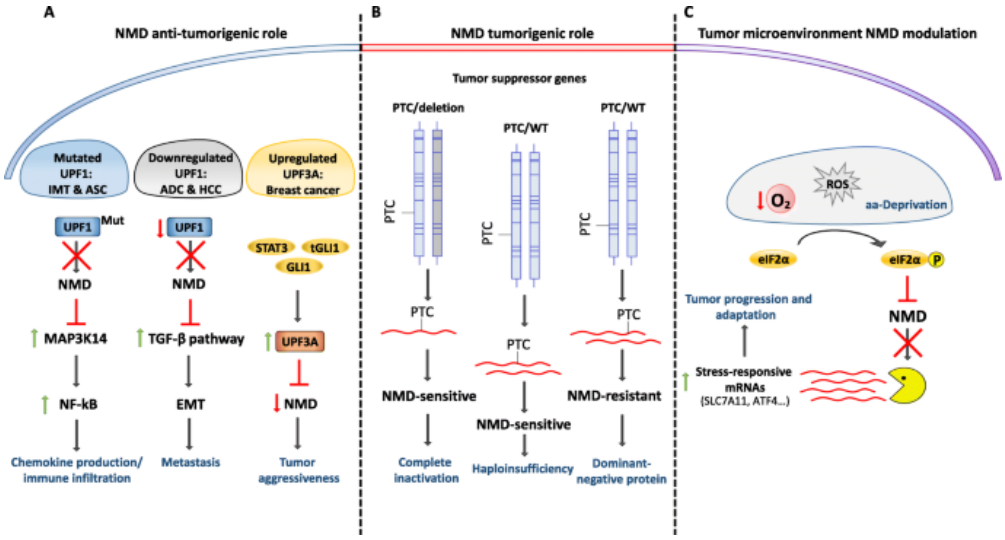

Figure 2.1: Pro- and anti-tumorigenic roles of NMD in cancer [50]

PTC acquisition leading to NMD degradation of TSGs results in either complete inactivation of the PTC+ allele or haploinsufficiency of the wild-type allele. On the contrary, if the TSG transcript acquires a mutation that renders it NMD-resistant, translation of that transcript produces a protein with a dominant negative function on the wild-type allele. In this way, partial or complete inactivation of TSG by NMD can be beneficial for cancer. Conclusively, the role of NMD in cancer is complex and controversial [50].

2.5 Modulating NMD Pathway in Cancer Treatment

The controversial, both pro- and anti-tumorigenic, behaviour of NMD in cancer imposes the greatest challenge in using NMD in cancer therapy. NMD inhibitors have been designed and clinically tried in cancer therapy. Likewise, NMD activation has been also proposed for treating cancer. However, the application of NMD modulation in cancer therapy is highly dependent on the genetic context of the

patient/disease, cancer type, and other related factors. Nevertheless, NMD modulation can be used as personalised therapy alongside traditional cancer therapies to enhance the efficacy of cancer treatment.

2.5.1 NMD Inhibition in Cancer therapy

Inhibition of NMD has been long speculated as an appropriate approach for enhancing the efficacy of cancer therapy. As NMD regulates transcription levels and, directly or indirectly, influences many cellular processes; it can be deduced that NMD could easily impact a plethora of cell responses. Cancer therapy usually depends upon inducing apoptosis, which is inhibited by NMD, it has been tested and observed that muting NMD promotes the efficacy of anti-tumor drugs in cancer cell lines. There are three major ways in which NMD inhibition induces apoptosis [70].

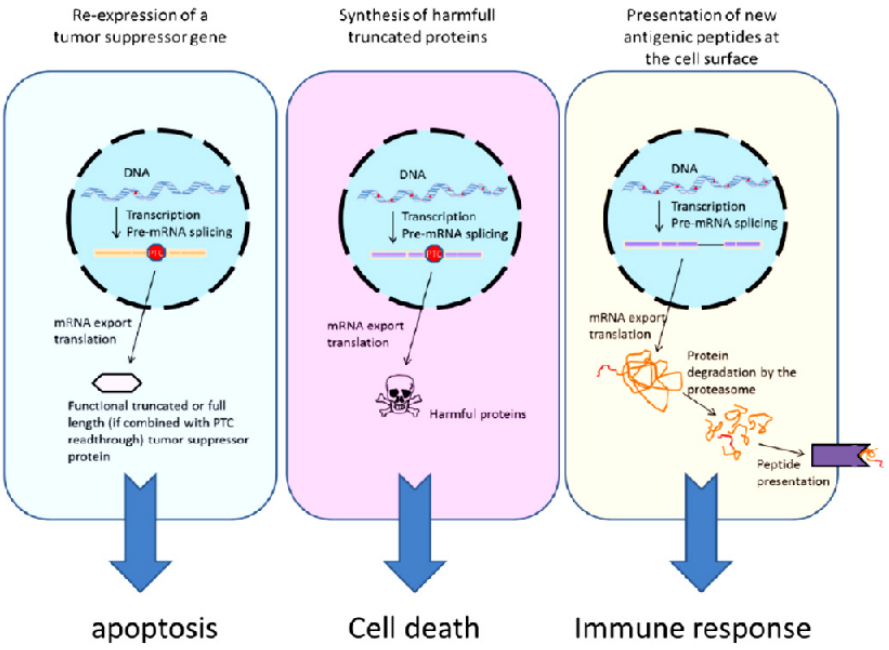

Figure 2.2: The triple effect of NMD inhibition on cancer cells [70]

Firstly, inhibition of NMD helps cell enter apoptosis by stabilising the PTC+ transcripts and increasing the proportion of truncated or harmful peptides in the cell. Some of these peptides can induce deleterious action, reinforcing apoptosis [71]. Moreover, apoptosis reinforces NMD inhibition by cleaving NMD factors (as discussed in section 2.4.1). Secondly, NMD inhibition can help restore the

expression of TSGs, leading to the inhibition of cell growth and eventual apoptosis. Additionally, TSGs reactivation increases the efficacy of cancer treatment by making the cell sensitive to chemotherapy as well [72]. Lastly, muting NMD helps elicit a stronger immune response by allowing the antigens which are otherwise degraded by NMD. Underactivity of NMD allows aberrant transcripts to be translated into truncated peptides which are then presented at the cell surface, flagging the cell to immune cells. Subsequently, the C-terminus end of the aberrant protein is recognised by immune cells as a foreign antigen, inducing an immune response that results in the elimination of that cancer cell [57].

2.5.2 NMD Activation in Cancer therapy

Some cancer types including HCC and ASC could benefit from NMD activation (as discussed in 2.4.2) by making use of the anti-tumor role of NMD. The main side effect of global NMD activation is that it could enhance the tumorigenic function of NMD as well. However, gene-specific activation of NMD is safer and the use of oligonucleotides to induce NMD of only oncogenic transcripts can overcome the aforementioned challenge[54]. Furthermore, it has been speculated that NMD activation can be used in combination with traditional immunotherapy to help protect normal or non-cancer cells [73]. Nevertheless, designing such NMD activators is a challenging task and NMD activating approaches, though proposed, are not yet fully developed.

CHAPTER 3: DESIGN AND METHODOLOGY

The study explores the role of NMD in epithelial cancers by identifying isoform switches in multiple carcinomas and investigating the splicing/switching patterns of NMD-sensitive transcripts. Another goal is to find an association between differential usage of NMD-sensitive transcripts and functional consequences leading to predisposition or inhibition of cancer. This can give major insights into how NMD-targeted genes may contribute to tumor progression or suppression. The overall methodology designed for accomplishing these goals is shown in 3.1

Figure 3.1: General Workflow of Project

3.1 Data Retrieval from TCGA

RNA-seq is a high throughput sequencing technology to study transcriptome-wide dynamics of gene expression analysis. Currently, RNA-seq datasets can be retrieved from public databases such as GEO (Gene Expression Omnibus) [74], Array Express [75] NCBI (The National Centre for Biotechnology Information) [76], ENA (The European Nucleotide Archive) [77], SRA (The Sequence Read Archive) [76] and TCGA (The cancer genome atlas) [78]. For this study, data was retrieved from TCGA.

TCGA is a well-coordinated omics data hub outlining the molecular basis of cancer in a comprehensive manner. TCGA legacy data accumulated at Genomic Data Commons (GDC) has been mapped with older genomes having UCSC annotations and contains batch effects. However, Toil workflow was developed to recompute raw data from TCGA, Pacific Pediatric Neuro-Oncology Consortium (PNOC:https://pnoc.us/), Genotype Tissue Expression Project (GTEx) [79], Therapeutically Applicable Research to Generate Effective Treatments (TARGET:https://ocg.cancer.gov/programs/target) was realigned to create a consistent and batch-effects free meta-analysis of all four datasets with new Ensembl annotations [80]. Toil recomputed data is publicly available at Xena browser (http://xena.ucsc.edu) and presents an excellent resource for bioinformatics analysis.

For this study, TOIL RSEM transcript expected count dataset (RNA-seq data quantified with RSEM) along with the corresponding probemap was downloaded from the same section of Xena browser. The dataset consists of log2 transformed expression data from 10530 samples with 198,620 isoforms for 33 TCGA cancers. For gene co-expression network analysis, gene expression dataset along with the corresponding gene probemap was downloaded from the TCGA PANCAN section of Xena browser. The dataset consisted of log2 transformed RSEM expected count data of 10530 samples with 60,499 genes/identifiers for same 33 TCGA cancers. As the TOIL RSEM expected count data form both transcript and gene expression has been aligned using hg38.p3, gencode v23 comprehensive model; corresponding GTF and FASTA sequence file were downloaded from Ensembl (https://www.gencodegenes.org/human/release$_2$3.*html*).

In addition to raw counts, clinical and phenotype data was also retrieved from the TCGA PANCAN section of Xena browser. Clinical data basically contains updated survival data for patients from whom samples were extracted. Phenotype dataset consisted of sample type (tumor or adjacent) and primary disease (cancer), and survival data contains curated clinical data of cancer patients such as their age, race, tumor stage etc.

3.2 Data Pre-processing and Sample Selection

3.2.1 TCGA Cancer and Samples Selection

The initially downloaded PANCAN transcript expression dataset contained samples from all 33 TCGA cancers. For the current study, carcinomas and adenocarcinomas with 20% or more samples harbouring NMD-elicit TSGs based on pan-can analysis [55] were selected. Selected tumor types were further narrowed down to cancers with >10 paired samples, finally leaving with only 10 TCGA cancers as listed in Table 3.1.

Table 3.1: List of selected epithelial cancer types

Cancer	Abbrev.
Bladder Urothelial Carcinoma	BLCA
Stomach Adenocarcinoma	STAD
Head & Neck Squamous Cell Carcinoma	HNSC
Kidney Clear Cell Carcinoma	KIRC
Colon Adenocarcinoma	COAD
Esophageal Carcinoma	ESCA
Lung Squamous Cell Carcinoma	LUSC
Lung Adenocarcinoma	LUAD
Liver Hepatocellular Carcinoma	LIHC
Breast Invasive Carcinoma	BRCA

For each selected cancer, subset of patients was obtained for transcript as well as gene expression separately, using sample IDs from the survival data. Samples with either 'Solid Tissue Normal' or Primary Tumor' type, here referred to as Adjacent and Tumor respectively, were extracted. In case of transcript expression data, paired samples were extracted for paired analysis, while for the unpaired analysis all the paired samples and additional 50 randomly selected 'Primary Tumor' samples were extracted. Corresponding clinical, phenotype and gene expression data were also retrieved for the selected patients. Summary of the selected samples for

each cancer is in Table 4.1.

3.2.2 Pre-processing of Expression Data

As the downloaded expression data was log2(x+1) transformed, antilog was taken, and pseudo count of 1 was subtracted from each of the read counts so that raw read counts could be obtained. The approach was applied for both transcript and gene expression data. However, gene expression data needed to be further pre-processed before feeding it to WGCNA pathway. Thus, VST transformation of Deseq2 was utilised. Transcript expression data, on the other hand, was directly fed into the IsoformSwitchAnalyzeR(ISAR) as the package requires raw read counts.

3.3 Differential Isoform Usage Analysis

The transcript expression files generated in the previous step were then analysed, individually for each cancer type, using a Bioconductor package called IsoformSwitchAnalyzeR [63]. IsoformSwitchAnalyzeR (ISAR) identifies isoform switches, by calculating isoform usage and applying statistical methods, from RNA-Seq derived quantification of full-length isoforms. It allows the integration of various annotations ranging from the addition of ORFs to prediction analysis from external tools. One distinguishing feature of ISAR is that it can predict the NMD-sensitivity of transcripts using the mammalian NMD rule. The analysis can be carried out either at genome-wide analysis or even at the individual gene level. By allowing the integration of analysis from external tools, it helps predict the functional consequences of the identified switches as well as study their AS patterns, thereby enhancing the useability of RNA-Seq data at the isoform resolution. The functions and steps implemented using ISAR are described in the coming subsections.

3.3.1 Importing Data into R

Isoform data was imported into R using `importRdata()` function of ISAR. It required previously retrieved (as mentioned in section 3.1) isoform quantification,

GTF and FASTA sequence files and manually curated design table obtained from phenotype file enlisting sample ID and its corresponding condition (Adjacent or Tumor). The `importRdata()` function takes in all these files and performs several computations required for further analysis.

Briefly, the function performs various roles including correction of some of the annotation problems such as unassigned transcripts and merged genes created when doing transcript assembly; and calculation of isoform abundance (and normalises data) in the form of FPKM or RPKM values. Furthermore, it obtains gene expression by adding the expression of isoforms belonging to the same gene. Subsequently, it calculates mean isoform and gene expression values and computes log2 fold change (FC), Isoform Fraction (IF), and dIF values from the mean gene and isoform expression values. Moreover, it also structures data into switchAnalyzeRlist by annotating each isoform with its genomic annotations (such as ORFs and genomic coordinates), conditions, and other computed values.

Here, IF is defined as the fraction of expression or quantifies the proportion of the parent gene expression originating from a specific isoform. Whereas dIF is the difference in the isoform usage or the difference in isoform fraction in each condition. It measures the effect size (like fold changes are in gene/isoform expression analysis).

$$IF1 = \frac{iso_{value1}}{gene_{value1}}$$
$$IF2 = \frac{iso_{value2}}{gene_{value2}}$$
$$dIF = IF2 - IF1$$

where *isovalue1* and *isovalue2* represent the expression value of transcript(n) in condition 1 and condition 2 respectively. Likewise, *genevalue1* and *genevalue2* represent the expression value of gene(n) in condition 1 and condition 2 respectively.

The obtained `switchAnalyzeRlist` concatenated with `isoformFeatures` and

other annotations are then further used to analyse isoform switches.

3.3.2 Pre-filtering

Lowly expressed transcripts or genes can bring uncertainty to analysis. Removal of isoforms with low IF values helps simplify analysis by filtering out isoforms that contribute too minimally to the gene expression. Genes with a single isoform cannot have differential isoform usage. Similarly, genes with very low change in isoform fraction cannot contribute significantly to functional consequences. Thus, such unwanted data must be removed in order to make further analysis better and simpler.

The `preFilter()` function was used to filter out uninteresting transcripts and genes including non-expressed or lowly expressed transcripts and genes, single isoform genes, genes without differential isoform usage, and isoforms with very low isoform fraction (IF).

Following cut-off values were utilised for pre-filtering: Gene expression = 1; Isoform expression = 0; IF value = 0.05; dIF value = 0.1.

3.3.3 Identifying Isoform Switches

The function `isoformswitchTestDEXSeq()` performed the differential isoform usage test, enabling the switch identification. The function has incorporated another Bioconductor statistical package called DEXSeq [81]. DEXSeq was initially designed to implement differential exon usage, but the concept has been updated for differential usage at isoform resolution in the ISAR package.

$$DEU = \frac{no.of exons from the containing transcript}{no.of all the transcripts from the gene}$$

Where DEU is the change in the relative usage of the exons caused by the experimental condition.

For a transcript, the formula calculates as: for each transcript or part of a transcript in each sample, and it counts how many reads map to this transcript and counts how many reads map to any other transcript of the same gene. It then

considers the ratio of the two counts and how it changes across conditions to infer changes relatively. This method is called bin read counting and can be carried out either at the exon or transcript level.

The ISAR tests the significance of the change in isoform usage with the Mann–Whitney U test. The isoform switch with absolute dIF value greater than the selected cut-off and FDR less than the selected level of significance represents a significant change in isoform usage. In other words, two parameters used by ISAR to define a significant isoform switch include the alpha argument (which indicates the statistical significance of switch between 2 conditions in the form of FDR corrected P-value also called the Q-value cut-off) and the `dIFcutoff` argument (which indicates the minimum change in isoform usage i.e., |dIF|).

For the current study, `isoformSwitchTestDEXSeq()` function was used to test significance of isoform switches. Non-switching genes were removed using the `'reduceToSwitchingGenes'` argument. A default q-value cut-off of 0.05 was used. The output Excel file from DEXSeq contains different columns enlisting annotated accession IDs of each of the transcripts and their corresponding meta-data, statistical calculations, and statistical parameters evaluated via the DEXSeq function.

3.3.4 Analysing Open Reading Frames (ORFs)

`IsoformSwitchAnalyzeR` provides three options for analysing open reading frames (ORFs). Firstly, it can analyse only known isoforms that have already been added from the annotation database (via GTF file) in the form of CDS. secondly, it can analyse novel isoforms only by predicting ORFs if a completely de-novo isoform reconstruction was performed and annotated ORFs are not available in the database. Thirdly, it can analyse both the novel and the known isoforms by first adding ORFs from GTF file and then predicting ORFs for isoforms with no ORF.

The third option was utilised for this study as the GTF file was available. The `AnalyzeNovelIsoformORF()` function was used to predict the open reading frames (ORF) and NMD-sensitivity of the isoforms. Out of the four methods available for the prediction of ORF, `'longest.AnnotatedWhenPossible'` (which identifies

the longest ORF in a novel isoform that has the same CDS start sites as one of the annotated isoforms) was applied. Moreover, for predicting the NMD-sensitivity of the given isoform PTC Distance of 50 was used.

3.3.5 Extracting FASTA Sequences

The `extractSequence()` function of ISAR was used to extract nucleotide (NT) and amino acid (AA) sequences of the transcripts corresponding to the genomic coordinates of the isoforms from the sequences of the reference genome. The sequences (NT and AA) are outputted as FASTA file formats and can be further utilised for external tools analysis.

3.3.6 Visualisation

Results of ISAR were visualised using switch plots, enrichment plots, and volcano plots. The `switchplot()` function was used for the final visualisation of the gene-specific switch plots. Similarly, the `extractSplicingEnrichment()` function of ISAR provided the AS events enrichment across conditions. The volcano plot was generated using `ggplot()` function.

3.4 WGCNA

WGCNA is an R package which develops a co-expression network utilising Pearson correlation by default or a customised distance measure. There are several other functions of WGCNA i.e., co-expression network analysis, module detection, Meta-correlation analysis. For this purpose, there are several libraries (functions) present in WGCNA. It uses hierarchical clustering and has several 'tree cutting' options for the identification of modules. Hence WGCNA is the most frequently used tool, well supported and documented [82].

3.4.1 Data Input and Pre-processing

Before inputting data into WGCNA workflow, data needed to be cleaned in order to make it suitable for network analysis. For each cancer type, genes with sig-

nificant isoform switches were extracted and their gene expression was retrieved from the PANCAN gene expression data. Pre-filtering was performed to keep only genes with total sum of 20 or more reads across the samples. Subsequently, expression dataset was subjected to variance stabilising transformation using the `vst()` function of DESeq2 package (Love, Huber, and Anders 2014). Data was cleaned by removing genes and samples with excessive missing values as well as samples regarded as outliers by sample clustering. Appropriate cut height was selected, and outlier samples were removed. Moreover, clinical data was loaded into R and traits were related with the sample dendrogram. Same process was performed for each cancer dataset.

3.4.2 Constructing Co-expression Network

Data prepared in the previous step was then used to run WGCNA workflow. Appropriate soft-thresholding power (), the minimum power value with scale-free topology fit index reaching threshold of 0.90, was chosen for each cancer dataset using `PickSoftThreshold()` function of WGCNA. Additionally, Pearson correlation for each gene pair was computed, raised to the selected power, and then used to calculate the signed adjacency matrix with the `adjacency()` function.

Subsequently, the Topological Overlap Matrix (TOM) was computed, via the `TOMsimilarity()` function, to use it as input for hierarchical clustering analysis, and gene modules were detected by using a dynamic tree-cutting algorithm.

$$TOM_{ij} = \frac{(\sum_u a_{iu} a_{uj} + a_{ij})}{min(k_i k_j + 1 - a_{ij})}$$

Where TOMi j denotes TOM matrix, aiu represents values between gene pairs (i, u), (i, j) and (u, j), respectively obtained for signed adjacency matrix. Whereas connectivity of genes i and j are represented by ki and kj respectively.

The TOM matrix was then converted to the `DistTOM` (dis-similarity TOM)

matrix. This dis-similarity TOM matrix was used for the hierarchical clustering algorithm to construct modules.

$$DistTOM_{ij} = 1 - TOM_{ij}$$

The `hclust()` function was used to construct a hierarchical clustering tree of genes which presents highly co-expressed genes in the form of closely interlinked branches with very small distances. Eigengenes for each module were calculated and then clustered based on their correlation in order to identify the most similar modules which can then be merged into larger modules. Merging of modules is crucial as it helps bring together the genes that have similar expression profiles.

3.4.3 Detecting Clinically Interesting Modules

After the detection of modules, the next step is to relate of modules with clinical traits in order to determine clinically significant modules. Detection of interesting modules and hub genes requires two parameters: Gene Significance (GS) which shows the correlation between the gene and the trait, and Module Membership (MM) which represents the similarity between genes of every module by correlating module eigengene with the gene expression profile.

Module eigengenes (MEs, the first principal component of a module) were correlated with external traits to identify highly correlated module with disease status (Adjacent vs Tumor) and to investigate the most significant associations, in the form of heat maps. Heat maps were generated with correlation and p-value for co-expression modules of each cancer by relating modules to disease status. Module(s) with highest significant (p-value<0.05) correlation with 'Tumor' as disease status was selected as the interesting module. Eventually, GS vs MM scatter plot for interesting module(s) was visualised. It should represent a high correlation with p-value less than the level of significance (0.05).

3.4.4 Identifying Hub Genes

Hub genes are defined as the genes which interact with many other genes in a gene network and thus potentially play significant role in biological system. In WGCNA, hub genes are characterised based on high MM and GS values. For this study, hub genes were determined if gene had high module connectivity (`cor.geneModuleMembership` $>= 0.75$) and high clinical trait relationship (`cor.geneTraitSignificance` $>= 0.2$).

3.5 Pan-Cancer-Wide Genes

For each TCGA cancer, genes were selected at the intersection between WGCNA hub genes and ISAR switching genes. The selected genes for individual cancer were integrated to get Pan-cancer-wide genes. Switching hub genes that reoccurred in more than 3 cancer types were selected as Pan-cancer wide (PANCAN) hub genes. These genes were then used for further analysis.

3.5.1 Enrichment Analysis

Gene-set enrichment analysis (GSEA) is a useful technique to help functionally characterise large gene lists, such as the results of gene expression experiments. This technique finds functionally coherent gene-sets, such as pathways, that are statistically over-represented in each gene list. The idea is to visualise the association between the switching genes and disease or biological pathway. Various software and web-based tools are available for performing functional analysis. For this study, gprofiler2 and Enrichr were used [83, 84, 85]. Users provide a list of genes of interest and select the annotation categories from different sources to determine enrichment. Go terms, KEGG pathways, Reactome pathways, and Uniprot key terms were explored.

3.5.2 Literature Search

The role of selected pan-cancer-wide hub genes in cancer was explored using extensive literature studies. Based on the findings, the genes were categorised as either oncogenes or tumor suppressor genes (TSGs). The potential role of the selected genes in tumor progression or suppression was accommodated based on their interactions and general functions as well.

3.5.3 AS-NMD Patterns

Alternative splicing, isoforms usage and NMD sensitivity of transcript variants of the finally selected genes were explored by visualising switch-plots.

CHAPTER 4: RESULTS

4.1 Data Selection and Preprocessing

Out of all TCGA cancers, data for only 10 epithelial cancer types was retrieved. Expression data for only a subset of patients was obtained for each selected cancer. For paired analysis, all available paired samples were extracted. All the paired samples and an additional 50 randomly selected 'Primary Tumor' samples were obtained for unpaired analysis. A summary of the selected samples for each cancer is in Table 4.1.

Table 4.1: Summary of the selected samples for each cancer

Cancer	Paired Samples	Unpaired Samples	
		Tumor	Adjacent
BLCA	19	88	19
STAD	33	116	33
HNSC	43	136	43
KIRC	72	194	72
COAD	26	102	26
ESCA	13	76	13
LUSC	50	150	50
LUAD	58	166	58
LIHC	50	150	50
BRCA	109	268	109

The selected samples were subjected to preprocessing. Log transformation of expression data resulted in raw expression count data obtained which was further utilised for switching and WGCNA analysis. The resulting transcript expression count was fed to ISAR while gene expression data was further subjected to Variance stabilising transformation (VST). The VST transformation helped generate

a matrix of values with stabilised variance across the range of mean values. In other words, the transformation resulted in normalised data suitable for WGCNA analysis.

4.2 Isoform Switching in Individual Cancer Types

In order to determine isoform switches in cancer, Toil recomputed RSEM transcript expression was utilised. Switching analysis was performed separately for each cancer type. Pre-filtering removed more than 80% of the transcripts in each dataset. For each cancer type, differential usage of remaining isoforms was tested using paired as well as unpaired approaches separately. Results obtained for individual cancers are summarised in tables 4.2 and 4.3.

Table 4.2: Summary of Paired Isoform Switching Analysis

Cancer	PAIRED			
	isoDTU	**gDTU**	**isoNMD**	**gNMD**
BLCA	8327	2760	1485	1105
LUAD	9416	3124	1722	1287
LUSC	13501	4470	2377	1795
LIHC	5050	1814	903	711
KIRC	9109	3070	1596	1237
STAD	8701	2726	1457	1086
ESCA	6362	1982	1127	845
COAD	9415	3143	1677	1257
HNSC	7566	2561	1312	1003
BRCA	7921	2630	1333	1011

For the paired analysis, on average, around 8,000 switching isoforms (isoDTU) and more than 2700 switching genes (gDTU) were found. Out of which, more than 1400 isoforms were NMD-sensitive (isoNMD), belonging to more than 1100 NMD-sensitive genes (gNMD) on average. For the unpaired analysis, around 10,000 switching isoforms (isoDTU) with 3200 switching genes (gNMD) were found on average across all 10 cancers. Moreover, on average, more than 1700 isoforms

Table 4.3: Summary of Unpaired Isoform Switching Analysis

Cancer	UNPAIRED			
	isoDTU	**gDTU**	**isoNMD**	**gNMD**
BLCA	9432	3060	1645	1219
LUAD	10277	3352	1866	1372
LUSC	13853	4479	2438	1821
LIHC	6154	2106	1100	831
KIRC	11618	3798	2026	1537
STAD	11615	3602	1984	1475
ESCA	8443	2623	1435	1081
COAD	10203	3409	1758	1335
HNSC	8041	2667	1412	1064
BRCA	9382	3092	1609	1198

were NMD-sensitive (isoNMD), belonging to more than 1200 NMD-sensitive genes (gNMD). On average, 8537 switching isoforms with 10427 NMD-sensitive genes and 9902 switching isoforms with 1299 NMD-sensitive genes were found using paired and unpaired isoform switching analysis respectively.

4.3 Overlapping Switches Across Cancer Types

Isoform switches of different tumors were compared in order to retrieve switches that are common among cancer types. The comparison was performed separately for the results of paired and unpaired analysis. In the case of paired analysis, only 19 switching genes were common across all cancer types; whereas in the case of unpaired analysis, 45 switching genes were common across all cancer types. These genes are listed in 4.4 and 4.5 respectively.

AC012593.1	FAM153A	RPGR	TRPV4
AKR1C2	MAPK1IP1L	RPP30	TTLL4
CCDC144A	POLB	TCAIM	WWP1
CLEC18C	RNF7	TMEM184C	ZNF775
FAM122C	RP11-408H1.3	TNFRSF10A	

Table 4.4: List of switching genes common across all cancers for paired analysis

AIM1L	CDC37L1-AS1	KCNAB2	PLS1	SSX5
AKR1C2	CORO7	KCNH2	POLB	TCEB1
ANKRD34C-AS1	FAIM	KCTD17	PRUNE	TMEM184C
ARV1	FAM122C	KIAA0020	PYCR2	TNFRSF10A
ATG4C	FAM153A	LINC00879	RET	TRPV4
ATXN10	FKBP1A	MBIP	RP11-408H1.3	WWP1
C5	FYN	MCFD2	RPGR	XPO7
CALCOCO2	IL6ST	MYH14	RPL37P6	ZNF208
CCDC144A	INTS6	ORMDL3	SPATA1	ZNF213

Table 4.5: List of switching genes common across all cancers for unpaired analysis

Overall commonality and recurrence of switching genes were visualised using Upset plots and bar plots (as shown in Figures 4.1 - 4.4) respectively. Upset plots were created using complexHeatmap package in R [86]. The mode as "intersect" was used and combinations with the degree of intersection greater than 7 were plotted. In other words, the frequency of switching genes common across different combinations of more than 7 cancer types was visualised.

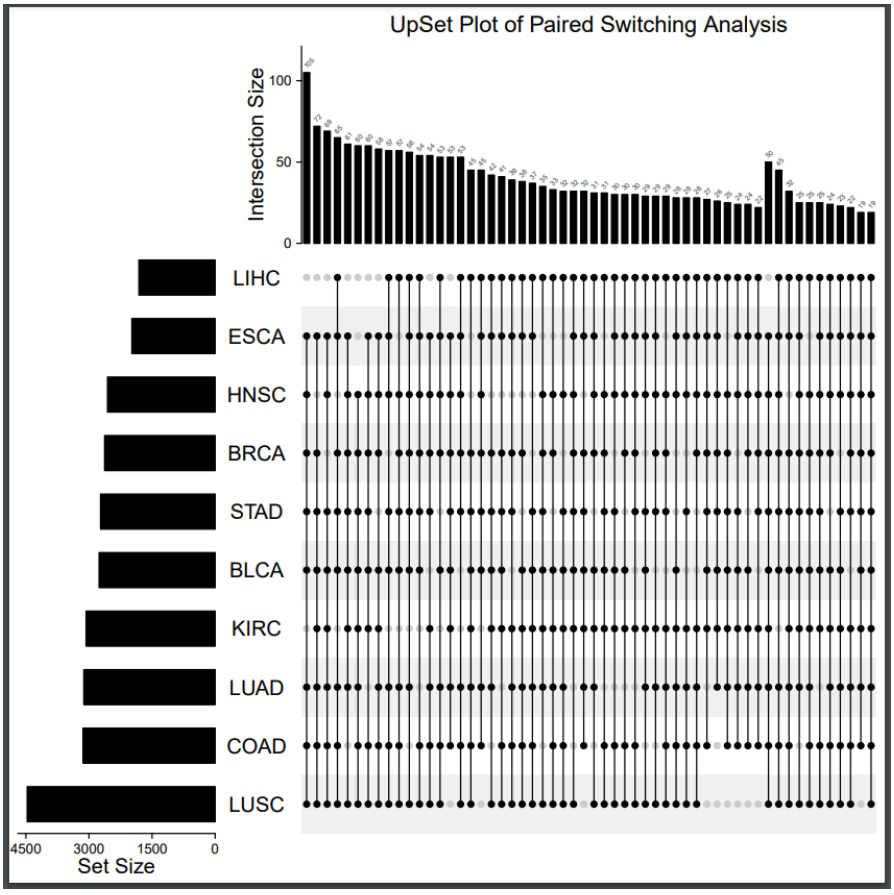

Figure 4.1: UpSet Plot of Paired Isoform Switching Analysis

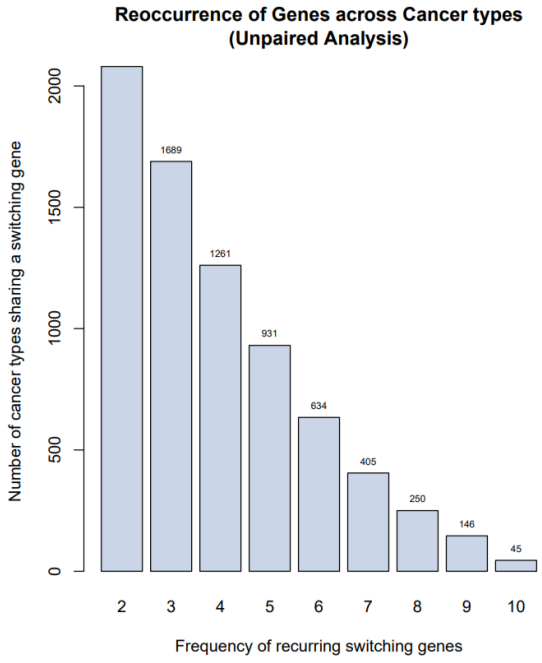

Figure 4.4: Recurrence Frequency Bar-plot of Switching Genes in Unpaired Analysis

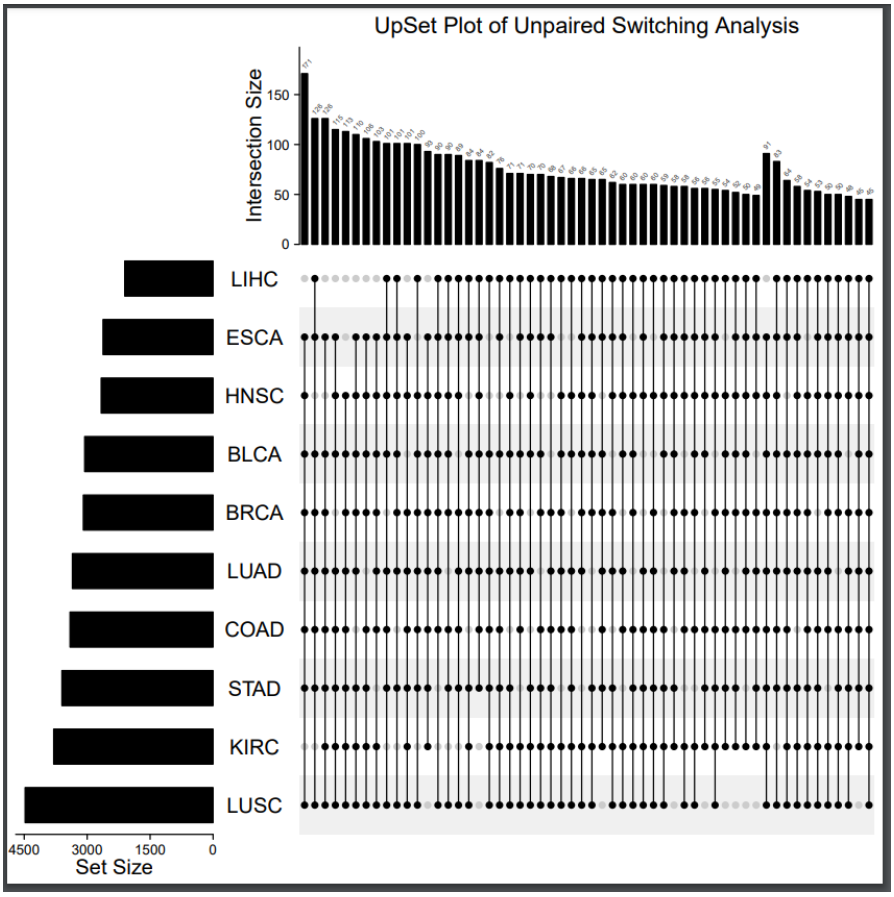

Figure 4.2: UpSet Plot of Unpaired Isoform Switching Analysis

Only 11 switching genes were common across all 10 cancers for both paired and unpaired analysis as illustrated by the Venn diagram in Figure 4.5. These genes are listed below in Table 4.6.

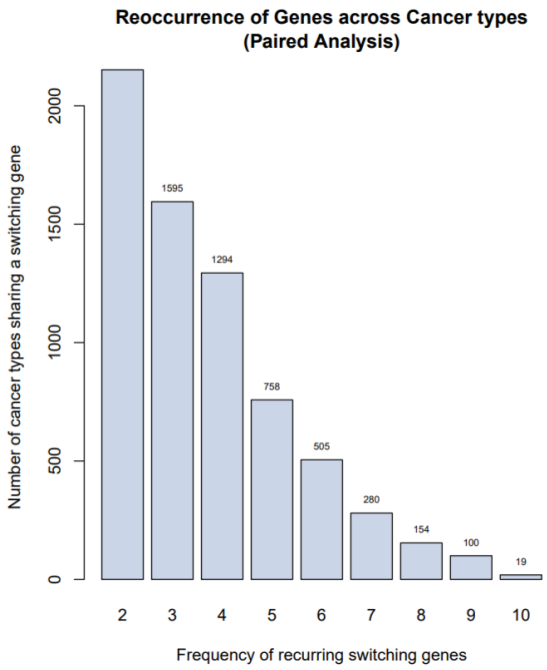

Figure 4.3: Recurrence Frequency Bar-plot of Switching Genes in Paired Analysis

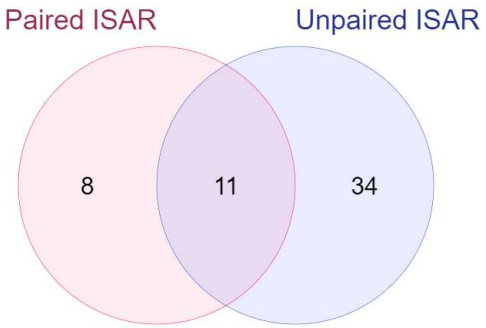

Figure 4.5: Venn diagram of paired and unpaired switching genes

Table 4.6: Common switching genes between paired and unpaired switching analysis

Gene Names	
AKR1C2	RPGR
CCDC144A	TMEM184C
FAM122C	TNFRSF10A
FAM153A	TRPV4
POLB	WWP1
RP11-408H1.3	

4.4 Weighted Gene Co-expression Network Analysis

4.4.1 Sample Clustering and Network Construction

Weighted gene co-expression network analysis was performed in order to get genes of interest that are highly correlated with tumor conditions. The normalised gene expression data was examined for outliers using hierarchical clustering and outlying samples were removed based on suitable cut-off height for each cancer dataset. Dendrograms obtained for all 10 cancers are provided in Figures A.11 - A.20.

For each cancer dataset, appropriate soft-thresholding power was selected for generating the weighted co-expression network based on suitable scale-free fit (threshold = 0.8) and mean connectivity (shown in Figures A.1 - A.10). Although 80% of the threshold was chosen, power was selected by observing individual graphs and choosing the power value at which the fit index reaches a plateau. The weighted gene co-expression network was generated individually for each cancer type and modules were obtained. Genes with the most similar expression profiles clustered into the same module. Modules with similar profiles were merged by using cut-height = 0.25. Dendrograms of genes clustered on the topological overlap matrix (TOM) based on dissimilarity, mapped with the original as well as merged module colours, are given in Figures A.31 - A.40. Parameters used for the network generation process are summarised in Table 4.7.

Table 4.7: Selection of Parameters for Running WGCNA. Switching genes represent the genes initially input for WGCNA. The respective number of samples were removed based on sample cut-height.

Cancer	Selected Genes	Sample cut-height	Samples removed	Power
BLCA	7510	130	7	12
LUAD	8222	102	9	9
LUSC	8109	125	1	5
LIHC	5976	100	5	12
KIRC	7850	110	7	9
STAD	9335	150	8	10
ESCA	8834	150	2	12
COAD	7282	90	8	8
HNSC	6690	- -	0	12
BRCA	8775	130	1	9

4.4.2 Relating Modules with Traits

The obtained modules were correlated with disease status (external trait) to find their biological significance. A module with the highest significant correlation with disease (tumor) status was selected for each cancer type as shown in Figures 4.6 - 4.15. The threshold of p-value<0.05 was set as a criterion to determine the significance of correlated modules.

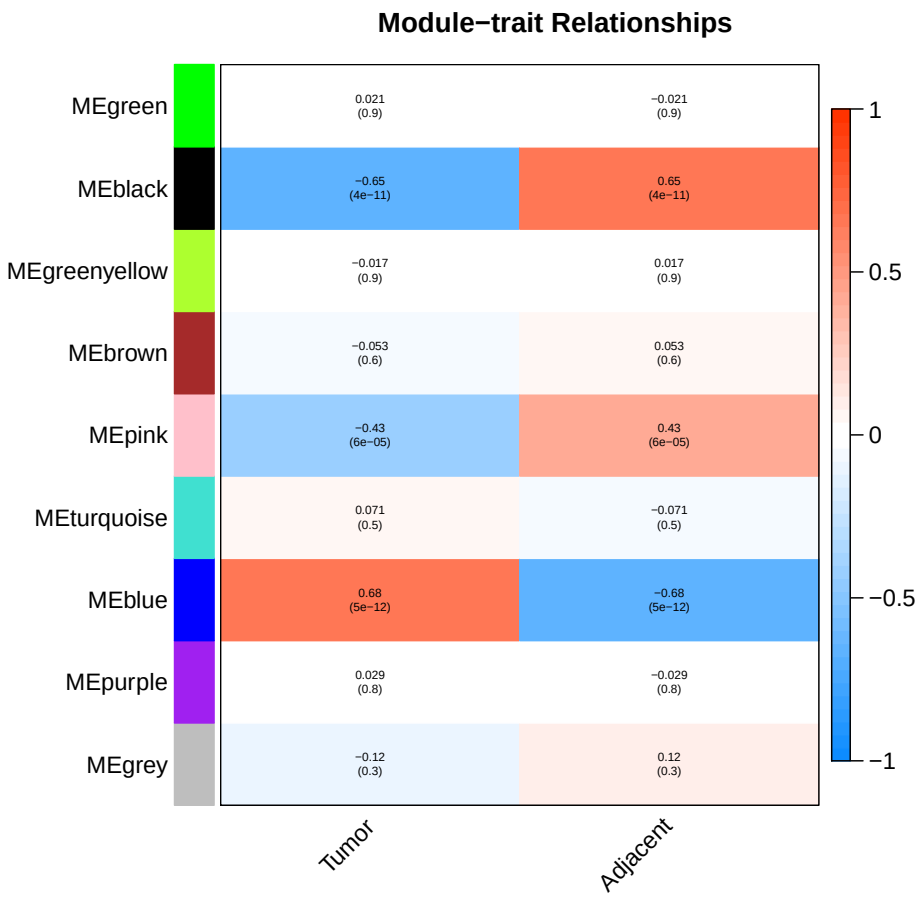

Figure 4.6: Module-trait Relationship for BLCA

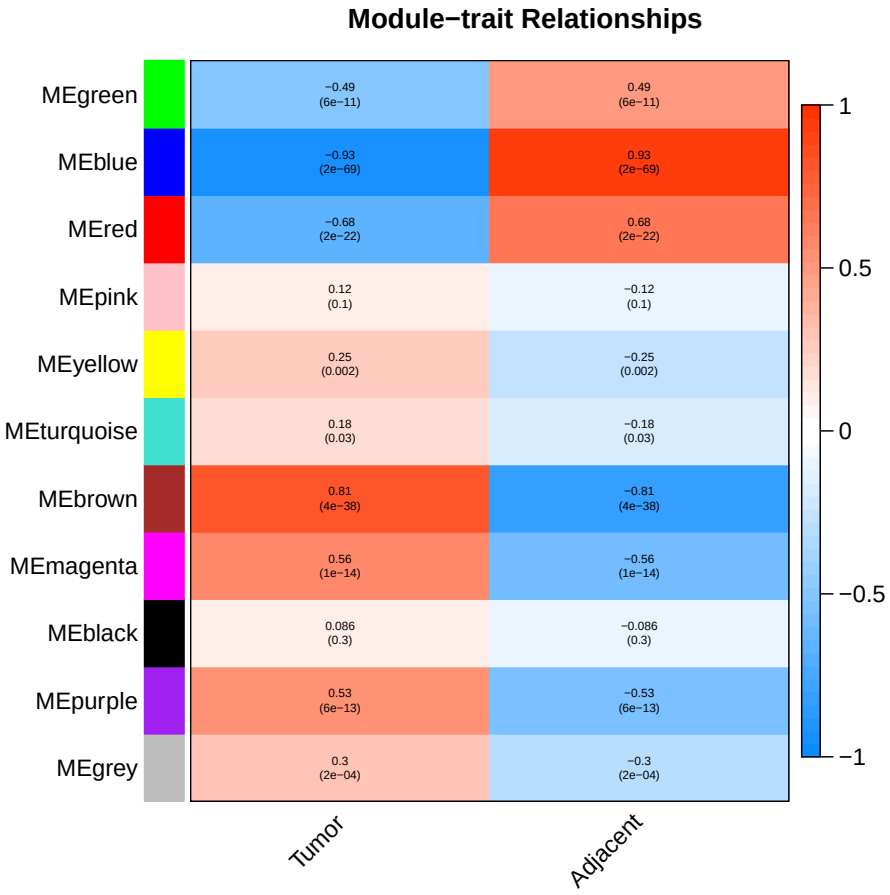

Figure 4.7: Module-trait Relationship for LUAD

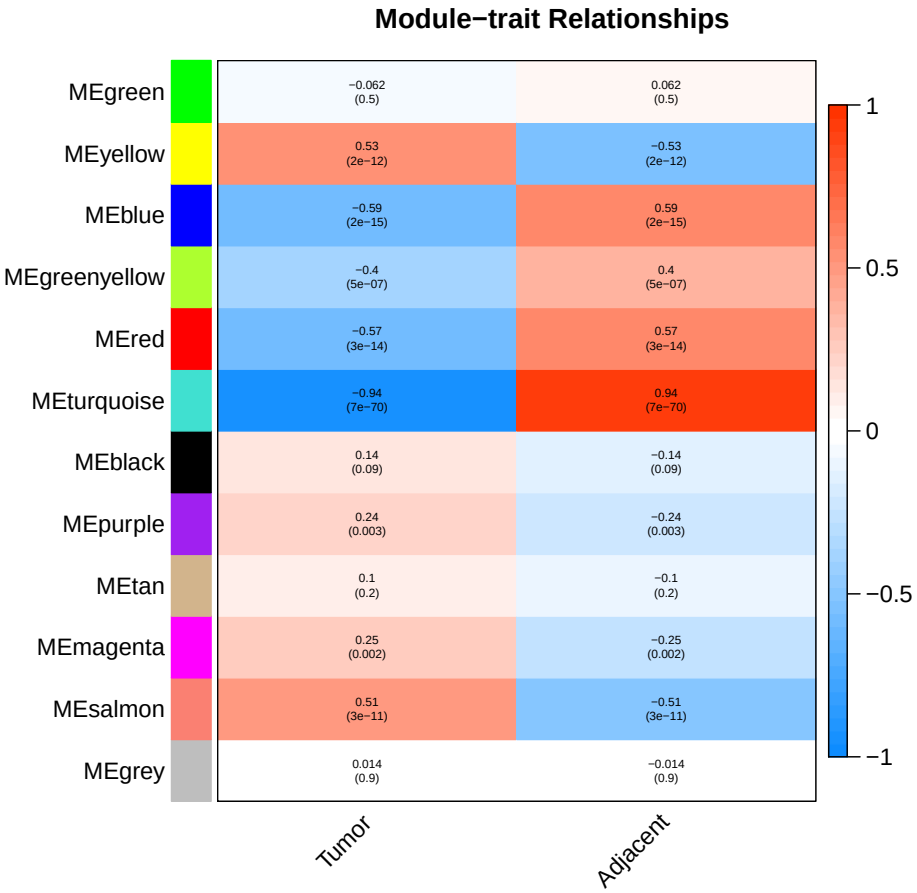

Figure 4.8: Module-trait Relationship for LUSC

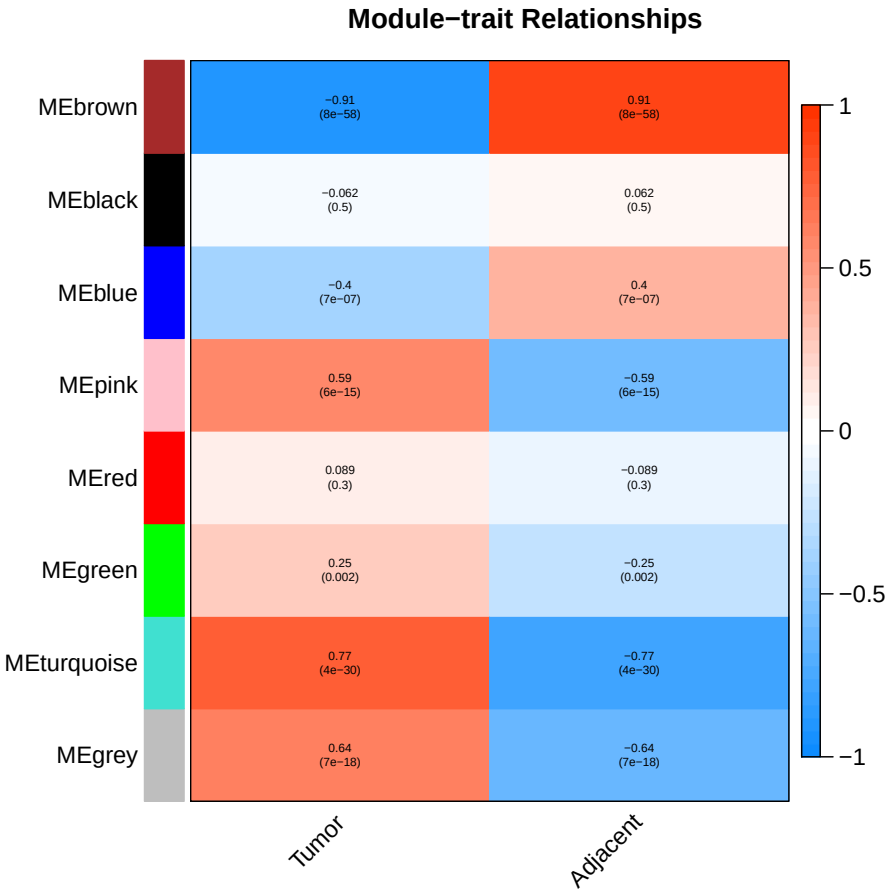

Figure 4.9: Module-trait Relationship for LIHC

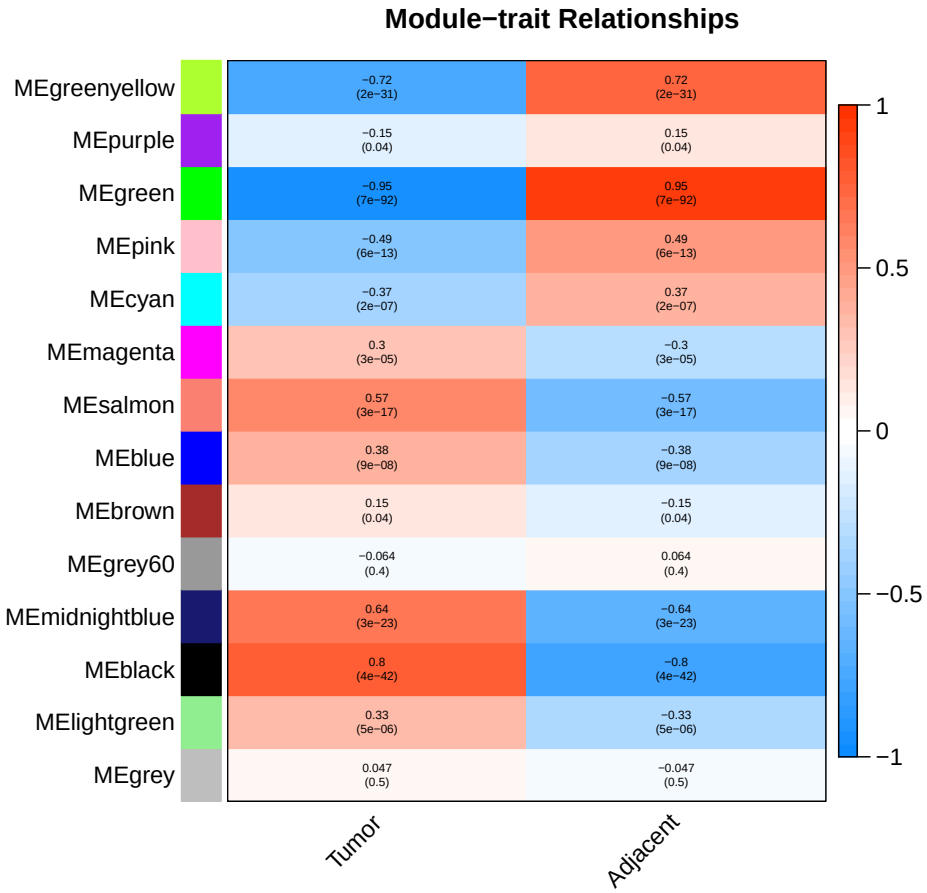

Figure 4.10: Module-trait Relationship for KIRC

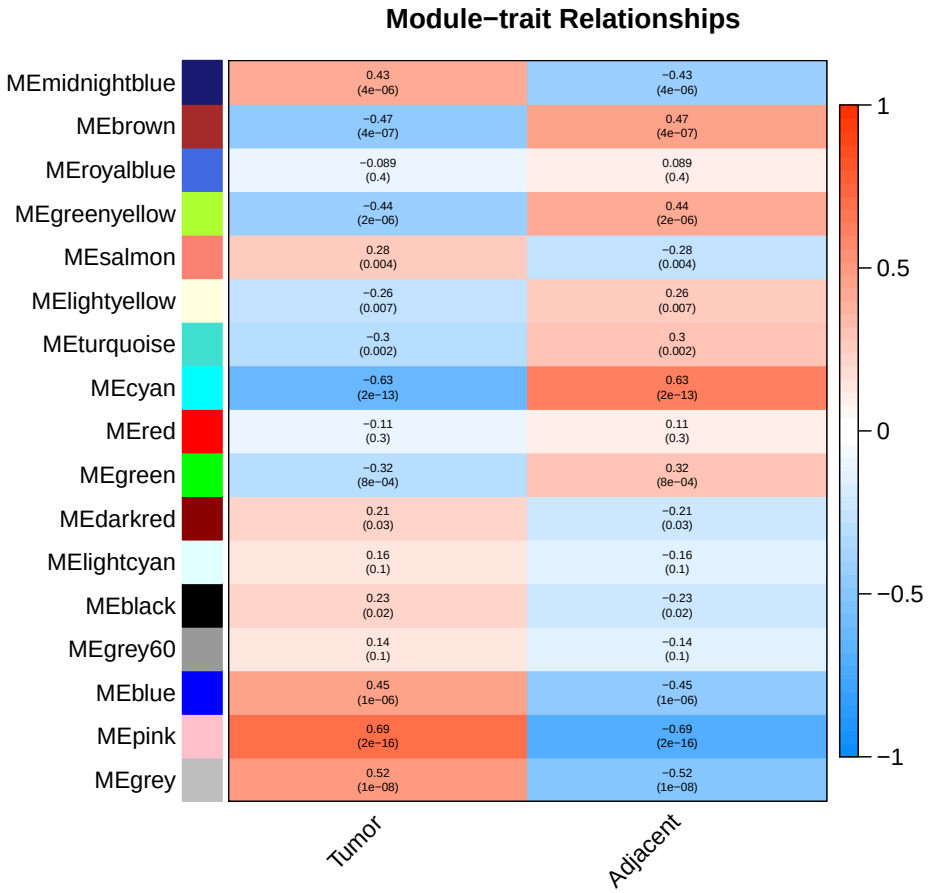

Figure 4.11: Module-trait Relationship for STAD

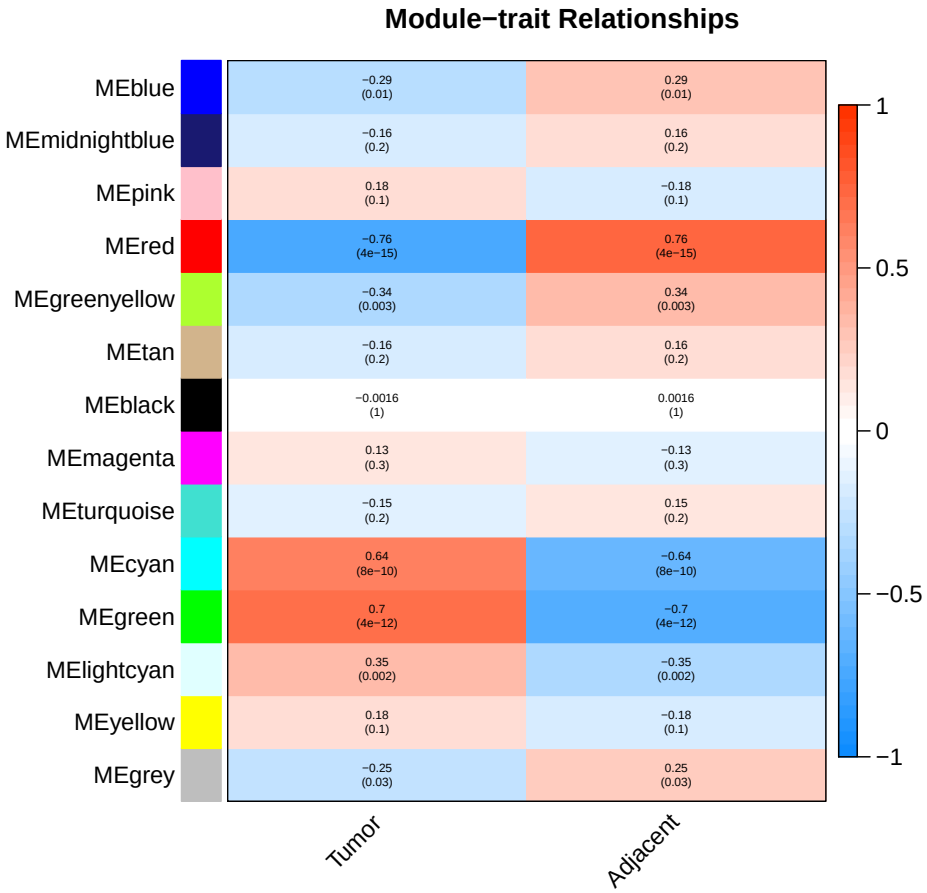

Figure 4.12: Module-trait Relationship for ESCA

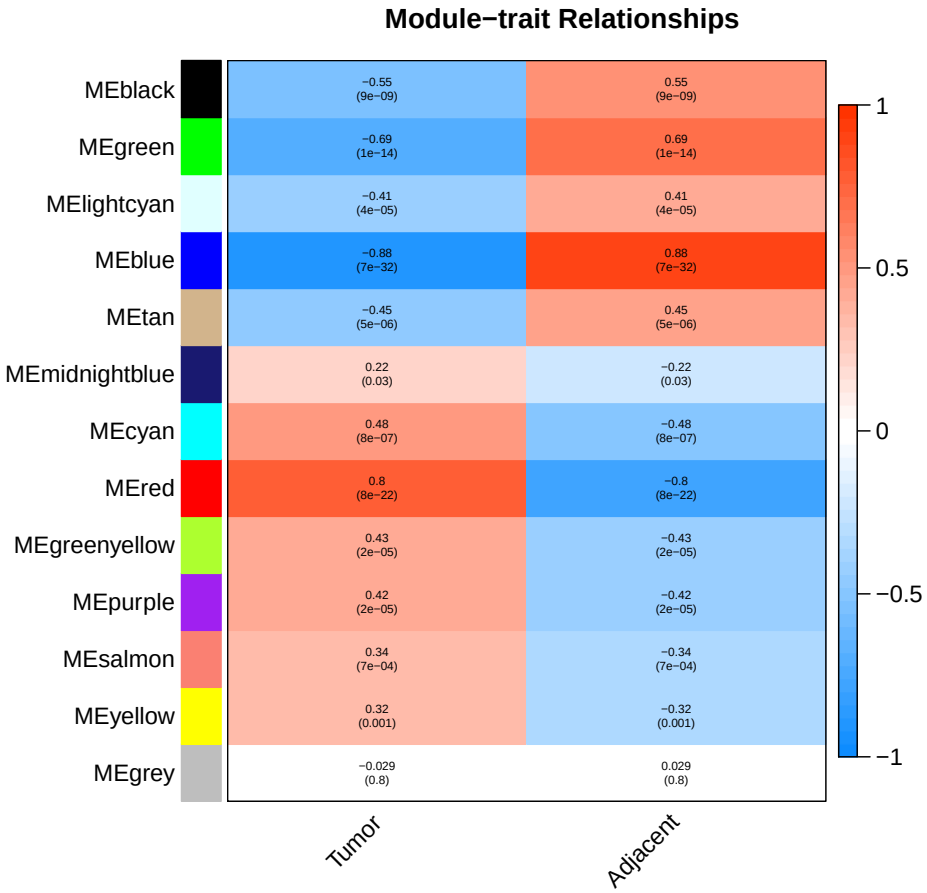

Figure 4.13: Module-trait Relationship for COAD

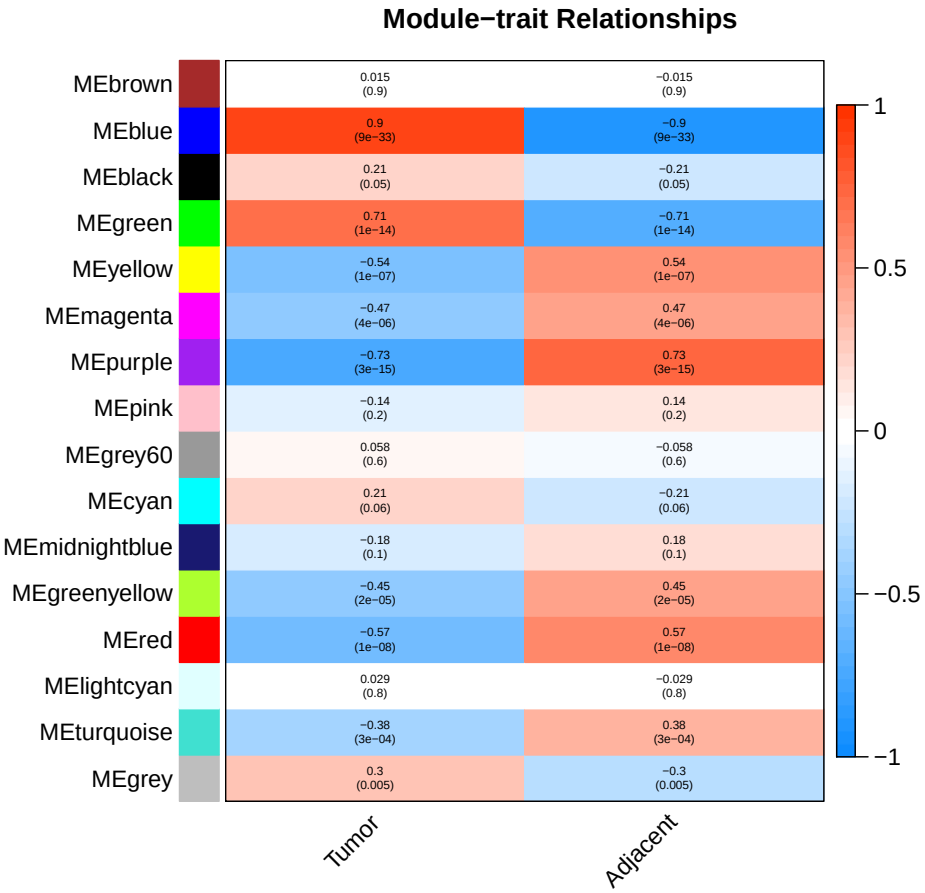

Figure 4.14: Module-trait Relationship for HNSC

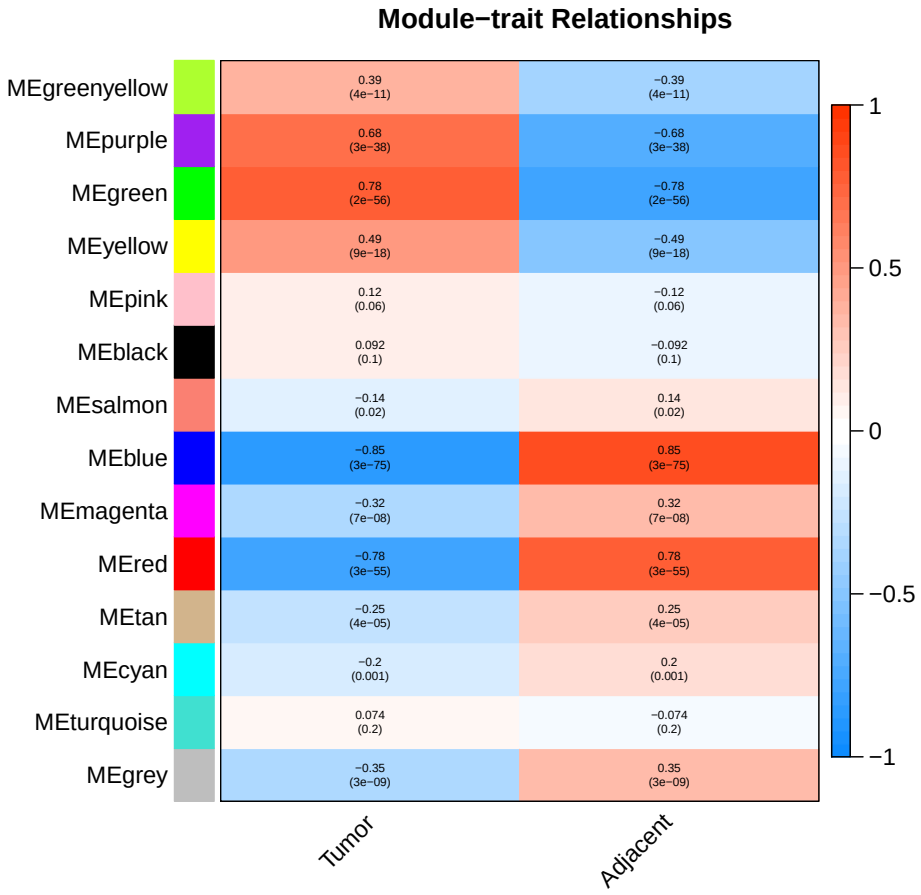

Figure 4.15: Module-trait Relationship for BRCA

Moreover, heatmap and eigengene network for module-trait association are given in Figures A.21 - A.40 for individual cancer types. Only one module was obtained for each cancer dataset except for LUSC which produced two interesting modules equally correlated with tumor status.

Finally, intramodular analysis was conducted and the association between Module membership and Gene significance was visualised for each selected module as shown in Figures 4.16 - 4.25.

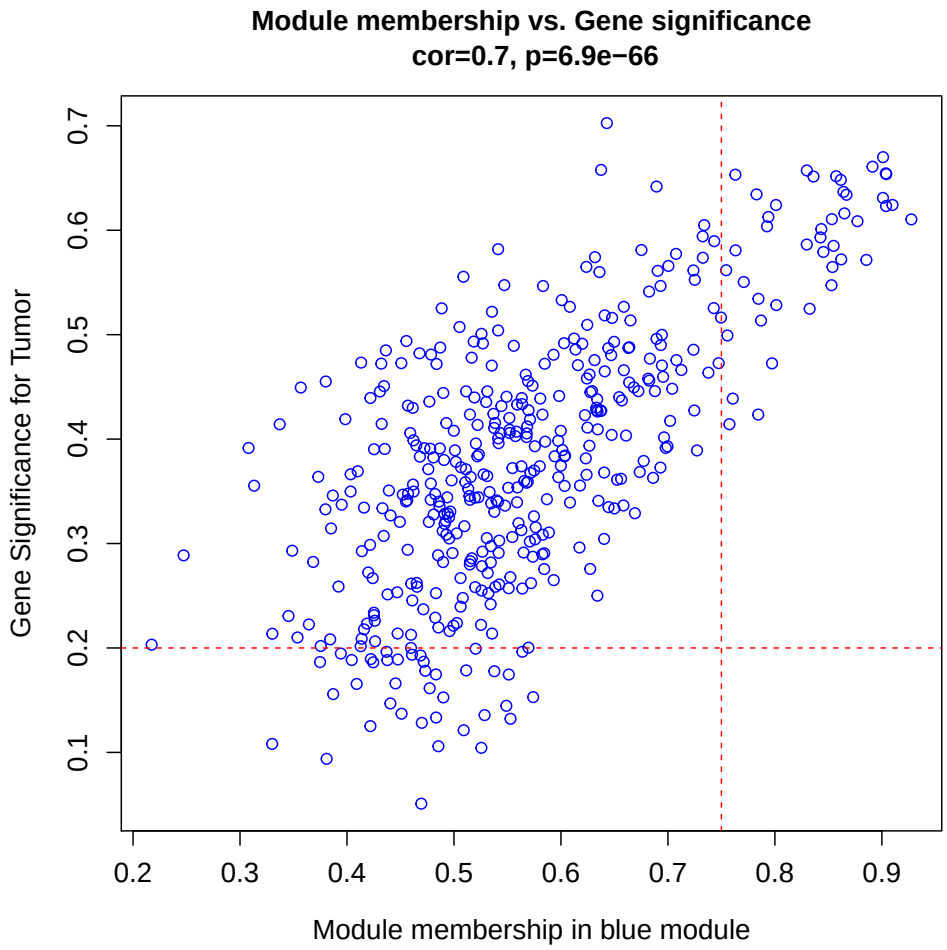

Figure 4.16: Module membership in the selected module for BLCA

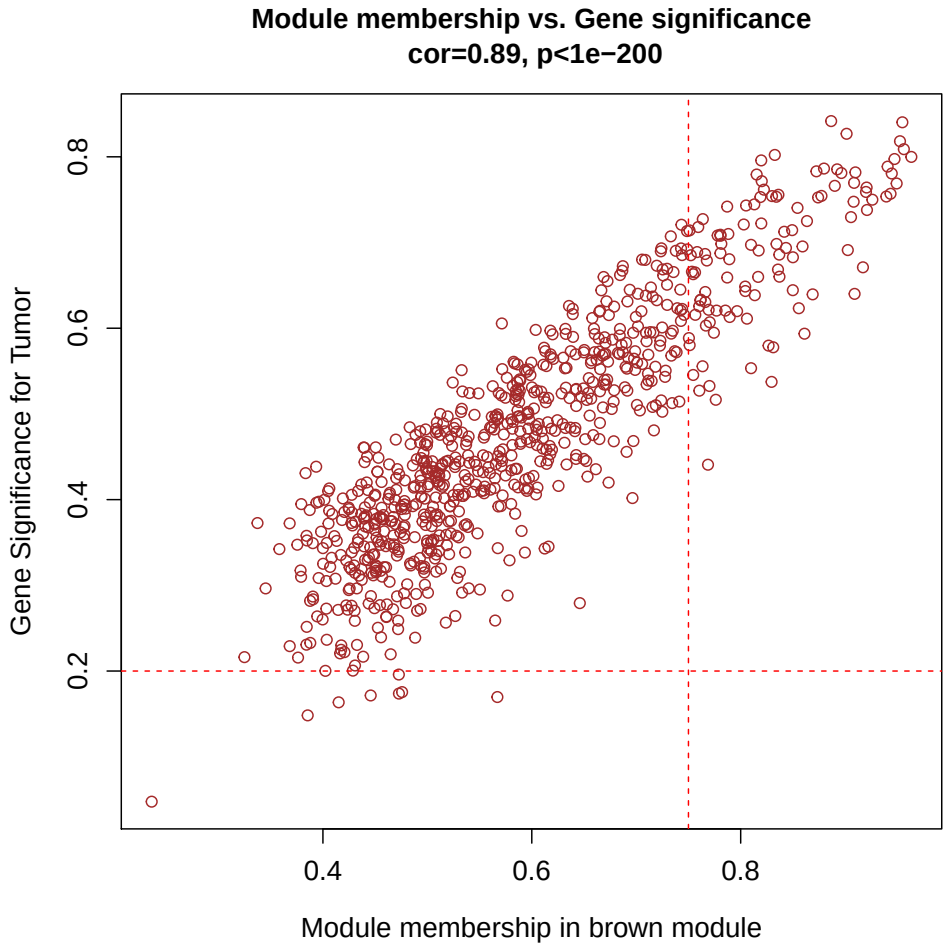

Figure 4.17: Module membership in the selected module for LUAD

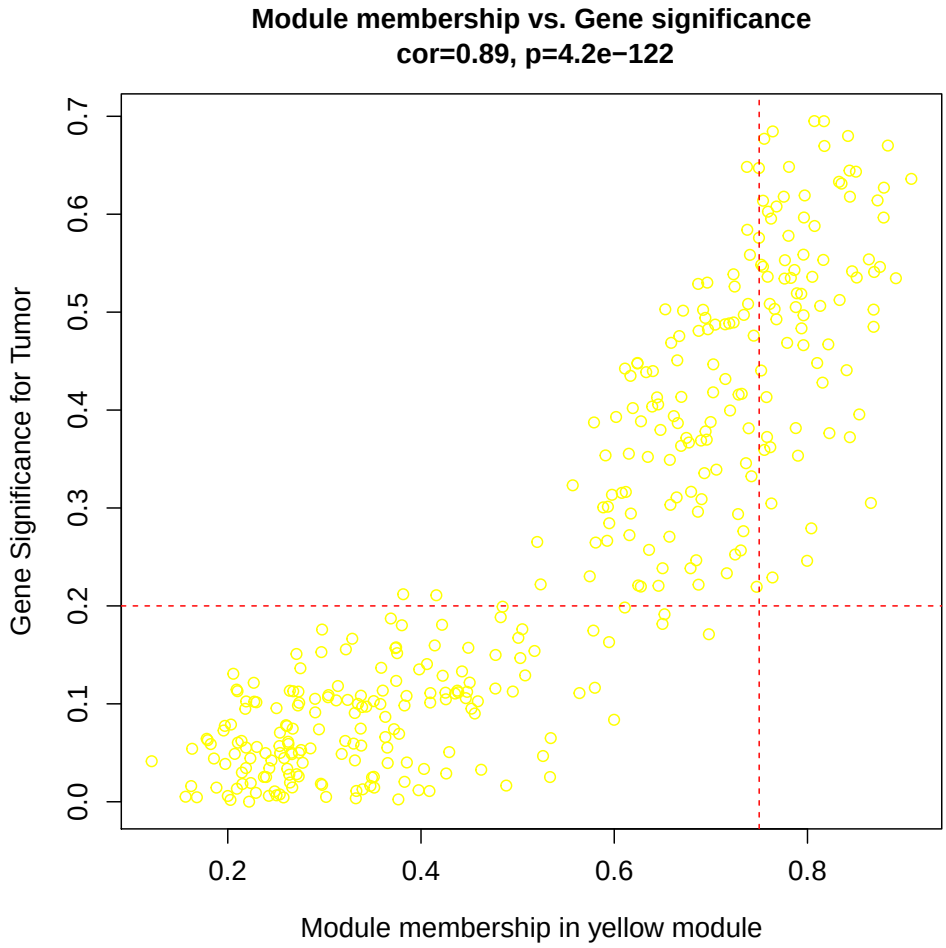

Figure 4.18: Module membership in the selected module for LUSC

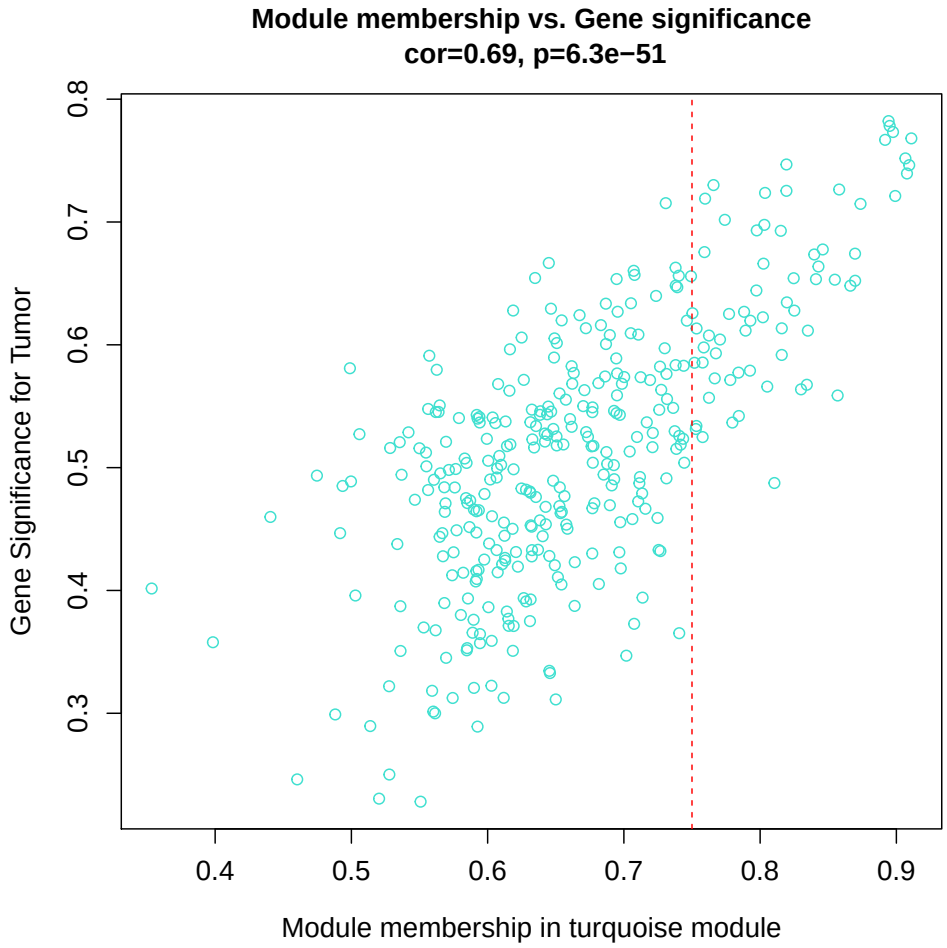

Figure 4.19: Module membership in the selected module for LIHC

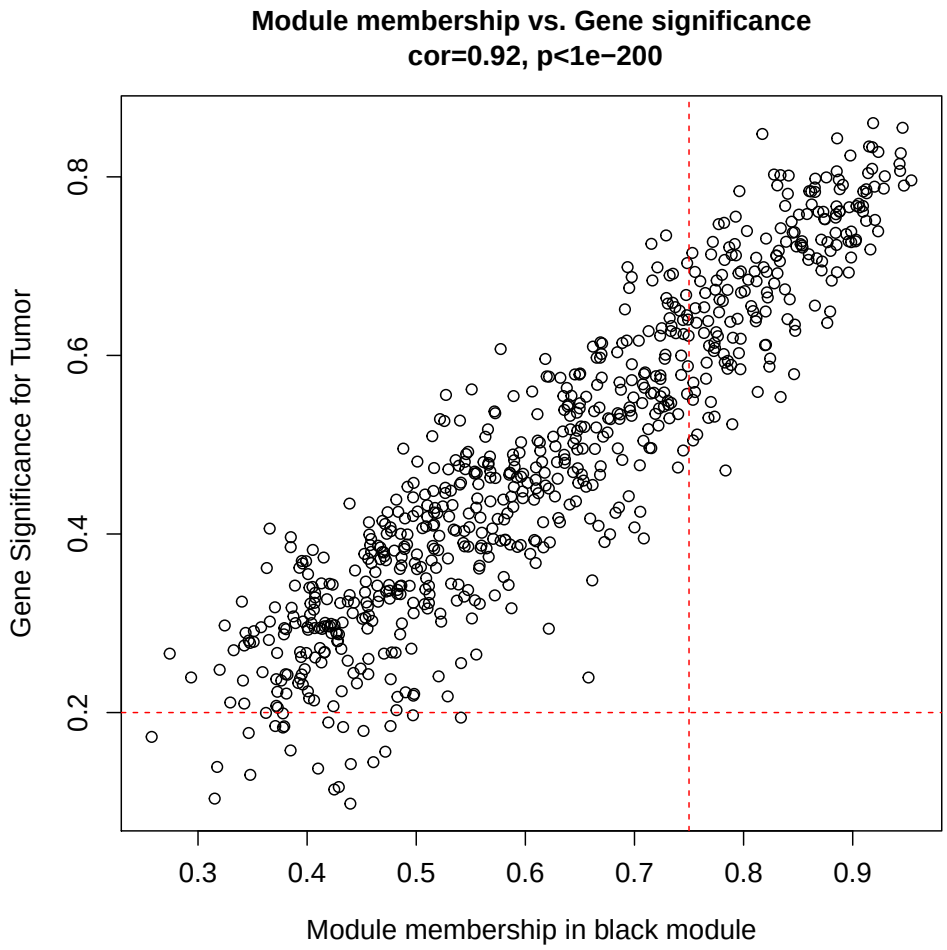

Figure 4.20: Module membership in the selected module for KIRC

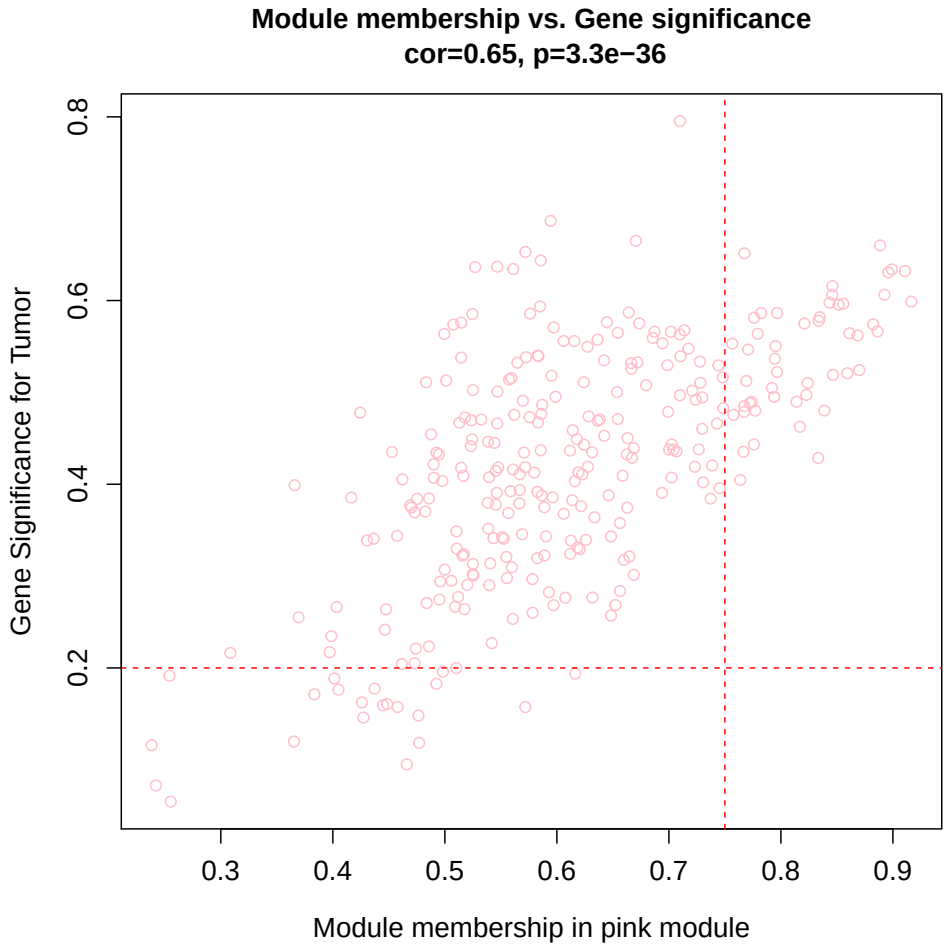

Figure 4.21: Module membership in the selected module for STAD

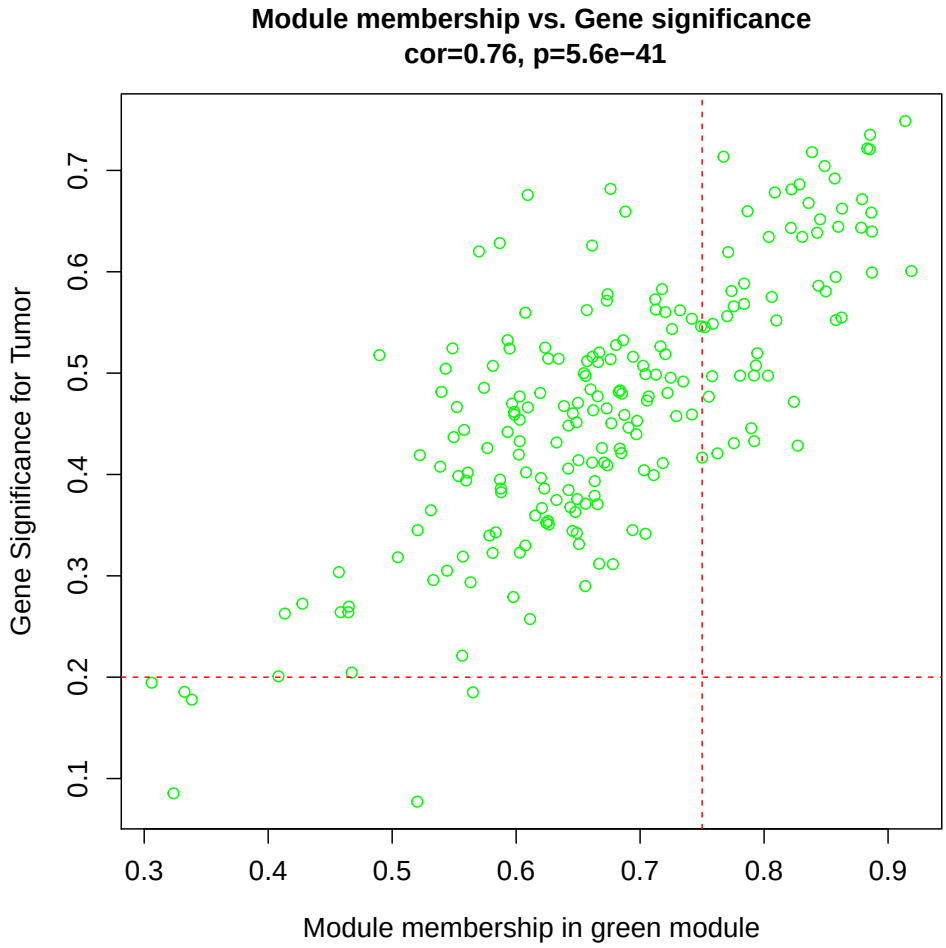

Figure 4.22: Module membership in the selected module for ESCA

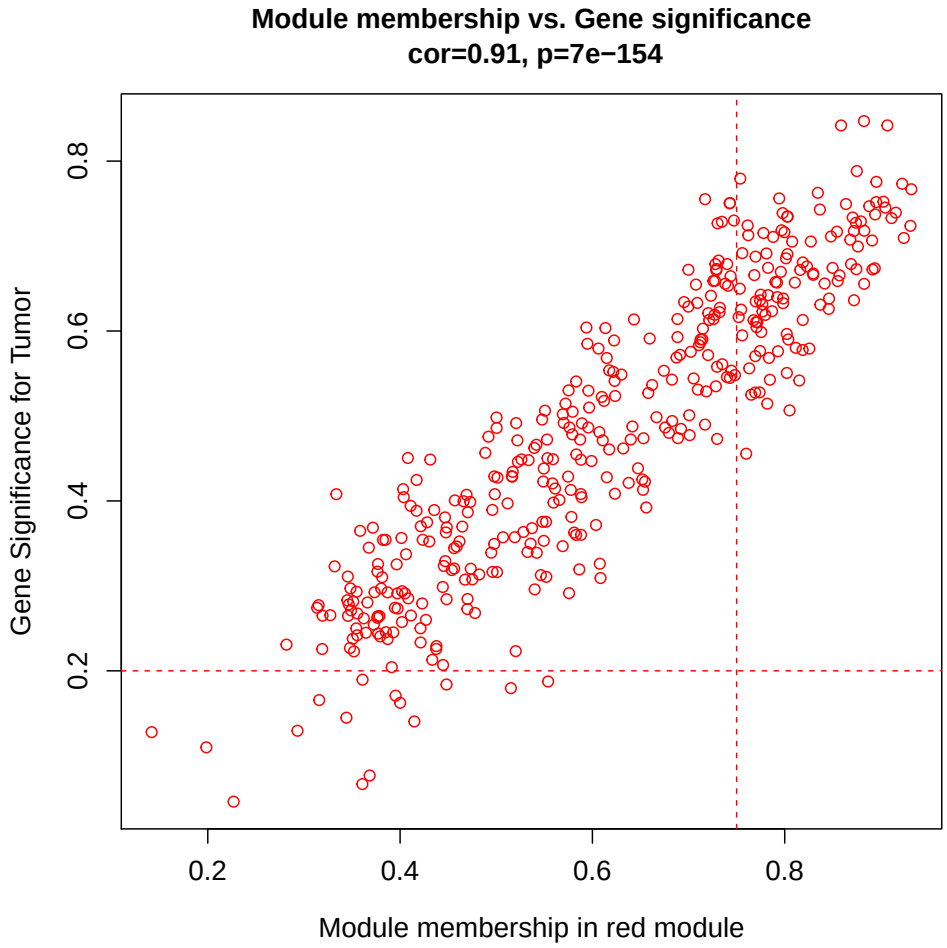

Figure 4.23: Module membership in the selected module for COAD

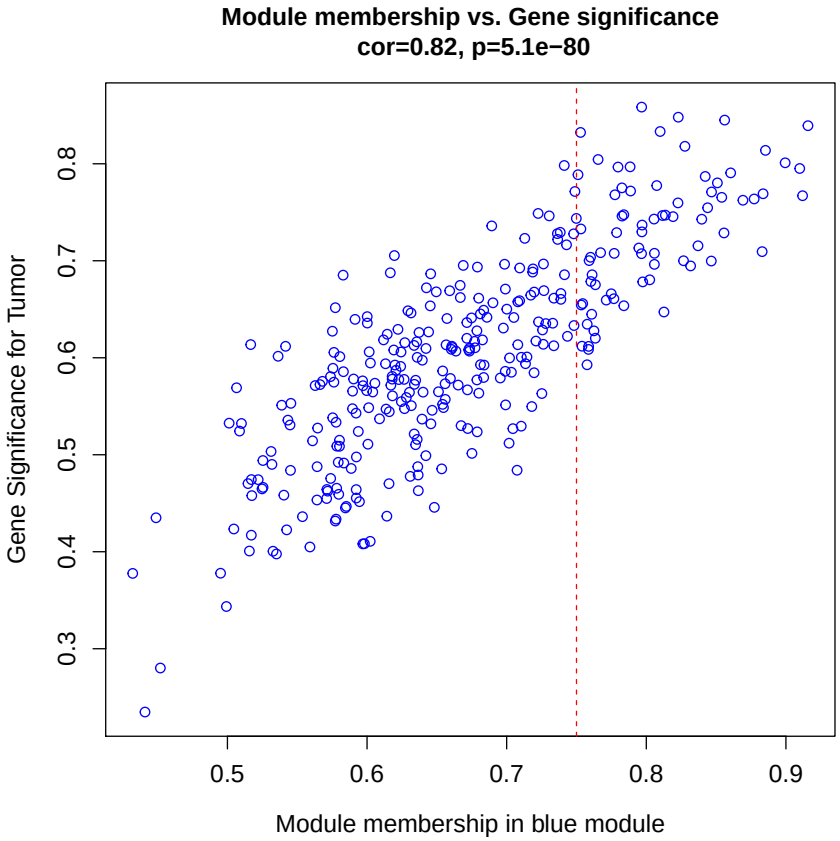

Figure 4.24: Module membership in the selected module for HNSC

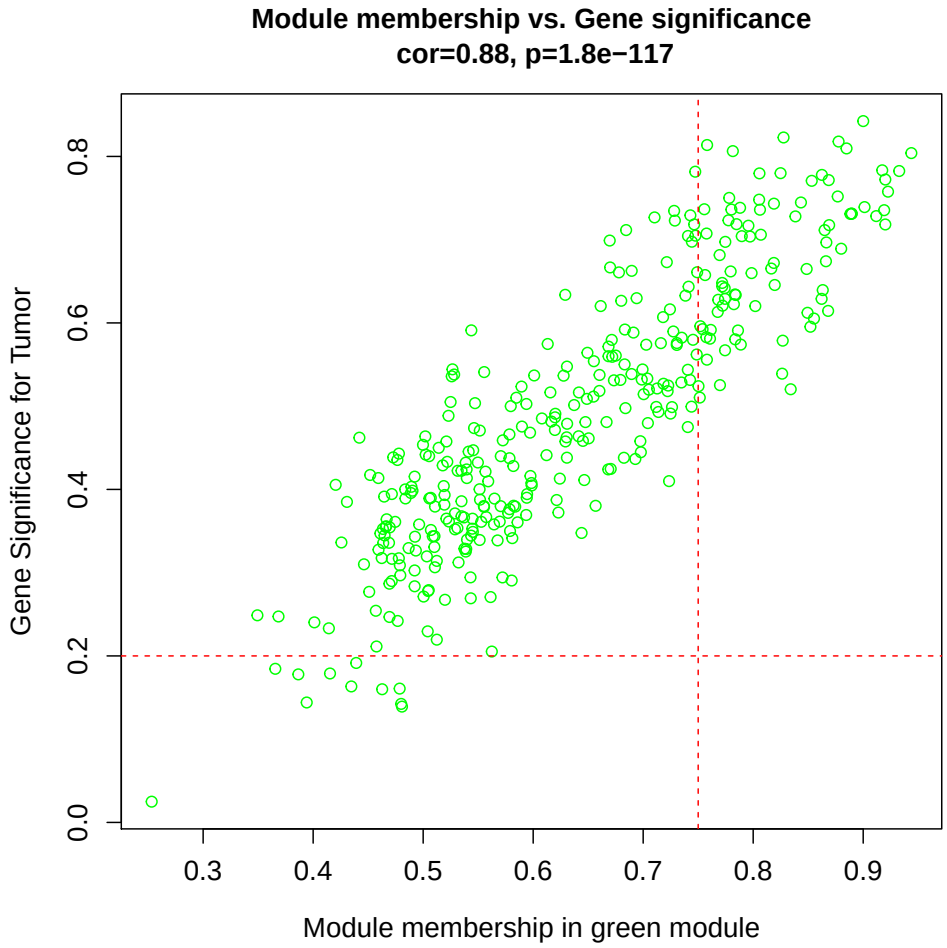

Figure 4.25: Module membership in the selected module for BRCA

A strong and statistically significant (p-value<0.05) correlation between Module membership vs. gene significance of the selected modules was expected. Selected modules along with their properties are provided in Table 4.8.

The aim was to assess the degree of association between these two parameters. A robust and statistically significant correlation, as denoted by a p-value less than 0.05, was observed. It signifies the extent to which the membership of genes within a given module aligns with their biological significance in the context of the studied phenomenon. In other words, the graphs show that the selected modules play a pivotal and statistically validated role in cancer.

Table 4.8: Selected WGCNA modules together with their properties

Cancer	Module	Genes in selected module	Module-trait Correlation	Intramodular GS-MM Correlation	Module hubGenes
BLCA	blue	439	0.68	0.7	43
LUAD	brown	794	0.81	0.89	94
LUSC	yellow	354	0.53	0.89	76
LIHC	turquoise	351	0.77	0.69	38
KIRC	black	231	0.8	0.92	203
STAD	pink	290	0.69	0.65	48
ESCA	green	211	0.7	0.76	54
COAD	red	399	0.8	0.91	120
HNSC	blue	324	0.9	0.82	53
BRCA	green	359	0.78	0.88	82

4.4.3 Hub Genes Extraction

After getting co-expressed genes from WGCNA, hub genes were extracted based on gene significance and module membership value of genes. For individual modules, hub genes were extracted using criteria: |GS|>=0.2, MM>=0.75. The hub genes selection criterion resulted in 38 to 203 hub genes from each module for individual cancer types. The perimeters and measures for individual cancer sets are shown in the Table 4.8.

In total, 811 hub genes were extracted with 488 unique genes. Only 122 hub genes showed re-occurrence across cancers. The frequency of hub genes is shown in Figure 4.26. As evident from the bar chart, only 68 hub genes were present in 3 or more different TCGA cancers.

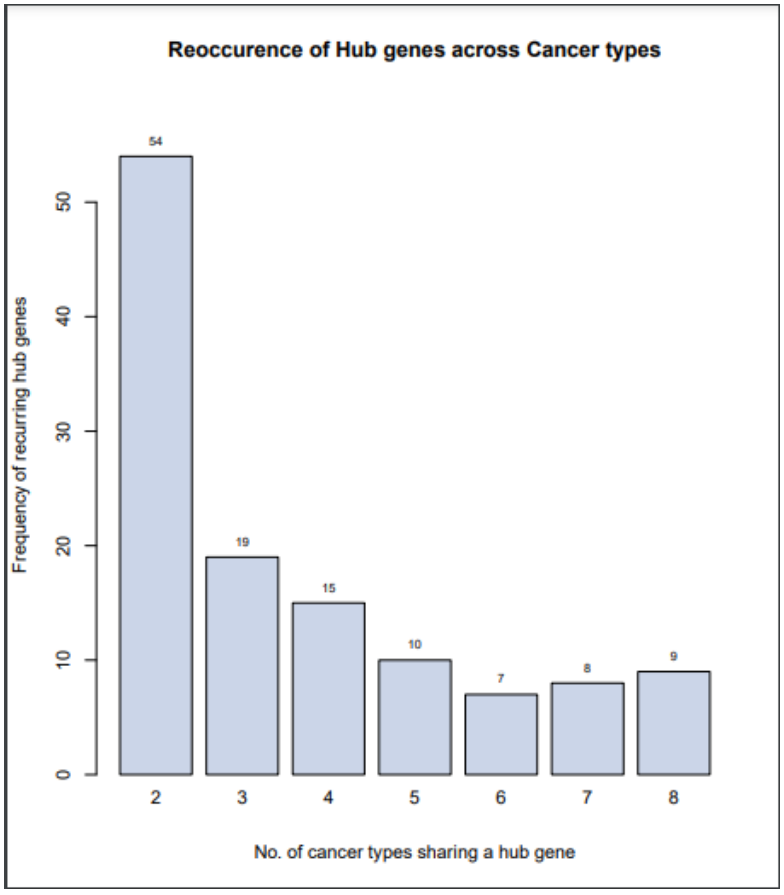

Figure 4.26: Recurrence Frequency Bar-plot of Hub Genes

4.5 Selection of Pan-cancer-wide genes

Pan-cancer-wide hub genes were extracted by integrating switching hub genes of individual cancer types. Intersection operation between hub genes and switching genes for individual cancers resulted in integrated 176 hub genes, out of which only 55 possessed NMD-sensitive transcripts. Out of the 68 pan-cancer-wide (>3 cancers) hub genes, only the switching hub genes that reoccurred in 3 or more cancer types were selected. The selection resulted in 20 hub genes that showed isoform switching in either paired or unpaired analysis. The commonality of hub genes across cancer types Figure 4.27. These pan-cancer-wide switching hub genes were extracted and used for further analysis. The list of the genes is given in Table 4.9.

Table 4.9: Pan-cancer-wide switching hub genes

Gene Name	
LIMS3	ZNRD1
C19orf43	BRCA1
ADAMTSL1	CDK20
FKBP8	COPB1
GGA1	FAM136A
FUNDC2P2	NEK9
MCAT	PAH
PSMA3	PPP1R9A
ST6GALNAC4	RPL23AP7
UQCRC1	TIMM17B

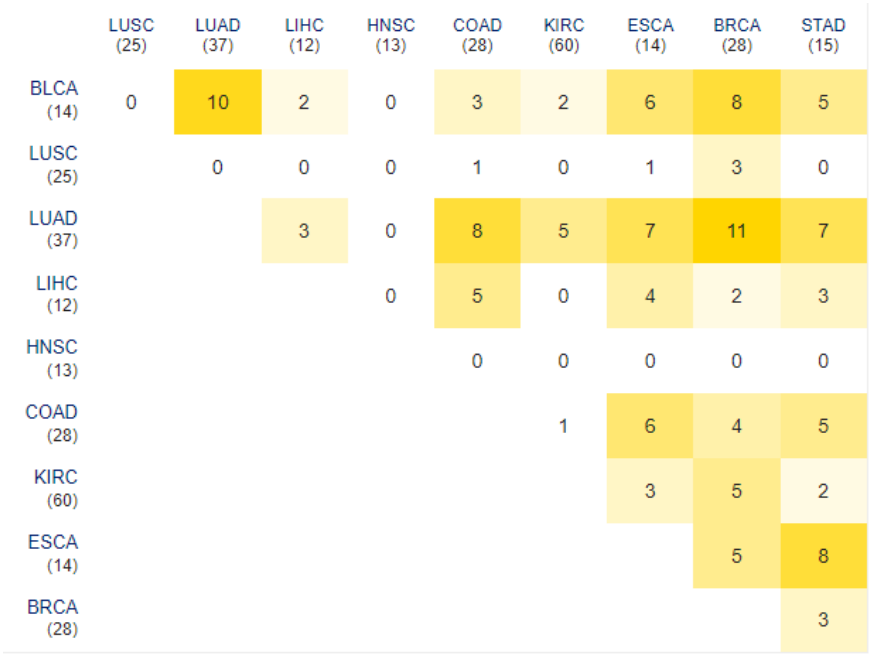

Figure 4.27: Heatmap showing commonality of hub genes across cancer types

4.6 NMD genes

Out of the selected 20 PANCAN switching hub genes, only 6 genes (C19orf43, ZNRD1, CDK20, FAM136A, LIMS3, and TIMM17B) involved NMD-sensitive transcripts during isoform switching. In other words, these genes showed an iso-

form switch of PTC+ transcript with protein-coding isoform.

Among the six genes under scrutiny, three genes exhibited a positive differential isoform fraction (dIF) for the NMD-sensitive transcript. In contrast, the remaining three genes demonstrated a divergent pattern, indicating a contrary behavior in terms of isoform regulation. The list of the selected NMD-sensitive switching hub genes along with their frequency across cancer types and isoform expression is shown in Table 4.10.

Table 4.10: NMD-sensitive Pan-cancer wide switching hub genes

Gene	Frequency	Present In	PTC+ transcript expression
C19orf43	6	BLCA, LUAD, STAD, KIRC, ESCA, BRCA	Down-regulated
ZNRD1	4	BLCA, LUAD, COAD, BRCA	Down-regulated
CDK20	3	BLCA, ESCA, BRCA	Down-regulated
FAM136A	3	LUAD, KIRC, ESCA	Up-regulated
LIMS3	3	LIHC, STAD, BRCA	Up-regulated
TIMM17B	3	LIHC, STAD, ESCA	Up-regulated

4.7 AS-NMD Patterns of Selected Genes

An extensive literature search was utilised to find out the role of the given genes in cancer and whether they acted as TSG or oncogene. Isoform switching pattern, NMD susceptibility, and phenotypic role of each gene were studied with respect to cancer. Lastly, AS-NMD and switching patterns of genes of interest were studied using ISAR switch plots illustrated in Figures 4.28 - 4.33.

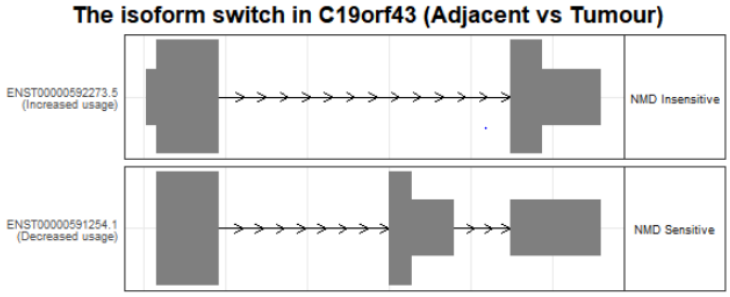

Figure 4.28: Switch Plot of C19orf43

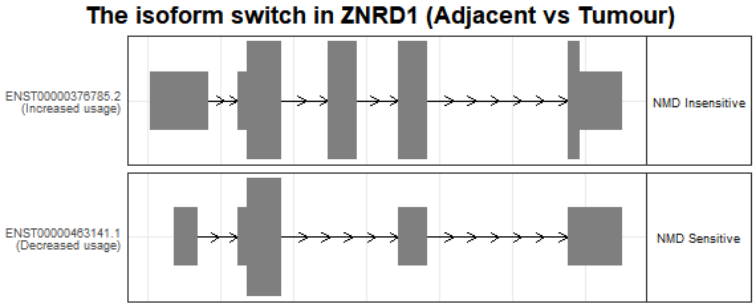

Figure 4.29: Switch Plot of ZNRD1

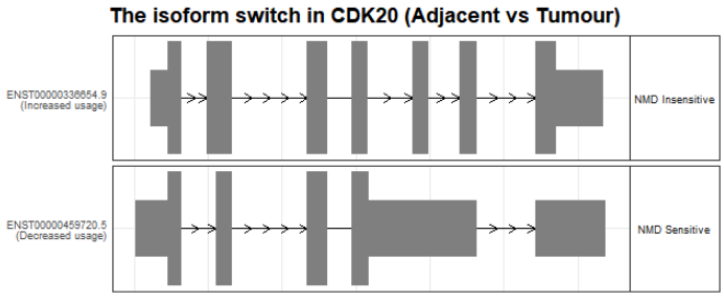

Figure 4.30: Switch Plot of CDK20

As evident from the switch plots (illustrated in Figures 4.28, 4.29 and 4.30), C19orf43, CDK20 and ZNRD1 genes showed significant isoform switches. In all three cases, the switching was characterised by a notable decrease in the isoform usage of the NMD-sensitive transcript, accompanied by a corresponding increase in the usage of the protein-coding transcript. On the contrary, the other 3 genes (FAM136A, LIMS3, and TIMM17B) showed the opposite transition. In other words, the isoform switch in each of these genes was marked by a significant increase in isoform usage of the NMD-sensitive transcript, accompanied by a cor-

responding decrease in the usage of the protein-coding transcript as shown in Figures 4.31, 4.32 and 4.33.

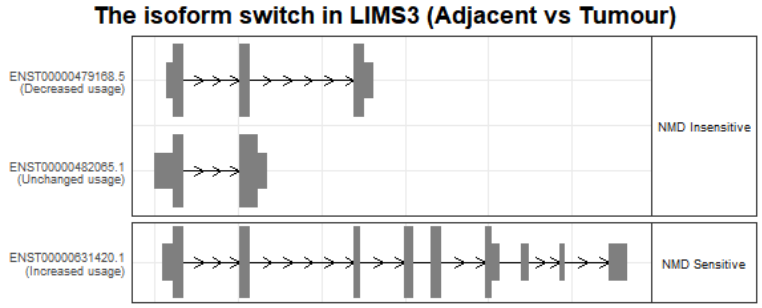

Figure 4.31: Switch Plot of LIMS3

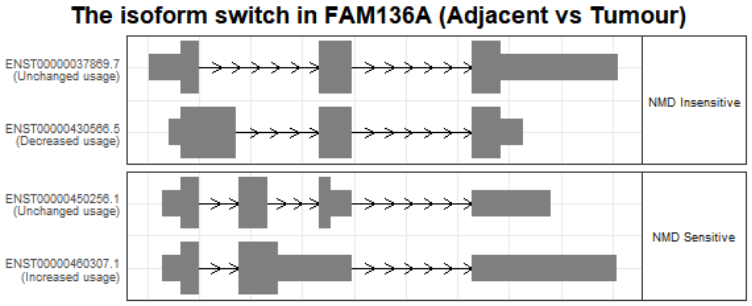

Figure 4.32: Switch Plot of FAMM136A

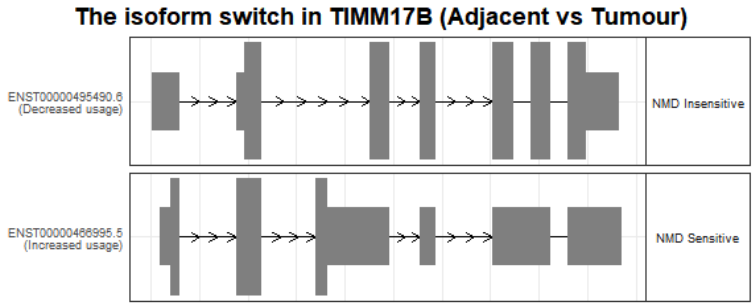

Figure 4.33: Switch Plot of TIMM17B

Such intriguing shifts in isoform usage suggest a dynamic regulatory mechanism that potentially influences the functional role of these genes in cellular processes. Further exploration of these isoform switches may unveil novel insights into the molecular mechanisms governing gene expression and their implications in physiological and pathological conditions.

4.8 Functional Analysis

The selected pan-cancer-wide switching hub genes (57 in total) were subjected to functional analysis using gprofiler2 and enriched GO terms and pathways were visualised. Protein binding activity and protein metabolic pathways were found to be among the enriched GO terms. Besides, enrichment was observed in REACTOME pathways associated with DNA repair mechanisms. The top 10 enriched pathways are shown in Figure 4.34.

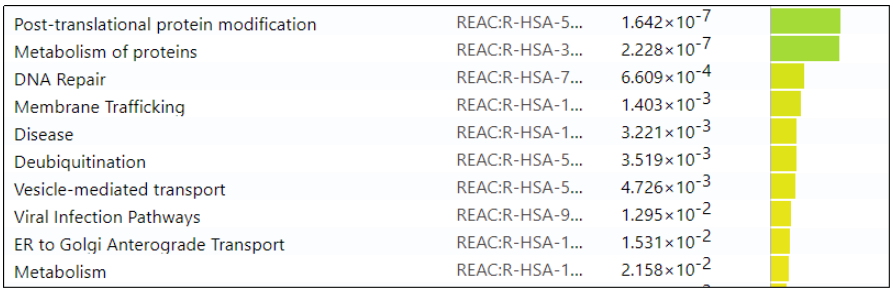

Figure 4.34: Enriched REACTOME Terms

CHAPTER 5: DISCUSSION

Epithelial cancers are the most prevalent among all cancers causing an increasing number of cancer-related deaths globally [87]. Cancer treatment has emerged as a serious challenge due to the heterogeneous nature of cancer as well as the complexity of the plethora of mechanisms associated with tumor progression [88]. In this study, we explored the multifaceted role of Nonsense-Mediated mRNA Decay (NMD) in epithelial cancers. Generally, nonsense-mediated mRNA decay (NMD) is an eukaryotic post-transcriptional surveillance and gene expression regulation mechanism that selectively degrades aberrant mRNA transcripts. Moreover, cancer-related pathways are frequently targeted by alternative splicing events leading to isoform switching. Alternative splicing coupled with NMD (AS-NMD) maintains the integrity of transcriptome and fine-tunes gene expression levels under normal cellular conditions. However, disruption in the normal execution of this pathway can lead to many diseases including cancer [43, 45, 49].

Although the role of NMD in cancer is important, its contribution to the disease is complex [50]. On the one hand, tumors down-regulate NMD factors in order to grow, survive, and evolve. Reports show that inhibition of NMD results in the up-regulation of certain growth factors, immune infiltration, and consequently metastasis. Moreover, NMD degrades many PTC-harbouring dominant-negative TSGs which can otherwise antagonise with the wild-type genes and compromise their role as tumor-suppressors [51, 52, 53, 54]. On the other hand, NMD aggravates cancer by inactivating or disrupting tumor-suppressor genes in many ways (adding NMD-resistant mutations in dominant-negative alleles, biallelic inactivation, NMD-stimulating mutations in wildtype alleles), making NMD a tumor-promoting mechanism [55, 56, 57, 58]. The conflict can be resolved by exploring the transcriptional and translational profiles of various cancers in more depth. Distinguishing whether NMD has a tumor-suppressing or tumor-promoting role in cancer is important for designing therapeutic targets against cancer. Our pri-

mary research objective was to elucidate the impact of AS-NMD on cancer progression, shedding light on whether NMD acts as an aggravating force through the degradation of tumor suppressor genes or as a suppressive mechanism by targeting oncogenes.

Our study delved into the role of Nonsense-Mediated mRNA Decay (NMD) in epithelial cancers by investigating isoform switches across multiple carcinomas. A key focus was to analyse the splicing patterns of NMD-sensitive genes and their association with functional consequences related to cancer predisposition or inhibition. For this purpose, transcript expression and gene expression data were extracted from the TOIL RSEM and TCGA PANCAN datasets, respectively. Individual cancer datasets for selected cancer types (BLCA, LUSC, LUAD, LIHC, KIRC, COAD, ESCA, HNSC, STAD, and BRCA) were extracted, pre-processed and annotated in order to perform isoform usage analysis (using isoformSwitchAnalyzeR) as well as gene co-expression analysis (using WGCNA) on transcript expression and gene expression data, respectively. After extraction of hub genes based on high module connectivity and adequate clinical trait relationship in WGCNA, switching hub genes common across 3 or more cancers were extracted resulting in 55 PANCAN switching hub genes. It was further explored that only 28 genes showed significant NMD sensitivity with at least one PTC-containing transcript showing switching with the protein-coding transcript.

Eventually, 6 NMD-sensitive hub genes exhibiting isoform switching that recurred in 3 or more cancers were extracted and in order to investigate the role of NMD in such genes, the switching pattern of the genes was observed as well as their role (i.e. oncogene or TSG) in cancer was studied. C19orf43, ZNRD1, and CDK20 possessed down-regulated PTC+ transcript being switched with protein-coding isoform. While, FAM136A, TIMM17B, and LIMS3 possessed up-regulated PTC+ transcript being switched with protein-coding isoform (with the exception of FAM136A showing negatively regulated PTC+ transcript in ESCA).

CDK20 (Cyclin-Dependent Kinase 20) is one of the most important regulators of cell cycle checkpoints, involved in major cell cycle regulatory pathways including Wnt, EZH2/NF-B, and other signaling pathways [89, 90]. Reports suggest the

involvement of CDK20 in cell proliferation via activation of CDK2 or increasing cyclin D expression, both of which promote cell growth and survival [91, 92]. Moreover, CDK20 has been reported to aid in chemotherapy resistance in lung cancer [93]. It has been suggested as one of the most crucial factors that contribute to cell survival and growth in other cancer types as well [94]. Concisely, the role of CDK20, in general terms, is tumorigenic.

ZNRD1 is a transcription-related gene that shows tumor-suppressing effects by disrupting the cell cycle and hindering cell growth [95]. It has been reported to be downregulated in a few cancers including esophageal cancer and gastric cancer [96, 97]. ZNRD1 suppresses cell growth by halting the G1 to S phase progression of the cell cycle via inhibition of transcriptional activity of cyclin D1 and CDK4 as well as upregulation of other apoptosis-inducing factors [98]. Moreover, knockdown of ZNRD1-AS1, the antisense lncRNA of ZNRD1, which shows expression opposite to that of ZNRD1 gene has been reported to help inhibit tumor proliferation [99]. Conclusively, findings suggest a tumor suppressor function for the ZNRD1 gene and a tumor contributor function for lncRNA ZNRD1-AS1 in the process of carcinogenesis [100].

FAM136A is a mitochondrial protein encoding gene that has been speculated to influence CDK4/6 expression thus playing a crucial role in tumor proliferation, particularly in lung cancer. Although the exact mechanism of action of the gene is unclear, it is regarded as a crucial prognostic factor in lung cancer. FAM136A is highly expressed in tumor cells [101]. FAM136A is also a target of the Myc gene which is a positive regulator of G1-specific CDKs. Myc gene activates CDK4/6, resulting in cell growth and survival. It has been speculated that Myc might drive FAM136A and Myc- FAM136A- CDK4/6 might be a new signalling pathway in lung cancer regulation. However, it has not been demonstrated yet. Nevertheless, its knockdown has been speculated to suppress cancer progression by inducing apoptosis [102].

TIMM17B gene encodes for translocase protein 17B found in the inner mitochondrial membrane. This transmembrane protein is an integral component of a mitochondrial protein complex. TIMM17B gene is not commonly reported in the

context of cancer. However, recent research sheds light on its potential role in cancer. TIMM17B is differentially expressed in breast cancer and its expression level is inversely linked with survival rate in breast cancer patients [103]. Moreover, it has been reported that the expression level of TIMM17B is directly related to negative immune infiltration and positive drug resistance in breast cancer, suggesting the potential oncogenic role of the gene [104]. Although the exact mechanism of action of the gene in the context of cancer is unknown, it has been speculated that it might affect tumor progression by influencing the cell cycle.

C19orf43 is a Telomerase RNA component interacting RNase (TRIR). Although C19orf43 has not been associated with cancer directly, reports suggest a role of TRIR in melanoma where it exhibits tumor-suppressive effects by inhibiting angiogenesis [105]. However, contrasting perspectives also exist [106]. Likewise, LIMS3 has not been directly linked with cancer. Conclusively, the role of these two genes in cancer remains unknown and unclear.

Based on a comprehensive review of the existing literature and the results obtained from our study, there is compelling evidence to suggest that Alternative Splicing- coupled with Nonsense-Mediated mRNA Decay (AS-NMD) plays a significant role in the regulatory dynamics of cancer cells. Although contrasting reports suggest a bipolar role of NMD in cancer, above mentioned results show more support for tumor suppressing function of NMD. As evident from the results, AS-NMD decreases isoform usage of tumorigenic genes or oncogenes (in the case of TIMM17B and FAM136A) thus affecting their overall expression. Moreover, isoform usage of TSGs like ZNRD1 is increased by AS-NMD.

However, it is crucial to highlight that our results align with the prevailing perspective that cancer genes possess the capacity to manipulate NMD processes, thereby contributing to tumor progression. This is evidenced by the observed increase in the isoform usage of the protein-coding isoform of CDK20, an oncogene.

This study, in particular, sheds light on the functional outcomes of AS-NMD in cancer. We observed that AS-NMD leads to the up-regulation of functional isoforms associated with Tumor Suppressor Genes (TSGs) while concurrently down-regulating functional isoforms related to Oncogenes. This dual regulatory effect

is noteworthy, however, our study supports the bipolar role of NMD in cancer.

Although our study broadens the understanding of the role of AS-NND in regulating gene expression, certain factors and methodical limitations undermine the generalisability of the results. The first and foremost limitation in the area is heterogeneity and context dependency of cancer. Likewise, certain NMD events are tissue-specific, and tumor-specific and might also depend upon tumor stage and type. In addition to that, other limitations including less availability of long isoform data, lack of complete transcript annotations, and certain methodological constraints also hinder further exploration in the area. Nevertheless, this study plays a crucial role in widening our understanding of the field. Studying cancer stage-wise switching of NMD-sensitive genes in the future might be helpful in understanding the context-dependent role of AS-NMD. The future application of advanced sequencing technologies and the development of novel computational models would enable further investigation into the mechanism of action of NMD in cancer.

CHAPTER 6: CONCLUSION AND FUTURE RECOMMENDATIONS

In conclusion, this study explored the multifaceted role of Nonsense-Mediated mRNA Decay (NMD) in epithelial cancers, addressing the complex and debated nature of NMD's contribution to cancer progression. Despite conflicting perspectives in the literature, our study provides comprehensive evidence supporting a nuanced and context-dependent role of NMD in cancer. Specifically, Alternative Splicing coupled with NMD (AS-NMD) was found to exert a dual regulatory effect, decreasing isoform usage of tumorigenic genes while increasing the expression of functional isoforms associated with Tumor Suppressor Genes (TSGs). This supports the notion of NMD acting as a tumor-suppressive mechanism. However, the study also acknowledges the adaptability of cancer genes in manipulating NMD processes, exemplified by the observed increase in the isoform usage of the functional isoform of an oncogene. This highlights the complexity of the regulatory landscape and the need for a context-specific understanding of NMD's role in cancer.

Despite methodological limitations and the heterogeneity of cancer, our findings contribute valuable insights into the functional outcomes of AS-NMD, emphasising its relevance in shaping the transcriptomic landscape of cancer cells. Moving forward, a thorough exploration of tissue-specific and stage-dependent NMD events, coupled with advancements in sequencing technologies, will deepen our understanding of the nuanced role of NMD in cancer. Utilising advanced techniques such as single-cell RNA sequencing and long-read sequencing will enable a more detailed examination of isoform diversity and alternative splicing patterns across diverse cancer types and stages. Furthermore, expanding research beyond epithelial cancers and integrating multi-omics data in the future can provide a comprehensive view of the impact of NMD on tumorigenesis. Moreover, advancement in computational models will enhance precision in unraveling the complex

molecular dynamics involved. In essence, this ongoing investigation holds the potential to identify targeted therapeutic strategies for more effective therapeutic targets as well as personalised cancer treatments.

REFERENCES

[1] Michel Morange. "The Central Dogma of molecular biology". In: *Resonance* 14.3 (2009), pp. 236–247.

[2] Brenton R Graveley. "Alternative splicing: increasing diversity in the proteomic world". In: *TRENDS in Genetics* 17.2 (2001), pp. 100–107.

[3] Lu Chen, Jaime M Tovar-Corona, and Araxi O Urrutia. "Alternative splicing: a potential source of functional innovation in the eukaryotic genome". In: *International journal of evolutionary biology* 2012 (2012).

[4] Paulo J da Costa, Juliane Menezes, and Luísa Romão. "The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease". In: *The International Journal of Biochemistry & Cell Biology* 91 (2017), pp. 168–175.

[5] Yan Wang et al. "Mechanism of alternative splicing and its regulation". In: *Biomedical reports* 3.2 (2015), pp. 152–158.

[6] Eddo Kim, Alon Magen, and Gil Ast. "Different levels of alternative splicing among eukaryotes". In: *Nucleic acids research* 35.1 (2007), pp. 125–131.

[7] Martin Pohl et al. "Alternative splicing of mutually exclusive exons—A review". In: *Biosystems* 114.1 (2013), pp. 31–38.

[8] Hadas Ner-Gaon et al. "Intron retention is a major phenomenon in alternative splicing in Arabidopsis". In: *The Plant Journal* 39.6 (2004), pp. 877–885.

[9] Supriya Sen. "Aberrant pre-mRNA splicing regulation in the development of hepatocellular carcinoma". In: *Hepatoma Research* 4 (2018), p. 37.

[10] Rotem Sorek, Ron Shamir, and Gil Ast. "How prevalent is functional alternative splicing in the human genome?" In: *TRENDS in Genetics* 20.2 (2004), pp. 68–71.

[11] Cuixia Di et al. “Function, clinical application, and strategies of Pre-mRNA splicing in cancer”. In: *Cell Death & Differentiation* 26.7 (2019), pp. 1181–1194.

[12] Lizhong Ding, Ethan Rath, and Yongsheng Bai. “Comparison of alternative splicing junction detection tools using RNASeq data”. In: *Current genomics* 18.3 (2017), pp. 268–277.

[13] Juan P Romero et al. “Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm”. In: *BMC genomics* 19.1 (2018), pp. 1–14.

[14] Arfa Mehmood et al. “Systematic evaluation of differential splicing tools for RNA-seq studies”. In: *Briefings in bioinformatics* 21.6 (2020), pp. 2052–2065.

[15] Cole Trapnell et al. “Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms”. In: *Nature biotechnology* 28.5 (2010), p. 511.

[16] ZhongFa Zhang et al. “Isoform level expression profiles provide better cancer signatures than gene level expression profiles”. In: *Genome medicine* 5.4 (2013), pp. 1–13.

[17] Charlotte Soneson et al. “Differential transcript usage from RNA-seq data: isoform pre-filtering improves performance of count-based methods”. In: *bioRxiv* (2015), p. 025387.

[18] Héctor Climente-González et al. “The functional impact of alternative splicing in cancer”. In: *Cell reports* 20.9 (2017), pp. 2215–2226.

[19] Pamela A Frischmeyer and Harry C Dietz. “Nonsense-mediated mRNA decay in health and disease”. In: *Human molecular genetics* 8.10 (1999), pp. 1893–1900.

[20] Kim M Keeling and David M Bedwell. “Suppression of nonsense mutations as a therapeutic approach to treat genetic diseases”. In: *Wiley Interdisciplinary Reviews: RNA* 2.6 (2011), pp. 837–852.

[21] Eileen Wagner and Jens Lykke-Andersen. “mRNA surveillance: the perfect persist”. In: *Journal of Cell Science* 115.15 (2002), pp. 3033–3038.

[22] Joshua T Mendell et al. “Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise”. In: *Nature genetics* 36.10 (2004), pp. 1073–1078.

[23] Ashton T Belew, Vivek M Advani, and Jonathan D Dinman. “Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast”. In: *Nucleic acids research* 39.7 (2011), pp. 2799–2808.

[24] JING ZHANG et al. “Intron function in the nonsense-mediated decay of β-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm”. In: *Rna* 4.7 (1998), pp. 801–815.

[25] Zhongxia Yi, Manu Sanjeev, and Guramrit Singh. “The branched nature of the nonsense-mediated mRNA decay pathway”. In: *Trends in Genetics* 37.2 (2021), pp. 143–159.

[26] Tatsuaki Kurosaki, Maximilian W Popp, and Lynne E Maquat. “Quality and quantity control of gene expression by nonsense-mediated mRNA decay”. In: *Nature reviews Molecular cell biology* 20.7 (2019), pp. 406–420.

[27] Eszter Nagy and Lynne E Maquat. “A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance”. In: *Trends in biochemical sciences* 23.6 (1998), pp. 198–199.

[28] Nele Hug, Dasa Longman, and Javier F Cáceres. “Mechanism and regulation of the nonsense-mediated decay pathway”. In: *Nucleic acids research* 44.4 (2016), pp. 1483–1495.

[29] Isao Kashima et al. “Binding of a novel SMG-1–Upf1–eRF1–eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay”. In: *Genes & development* 20.3 (2006), pp. 355–367.

[30] Daniel R Schoenberg and Lynne E Maquat. “Regulation of cytoplasmic mRNA decay”. In: *Nature Reviews Genetics* 13.4 (2012), pp. 246–259.

[31] Tetsuo Ohnishi et al. “Phosphorylation of hUPF1 induces formation of mRNA surveillance complexes containing hSMG-5 and hSMG-7”. In: *Molecular cell* 12.5 (2003), pp. 1187–1200.

[32] Fabrice Lejeune, Xiaojie Li, and Lynne E Maquat. “Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities”. In: *Molecular cell* 12.3 (2003), pp. 675–687.

[33] Belinda Loh, Stefanie Jonas, and Elisa Izaurralde. “The SMG5–SMG7 heterodimer directly recruits the CCR4–NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2”. In: *Genes & development* 27.19 (2013), pp. 2125–2138.

[34] James PB Lloyd. “The evolution and diversity of the nonsense-mediated mRNA decay pathway”. In: *F1000Research* 7 (2018).

[35] Lulu Huang and Miles F Wilkinson. “Regulation of nonsense-mediated mRNA decay”. In: *Wiley Interdisciplinary Reviews: RNA* 3.6 (2012), pp. 807–828.

[36] Liat Linde et al. “The efficiency of nonsense-mediated mRNA decay is an inherent character and varies among different cells”. In: *European journal of human genetics* 15.11 (2007), pp. 1156–1162.

[37] Almoutassem B Zetoune et al. “Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues”. In: *BMC genetics* 9.1 (2008), pp. 1–11.

[38] Liat Linde et al. “Nonsense-mediated mRNA decay affects nonsense transcript levels and governs response of cystic fibrosis patients to gentamicin”. In: *The Journal of clinical investigation* 117.3 (2007), pp. 683–692.

[39] Lulu Huang et al. “RNA homeostasis governed by cell type-specific and branched feedback loops acting on NMD”. In: *Molecular cell* 43.6 (2011), pp. 950–961.

[40] A Sureau et al. “SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs”. In: *The EMBO journal* 20.7 (2001), pp. 1785–1796.

[41] Juan Valcárcel and Fátima Gebauer. “Post-transcriptional regulation: the dawn of PTB”. In: *Current Biology* 7.11 (1997), R705–R708.

[42] Matthew C Wollerton et al. “Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay”. In: *Molecular cell* 13.1 (2004), pp. 91–100.

[43] David AW Soergel, Liana F Lareau, and Steven E Brenner. “Regulation of gene expression by coupling of alternative splicing and NMD”. In: *Madame Curie Bioscience Database [Internet]*. Landes Bioscience, 2013.

[44] Benjamin P Lewis, Richard E Green, and Steven E Brenner. “Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans”. In: *Proceedings of the National Academy of Sciences* 100.1 (2003), pp. 189–192.

[45] Juan F García-Moreno and Luísa Romão. “Perspective in alternative splicing coupled to nonsense-mediated mRNA decay”. In: *International Journal of Molecular Sciences* 21.24 (2020), p. 9424.

[46] Liana F Lareau et al. “The coupling of alternative splicing and nonsense-mediated mRNA decay”. In: *Alternative splicing in the Postgenomic Era* 623 (2007), pp. 190–211.

[47] Liana F Lareau and Steven E Brenner. “Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible”. In: *Molecular biology and evolution* 32.4 (2015), pp. 1072–1079.

[48] Jennifer Winter et al. “Regulation of the MID1 protein function is fine-tuned by a complex pattern of alternative splicing”. In: *Human genetics* 114.6 (2004), pp. 541–552.

[49] Seyed M Ghiasi and Guy A Rutter. “Consequences for Pancreatic β-Cell Identity and Function of Unregulated Transcript Processing”. In: *Frontiers in Endocrinology* 12 (2021), p. 111.

[50] Gonçalo Nogueira et al. “Nonsense-mediated RNA decay and its bipolar function in cancer”. In: *Molecular Cancer* 20.1 (2021), pp. 1–19.

[51] Lei Chang et al. “The human RNA surveillance factor UPF1 regulates tumorigenesis by targeting Smad7 in hepatocellular carcinoma”. In: *Journal of Experimental & Clinical Cancer Research* 35.1 (2016), pp. 1–12.

[52] Chen Liu et al. “The UPF1 RNA surveillance gene is commonly mutated in pancreatic adenosquamous carcinoma”. In: *Nature medicine* 20.6 (2014), pp. 596–598.

[53] Saijun Fan et al. “Mutant BRCA1 genes antagonize phenotype of wild-type BRCA1”. In: *Oncogene* 20.57 (2001), pp. 8215–8235.

[54] Maximilian W Popp and Lynne E Maquat. “Nonsense-mediated mRNA decay and cancer”. In: *Current opinion in genetics & development* 48 (2018), pp. 44–50.

[55] Zhiyuan Hu, Christopher Yau, and Ahmed Ashour Ahmed. “A pan-cancer genome-wide analysis reveals tumour dependencies by induction of nonsense-mediated decay”. In: *Nature communications* 8.1 (2017), pp. 1–9.

[56] Rik GH Lindeboom, Fran Supek, and Ben Lehner. “The rules and impact of nonsense-mediated mRNA decay in human cancers”. In: *Nature genetics* 48.10 (2016), pp. 1112–1118.

[57] Fernando Pastor et al. “Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay”. In: *Nature* 465.7295 (2010), pp. 227–230.

[58] Fran Supek, Ben Lehner, and Rik GH Lindeboom. “To NMD or Not To NMD: nonsense-mediated mRNA decay in cancer and other genetic diseases”. In: *Trends in Genetics* 37.7 (2021), pp. 657–668.

[59] Kristian M Hargadon, Coleman E Johnson, and Corey J Williams. “Immune checkpoint blockade therapy for cancer: An overview of FDA-approved immune checkpoint inhibitors”. In: *International immunopharmacology* 62 (2018), pp. 29–39.

[60] Mike Bogetofte Barnkob, Kristoffer Vitting-Seerup, and Lars Rønn Olsen. “Target isoforms are an overlooked challenge and opportunity in chimeric antigen receptor cell therapy”. In: *Immunotherapy Advances* 2.1 (2022), ltac009.

[61] Zhi-xiang Lu et al. “Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization”. In: *Molecular Cancer Research* 13.2 (2015), pp. 305–318.

[62] Endre Sebestyén, Michał Zawisza, and Eduardo Eyras. “Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer”. In: *Nucleic acids research* 43.3 (2015), pp. 1345–1356.

[63] Kristoffer Vitting-Seerup and Albin Sandelin. “The landscape of isoform switches in human cancers”. In: *Molecular Cancer Research* 15.9 (2017), pp. 1206–1220.

[64] Abdullah Kahraman et al. “Pathogenic impact of transcript isoform switching in 1,209 cancer samples covering 27 cancer types using an isoform-specific interaction network”. In: *Scientific reports* 10.1 (2020), p. 14453.

[65] Tülay Karakulak et al. “Probing isoform switching events in various cancer types: lessons from pan-cancer studies”. In: *Frontiers in Molecular Biosciences* 8 (2021), p. 726902.

[66] Evangelos D Karousis et al. “Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells”. In: *Genome biology* 22.1 (2021), pp. 1–23.

[67] Matthew Mort et al. “A meta-analysis of nonsense mutations causing human genetic disease”. In: *Human mutation* 29.8 (2008), pp. 1037–1047.

[68] Joachim Weischenfeldt et al. “Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns”. In: *Genome biology* 13.5 (2012), pp. 1–19.

[69] Roberta Bongiorno, Mario Paolo Colombo, and Daniele Lecis. “Deciphering the nonsense-mediated mRNA decay pathway to identify cancer cell vulnerabilities for effective cancer therapy”. In: *Journal of Experimental & Clinical Cancer Research* 40 (2021), pp. 1–11.

[70] F Lejeune. “Triple effect of nonsense-mediated mrna decay inhibition as a therapeutic approach for cancer”. In: *Single Cell Biol* 5.2 (2016), p. 1.

[71] Maximilian W Popp and Lynne E Maquat. “Attenuation of nonsense-mediated mRNA decay facilitates the response to chemotherapeutics”. In: *Nature communications* 6.1 (2015), p. 6632.

[72] Jayanthi P Gudikote et al. “Inhibition of nonsense-mediated decay rescues functional p53β/γ isoforms in MDM2-amplified cancers”. In: *bioRxiv* (2020), pp. 2020–04.

[73] Alexandra E Goetz and Miles Wilkinson. “Stress and the nonsense-mediated RNA decay pathway”. In: *Cellular and Molecular Life Sciences* 74 (2017), pp. 3509–3531.

[74] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic acids research* 41.D1 (2012), pp. D991–D995.

[75] Helen Parkinson et al. “ArrayExpress—a public database of microarray experiments and gene expression profiles”. In: *Nucleic acids research* 35.suppl_1 (2007), pp. D747–D750.

[76] David L Wheeler et al. “Database resources of the national center for biotechnology information”. In: *Nucleic acids research* 33.suppl_1 (2005), pp. D39–D45.

[77] Rasko Leinonen et al. “The European nucleotide archive”. In: *Nucleic acids research* 39.suppl_1 (2010), pp. D28–D31.

[78] John N Weinstein et al. “The cancer genome atlas pan-cancer analysis project”. In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[79] GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.

[80] John Vivian et al. “Toil enables reproducible, open source, big biomedical data analyses”. In: *Nature biotechnology* 35.4 (2017), pp. 314–316.

[81] Simon Anders, Alejandro Reyes, and Wolfgang Huber. “Detecting differential usage of exons from RNA-seq data”. In: *Nature Precedings* (2012), pp. 1–1.

[82] Peter Langfelder and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis”. In: *BMC bioinformatics* 9.1 (2008), pp. 1–13.

[83] Uku Raudvere et al. “g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic acids research* 47.W1 (2019), W191–W198.

[84] Maxim V Kuleshov et al. “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update”. In: *Nucleic acids research* 44.W1 (2016), W90–W97.

[85] Liis Kolberg et al. “gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler”. In: *F1000Research* 9 (2020).

[86] Zuguang Gu, Roland Eils, and Matthias Schlesner. “Complex heatmaps reveal patterns and correlations in multidimensional genomic data”. In: *Bioinformatics* 32.18 (2016), pp. 2847–2849.

[87] Geoffrey M Cooper and Kenneth W Adams. *The cell: a molecular approach*. Oxford University Press, 2023.

[88] Nader El-Sayes, Alyssa Vito, and Karen Mossman. “Tumor heterogeneity: a great barrier in the age of cancer immunotherapy”. In: *Cancers* 13.4 (2021), p. 806.

[89] Sowmya Chivukula and Vasavi Malkhed. “The Role of CDK20 Protein in Carcinogenesis”. In: *Current Drug Targets* 24.10 (2023), pp. 790–796.

[90] Prince K Awuah et al. “Cell cycle–related kinase links androgen receptor and β-catenin signaling in hepatocellular carcinoma: Why are men at a loss?” In: *Hepatology* 55.3 (2012), pp. 970–974.

[91] Lo Lai, Ga Yoon Shin, and Hongyu Qiu. “The role of cell cycle regulators in cell survival—dual functions of cyclin-dependent kinase 20 and p21cip1/waf1”. In: *International journal of molecular sciences* 21.22 (2020), p. 8504.

[92] Guo-Qing Wu et al. “Cell cycle-related kinase supports ovarian carcinoma cell proliferation via regulation of cyclin D1 and is a predictor of outcome in patients with ovarian carcinoma”. In: *International journal of cancer* 125.11 (2009), pp. 2631–2642.

[93] Q Wang et al. “CDK20 interacts with KEAP1 to activate NRF2 and promotes radiochemoresistance in lung cancer cells”. In: *Oncogene* 36.37 (2017), pp. 5321–5330.

[94] Samuel SM Ng et al. “Cell Cycle–Related Kinase: A Novel Candidate Oncogene in Human Glioblastoma”. In: *Journal of the National Cancer Institute* 99.12 (2007), pp. 936–948.

[95] Jin Wang et al. “lncRNA ZNRD1-AS1 promotes malignant lung cell proliferation, migration, and angiogenesis via the miR-942/TNS1 axis and is positively regulated by the m6A reader YTHDC2”. In: *Molecular Cancer* 21.1 (2022), pp. 1–19.

[96] Yunping Zhao et al. “Expression and prognostic value of ZNRD1 in esophageal squamous cell carcinoma”. In: *Digestive diseases and sciences* 54 (2009), pp. 586–592.

[97] Liu Hong et al. “ZNRD1 gene suppresses cell proliferation through cell cycle arrest in G1 phase”. In: *Cancer biology & therapy* 4.1 (2005), pp. 67–71.

[98] Liu Hong et al. “Mechanisms of growth arrest by zinc ribbon domain-containing 1 in gastric cancer cells”. In: *Carcinogenesis* 28.8 (2007), pp. 1622–1628.

[99] Lanwei Guo et al. “Expression quantitative trait loci in long non-coding RNA ZNRD1-AS1 influence cervical cancer development”. In: *American journal of cancer research* 5.7 (2015), p. 2301.

[100] Dan Li et al. “Strong evidence for LncRNA ZNRD1-AS1, and its functional Cis-eQTL locus contributing more to the susceptibility of lung cancer”. In: *Oncotarget* 7.24 (2016), p. 35813.

[101] Paul Blakeley et al. “Investigating protein isoforms via proteomics: a feasibility study”. In: *Proteomics* 10.6 (2010), pp. 1127–1140.

[102] Liufang Zhao et al. “FAM136A immunoreactivity is associated with nodal involvement and survival in lung adenocarcinoma in a Chinese case series.” In: *Bioengineered* 11.1 (2020), pp. 261–271.

[103] Shahan Mamoor. “TIMM17B is a differentially expressed gene in human metastatic breast cancer, in the brain and in the lymph nodes.” In: ().

[104] Mingting Duan et al. “High Expression of TIMM17B Is a Potential Diagnostic and Prognostic Marker of Breast Cancer”. In: *Cellular and Molecular Biology* 69.3 (2023), pp. 169–176.

[105] Jeff S Isenberg, Christine Yu, and David D Roberts. “Differential effects of ABT-510 and a CD36-binding peptide derived from the type 1 repeats of thrombospondin-1 on fatty acid uptake, nitric oxide signaling, and caspase activation in vascular cells”. In: *Biochemical pharmacology* 75.4 (2008), pp. 875–882.

[106] Xingda Zhang et al. “Revealing the Potential Markers of N (4)-Acetylcytidine through acRIP-seq in Triple-Negative Breast Cancer”. In: *Genes* 13.12 (2022), p. 2400.

APPENDIX A: SUPPLEMENTARY FIGURES

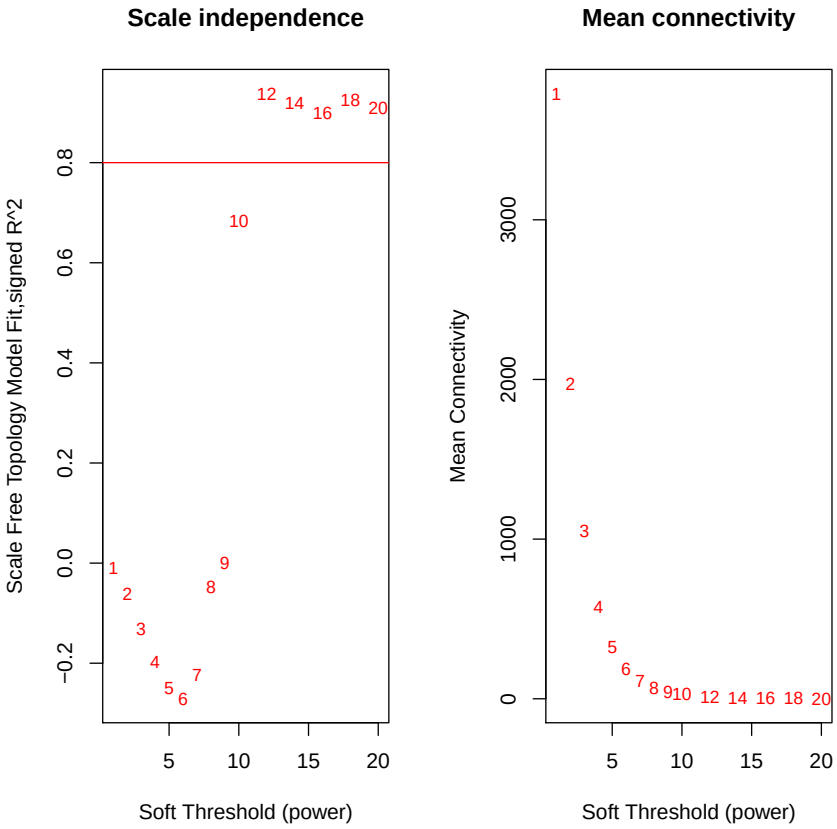

Figure A.1: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for BLCA

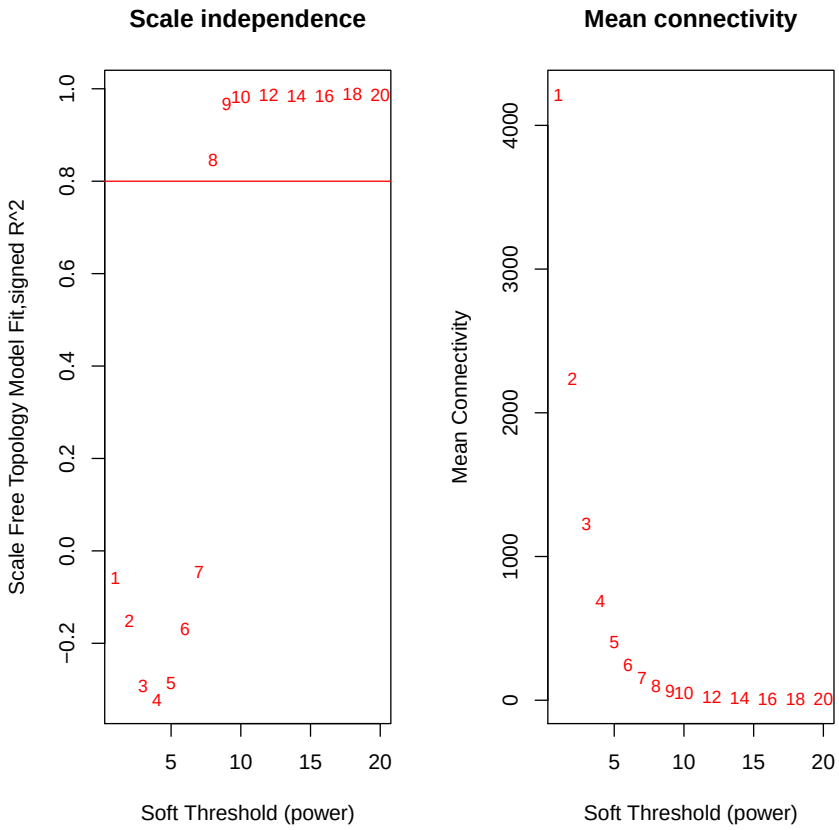

Figure A.2: Choosing soft threshold power for LUAD

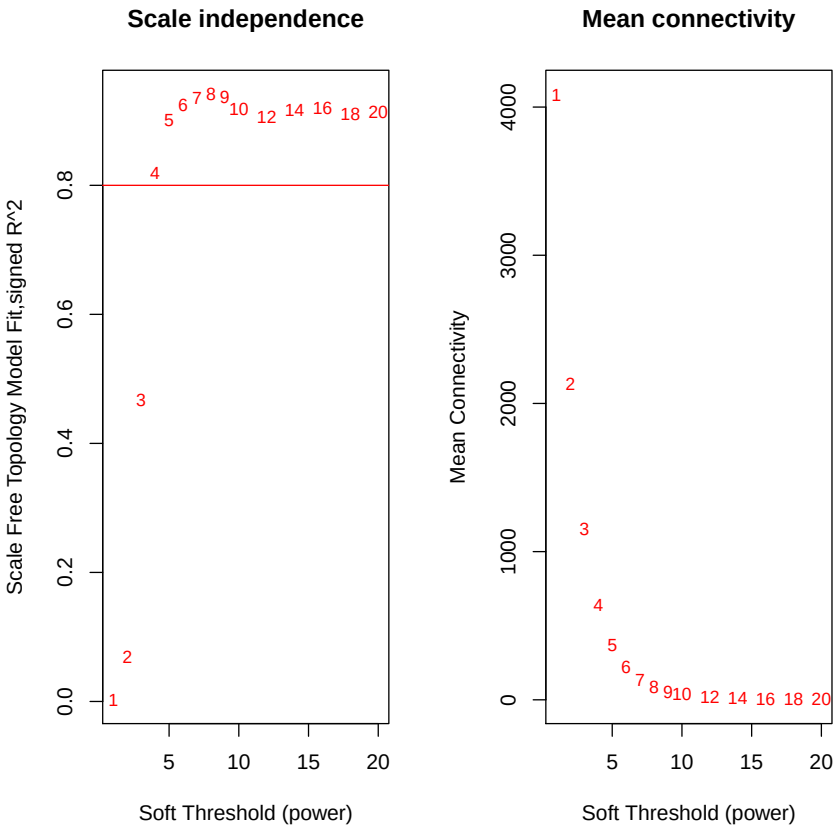

Figure A.3: Choosing soft threshold power for LUSC

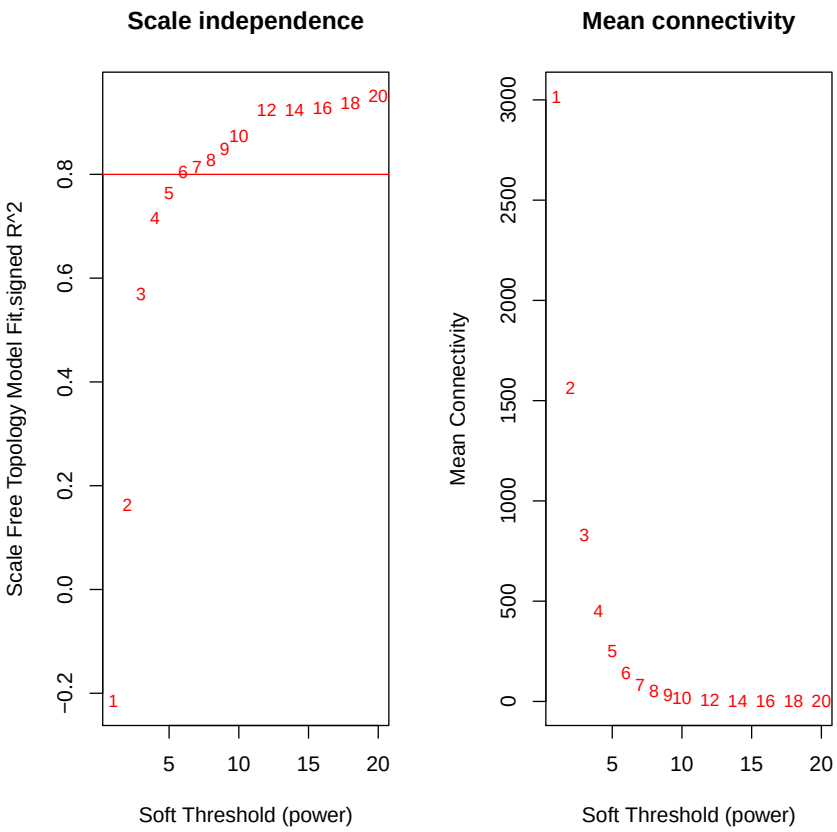

Figure A.4: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for LIHC

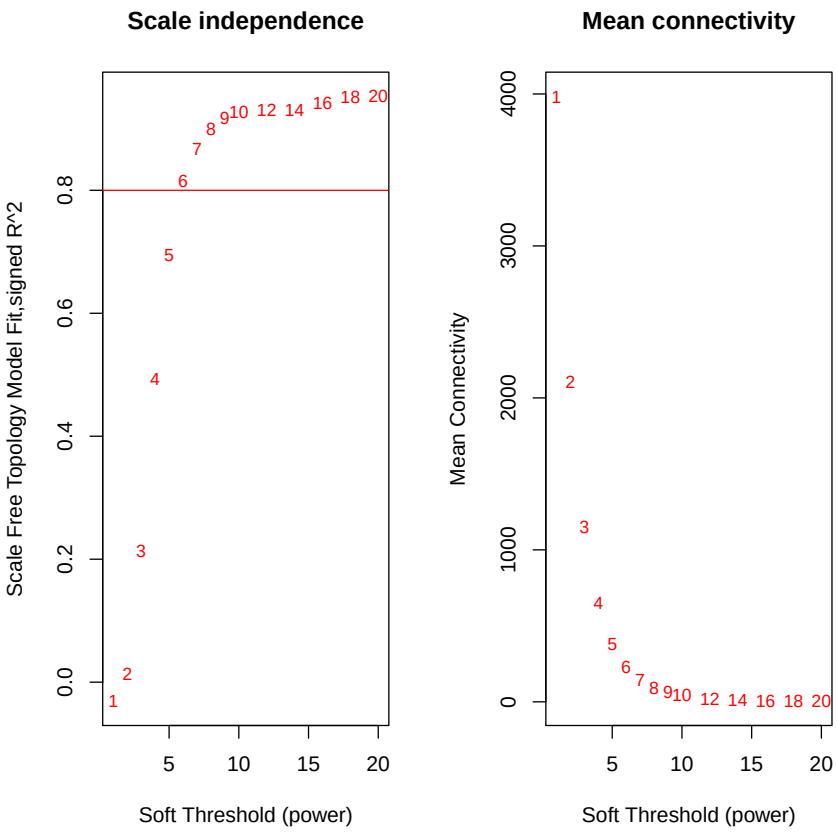

Figure A.5: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for KIRC

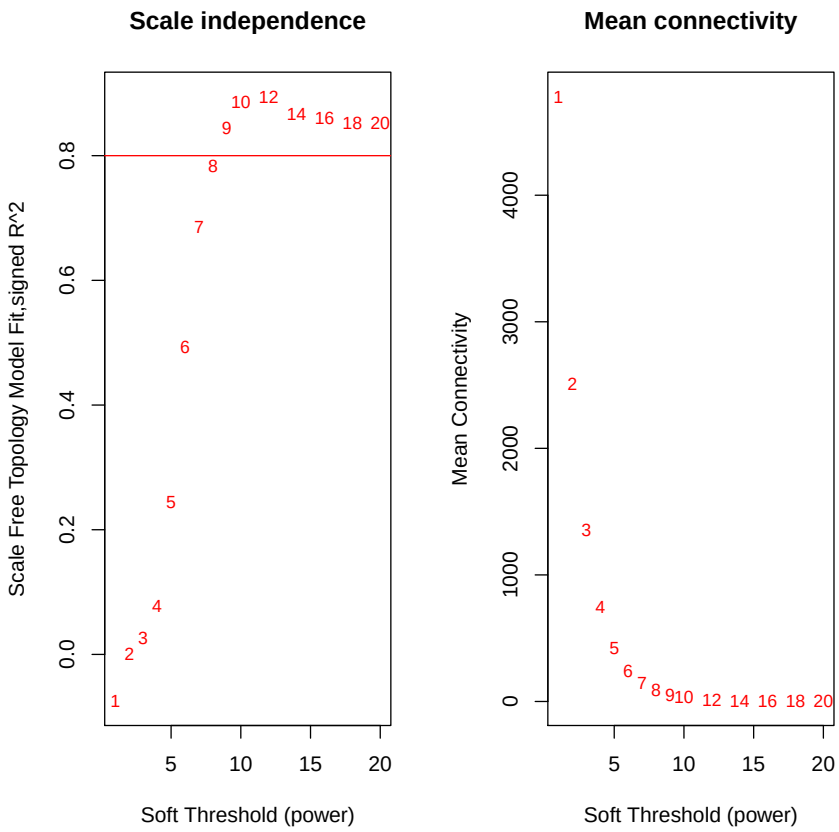

Figure A.6: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for STAD

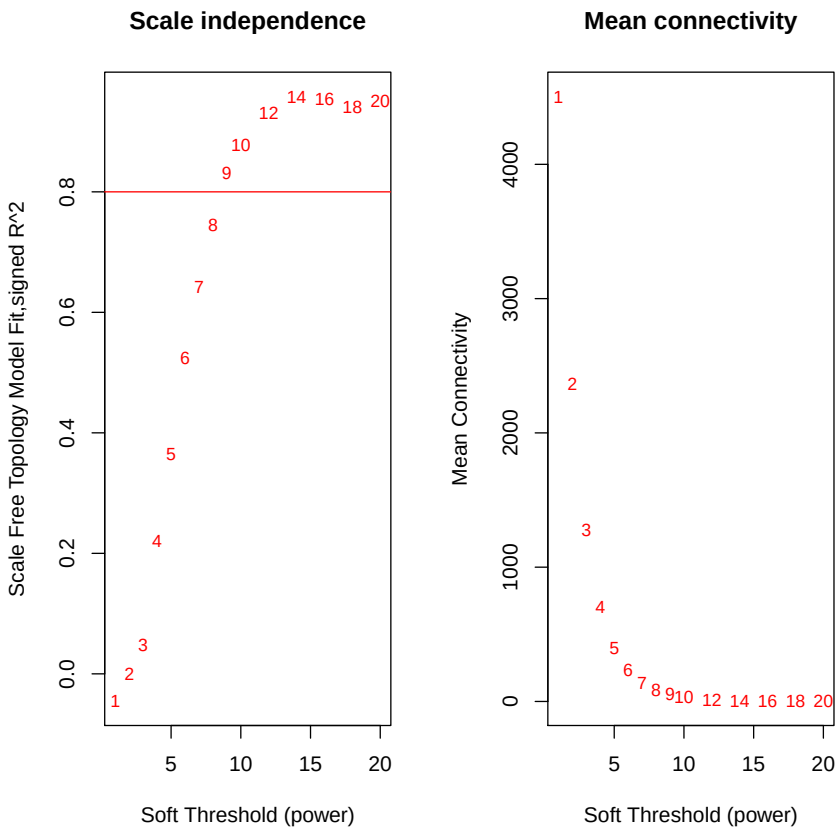

Figure A.7: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for ESCA

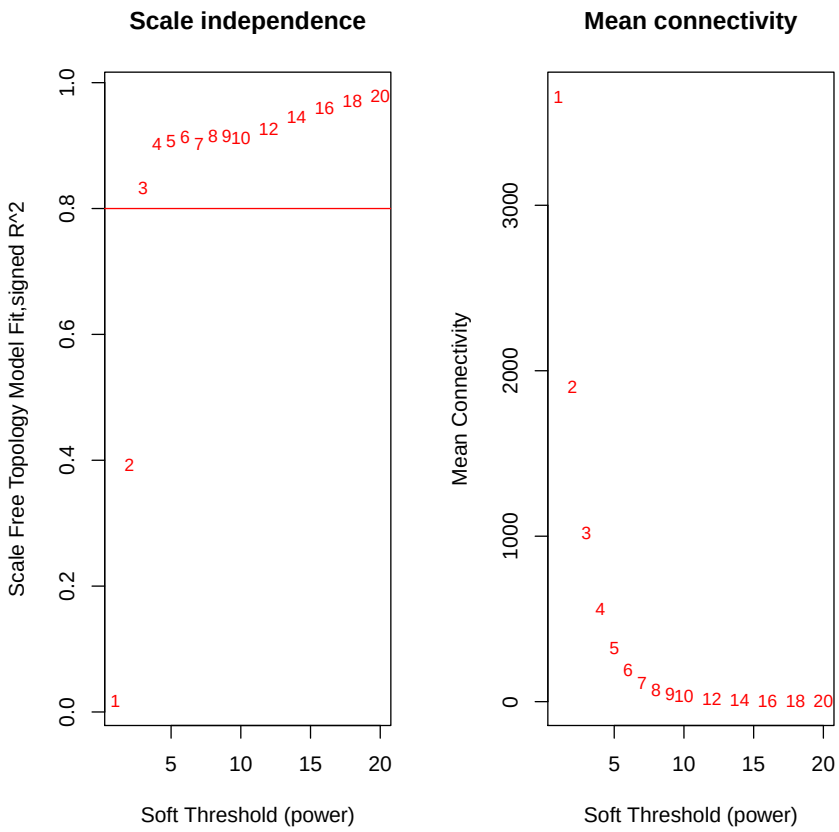

Figure A.8: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for COAD

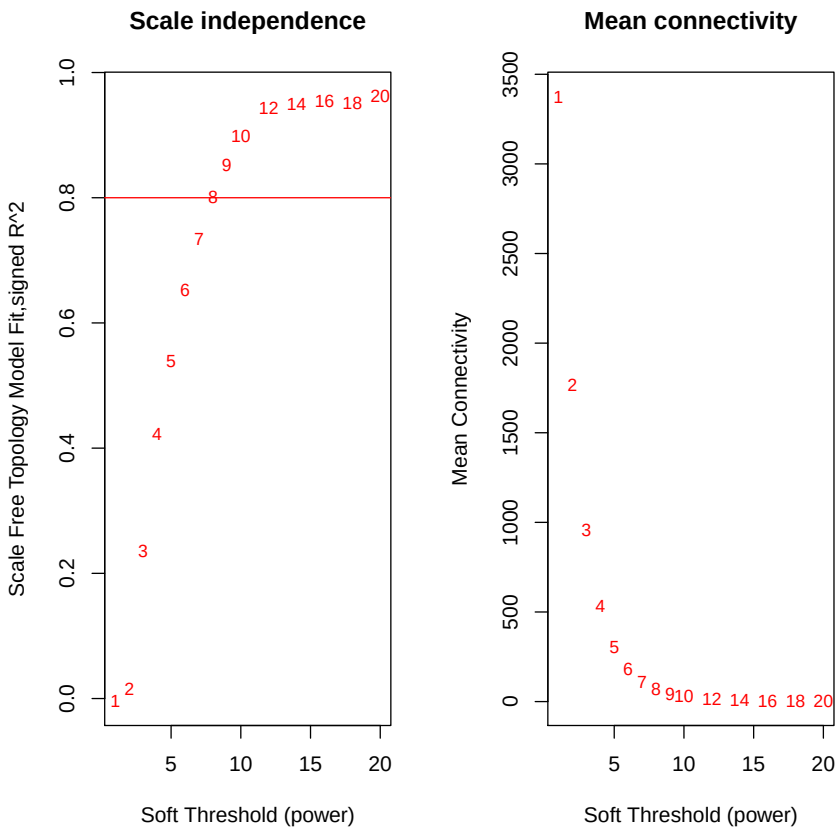

Figure A.9: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for HNSC

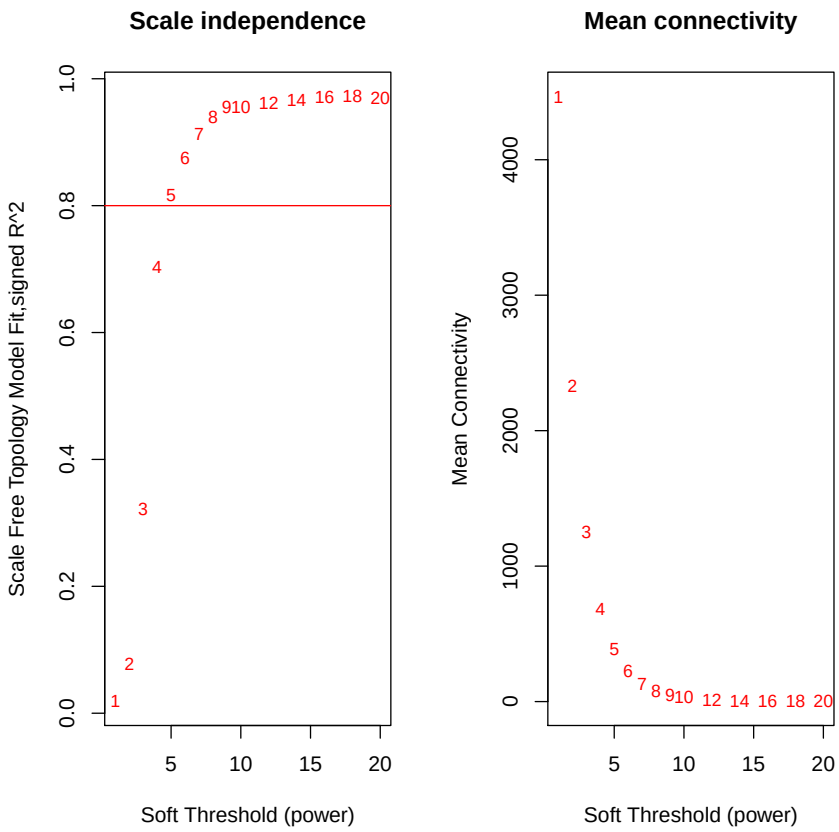

Figure A.10: Impact of power values on the scale independence and mean connectivity of genes in co-expression modules for BRCA

Appendix A: SUPPLEMENTARY FIGURES

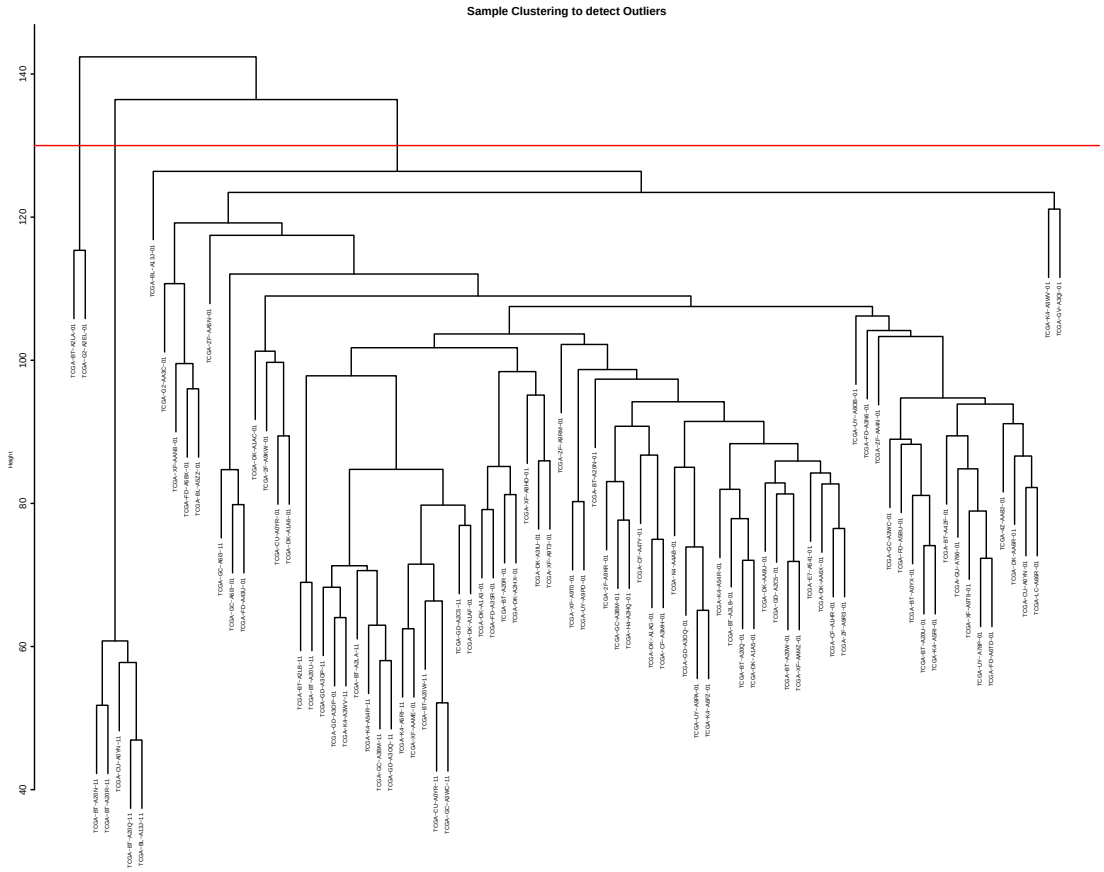

Figure A.11: Sample Clustering Dendrogram for BLCA

Appendix A: SUPPLEMENTARY FIGURES

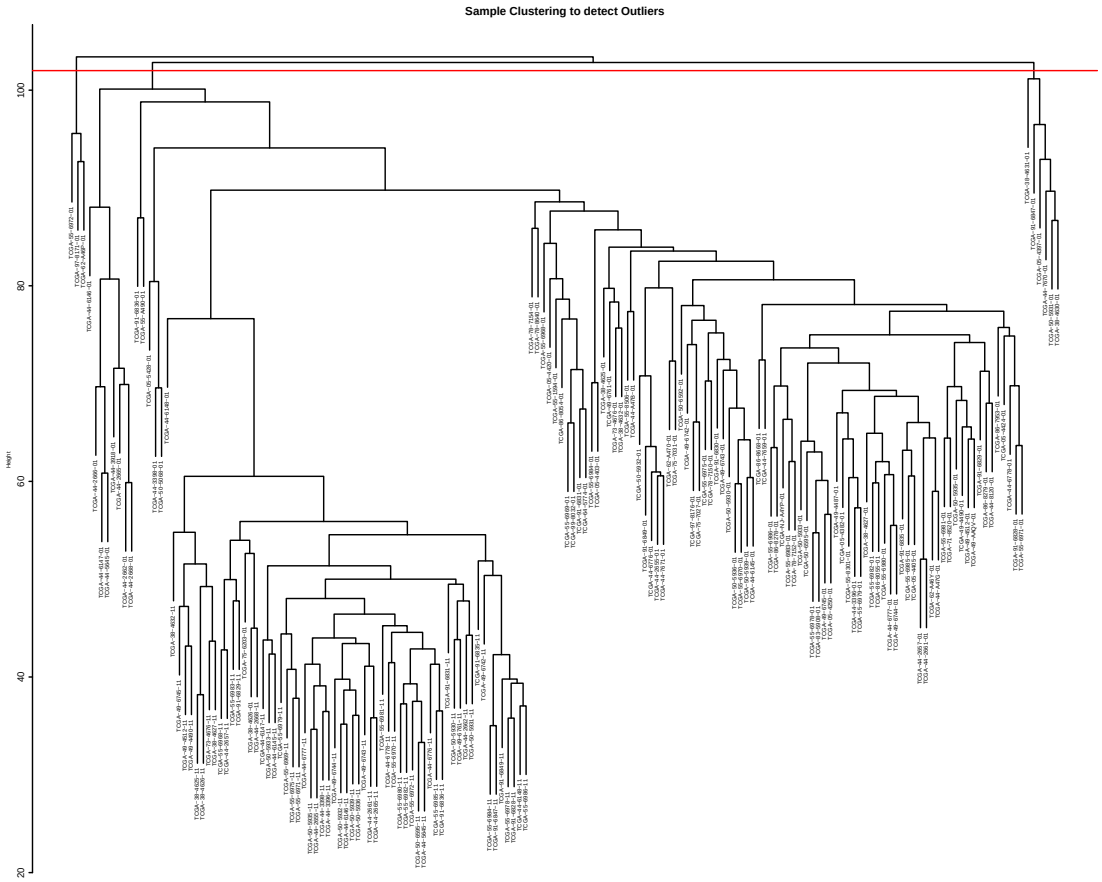

Figure A.12: Sample Clustering Dendrogram for LUAD

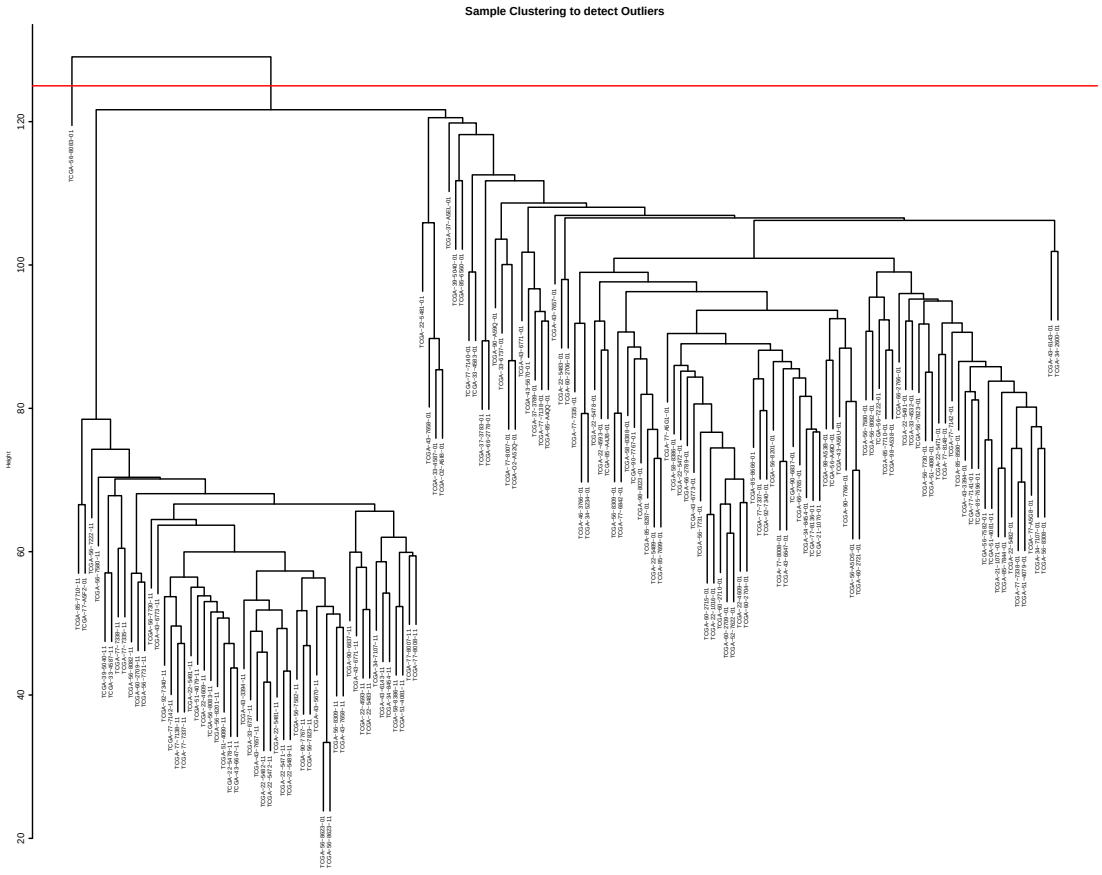

Figure A.13: Sample Clustering Dendrogram for LUSC

Appendix A: SUPPLEMENTARY FIGURES

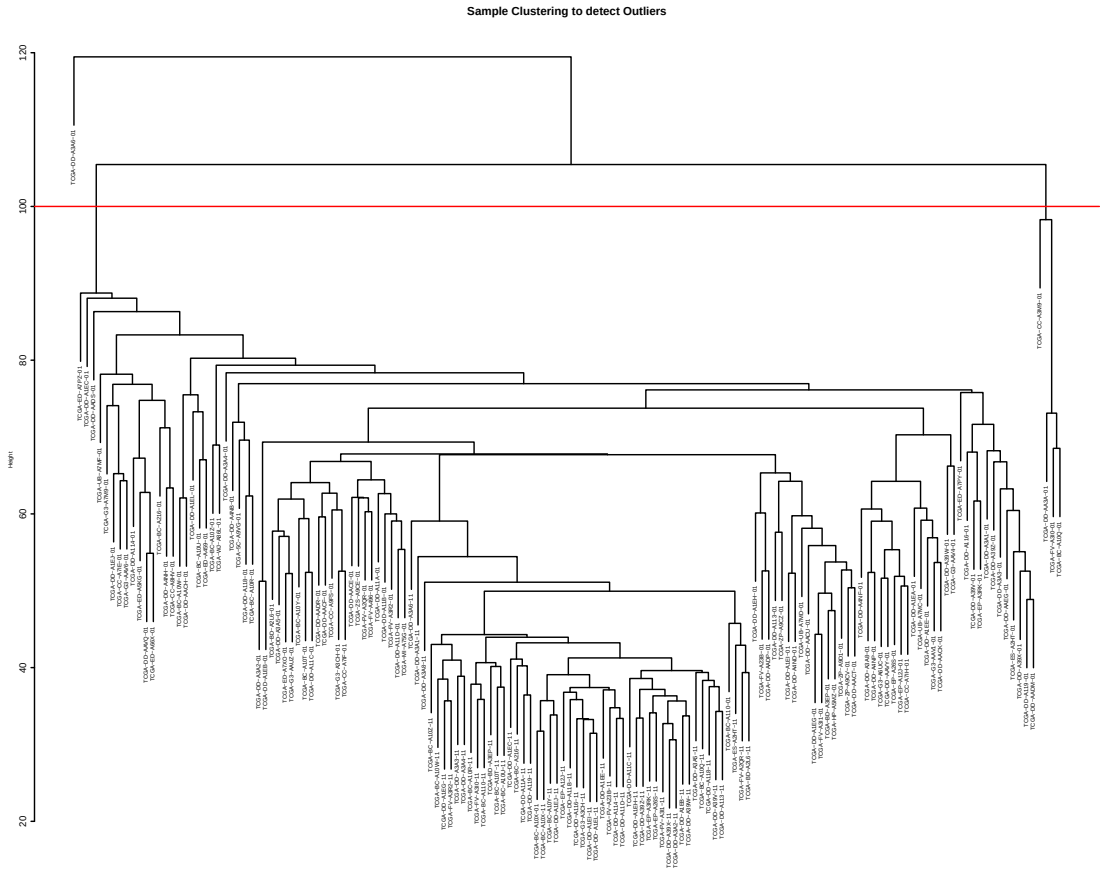

Figure A.14: Sample Clustering Dendrogram for LIHC

Appendix A: SUPPLEMENTARY FIGURES

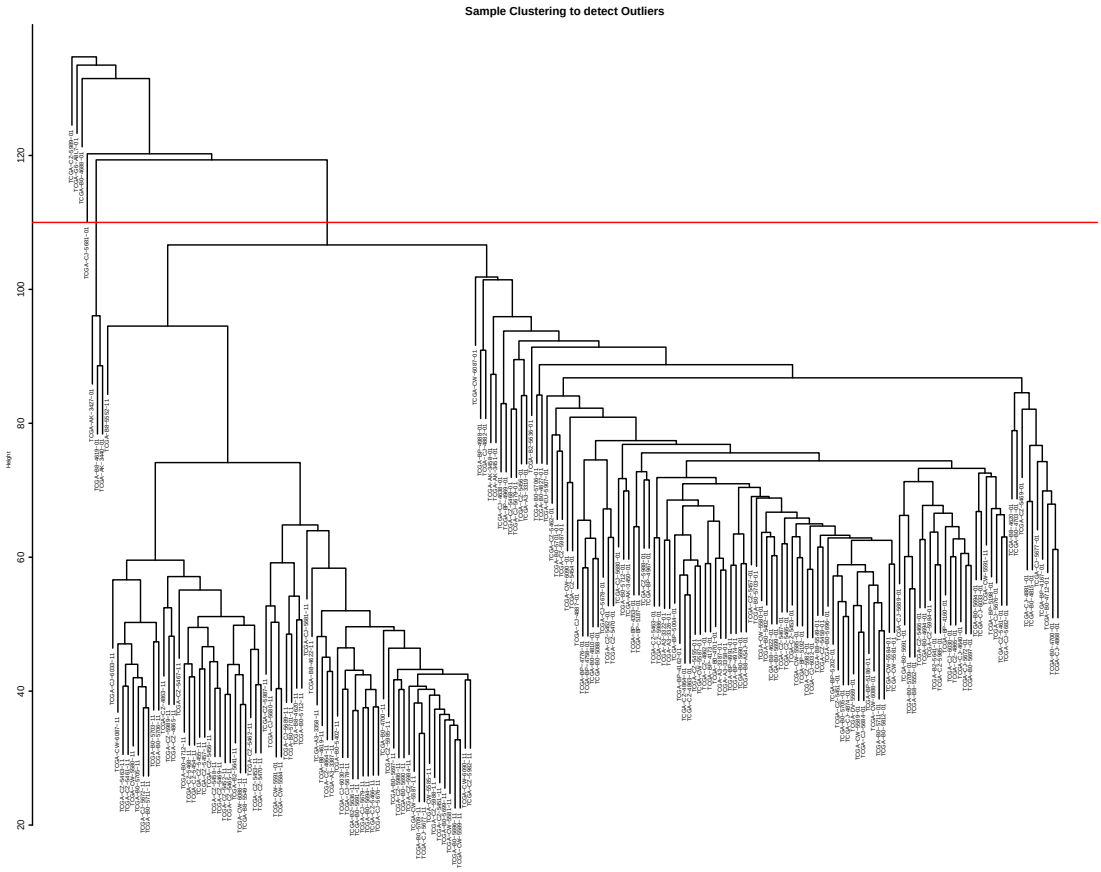

Figure A.15: Sample Clustering Dendrogram for KIRC

Appendix A: SUPPLEMENTARY FIGURES

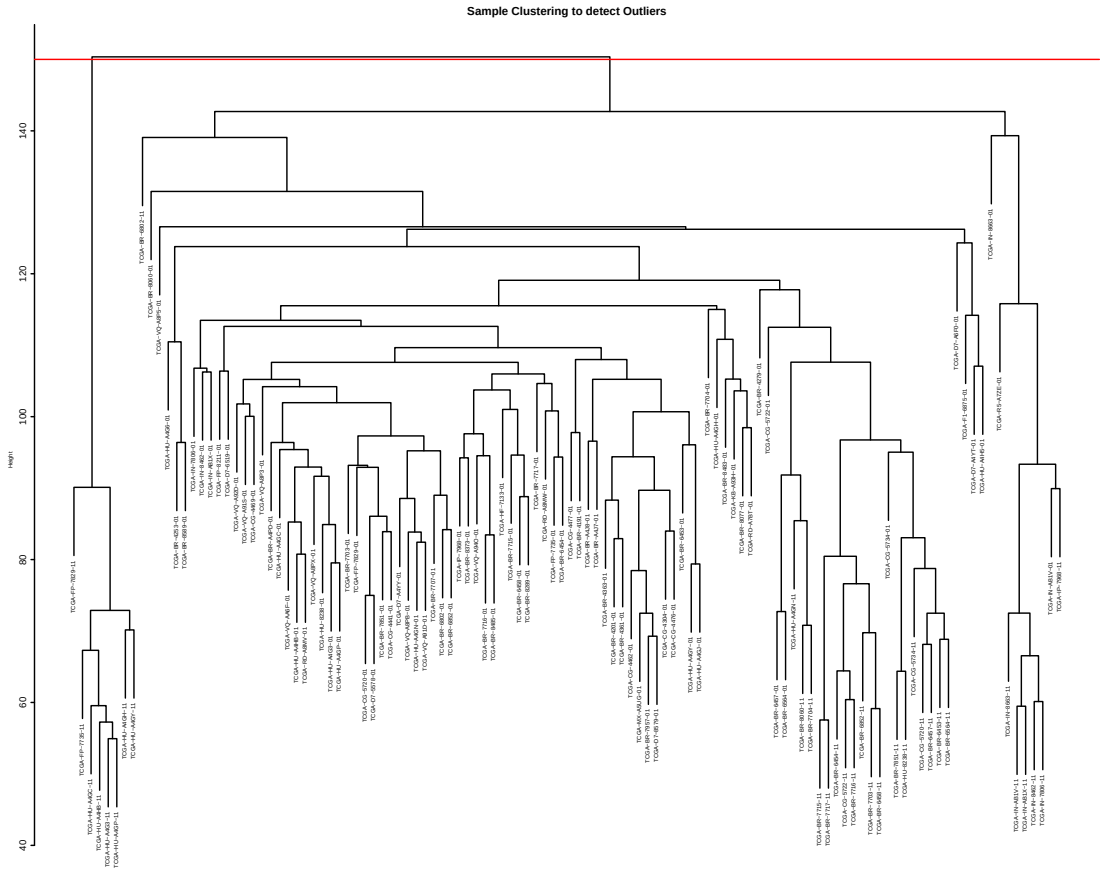

Figure A.16: Sample Clustering Dendrogram for STAD

Appendix A: SUPPLEMENTARY FIGURES

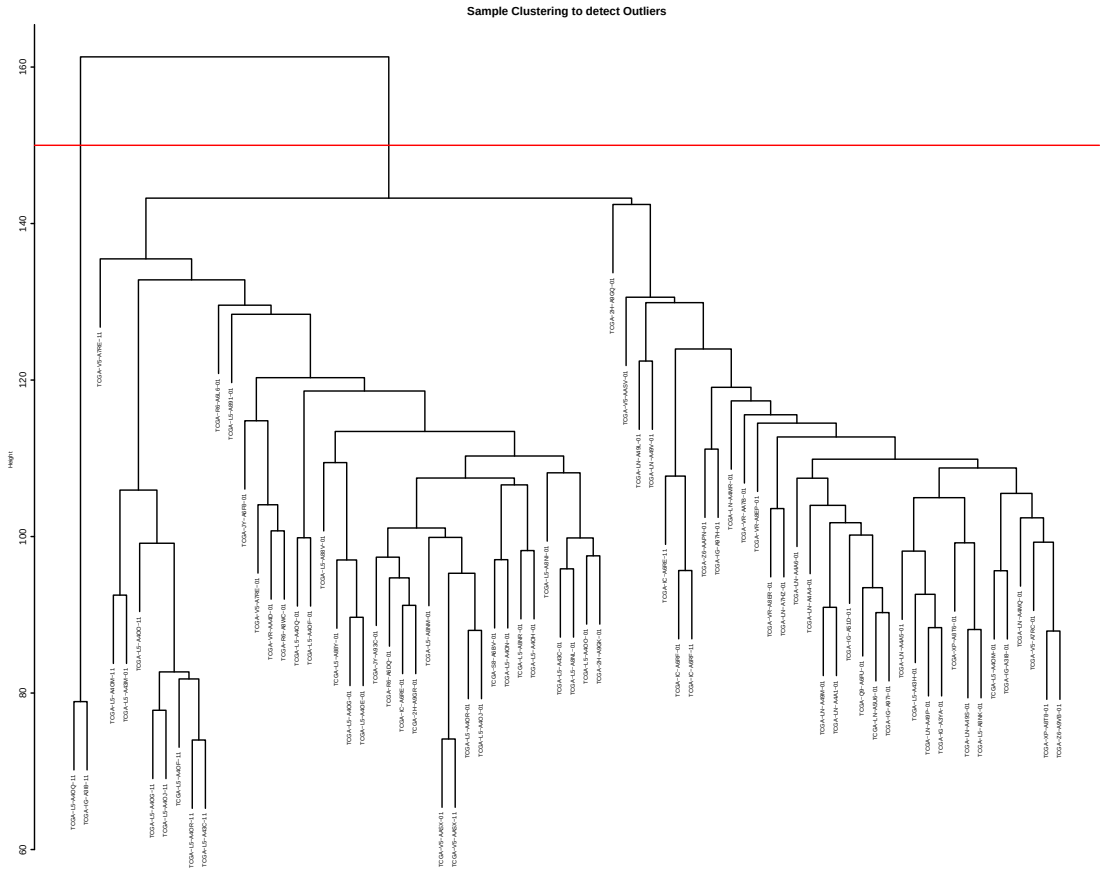

Figure A.17: Sample Clustering Dendrogram for ESCA

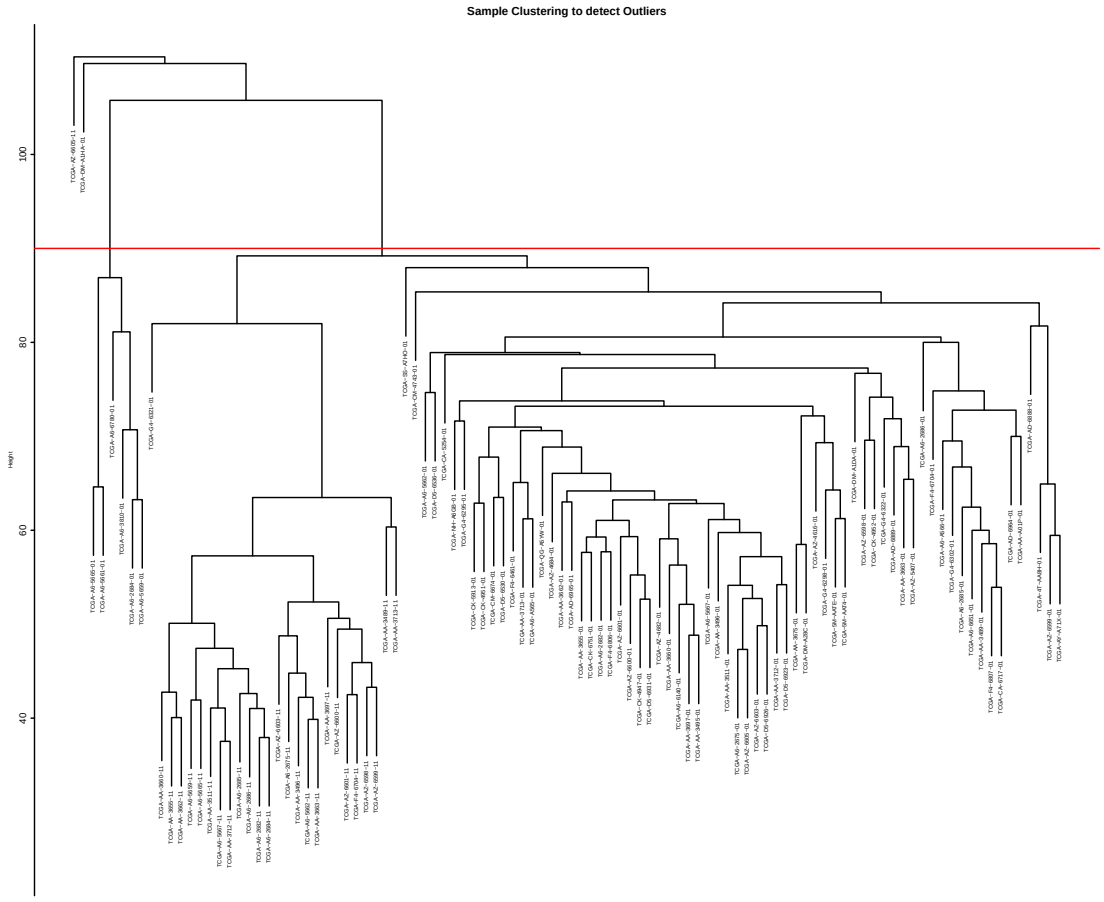

Figure A.18: Sample Clustering Dendrogram for COAD

Appendix A: SUPPLEMENTARY FIGURES

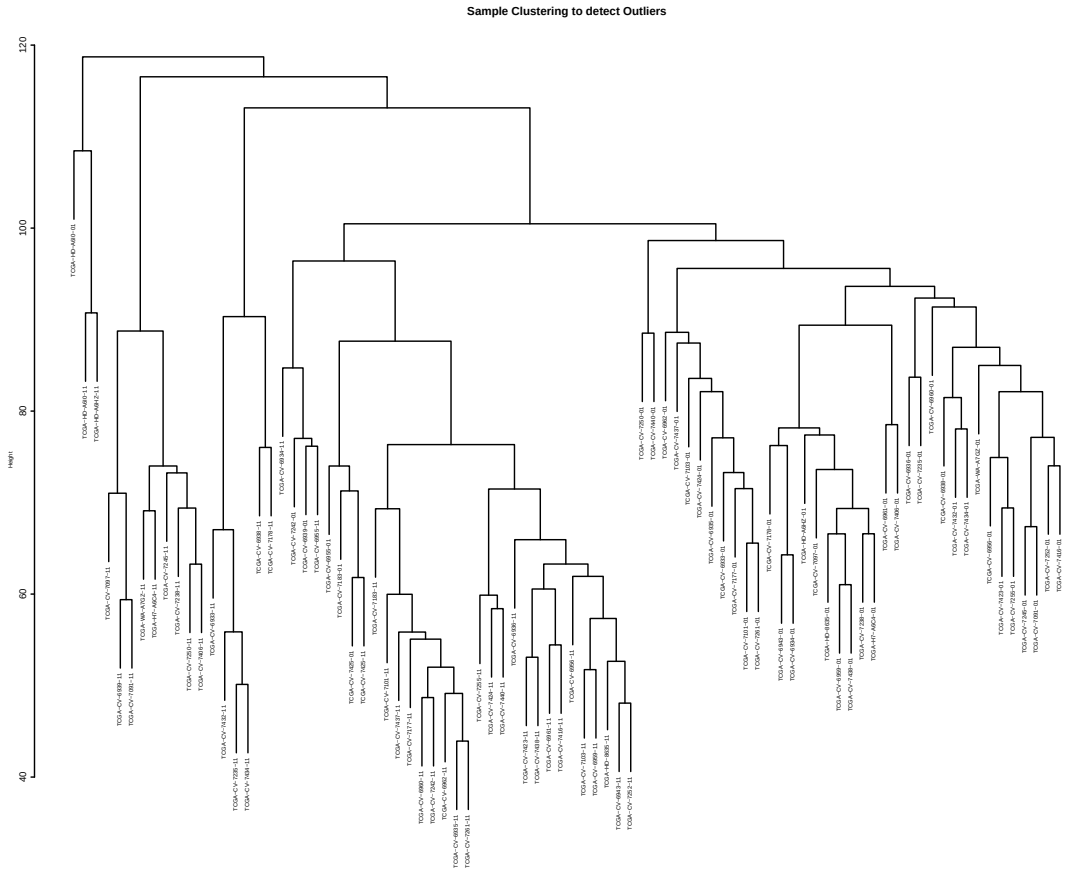

Figure A.19: Sample Clustering Dendrogram for HNSC

Appendix A: SUPPLEMENTARY FIGURES

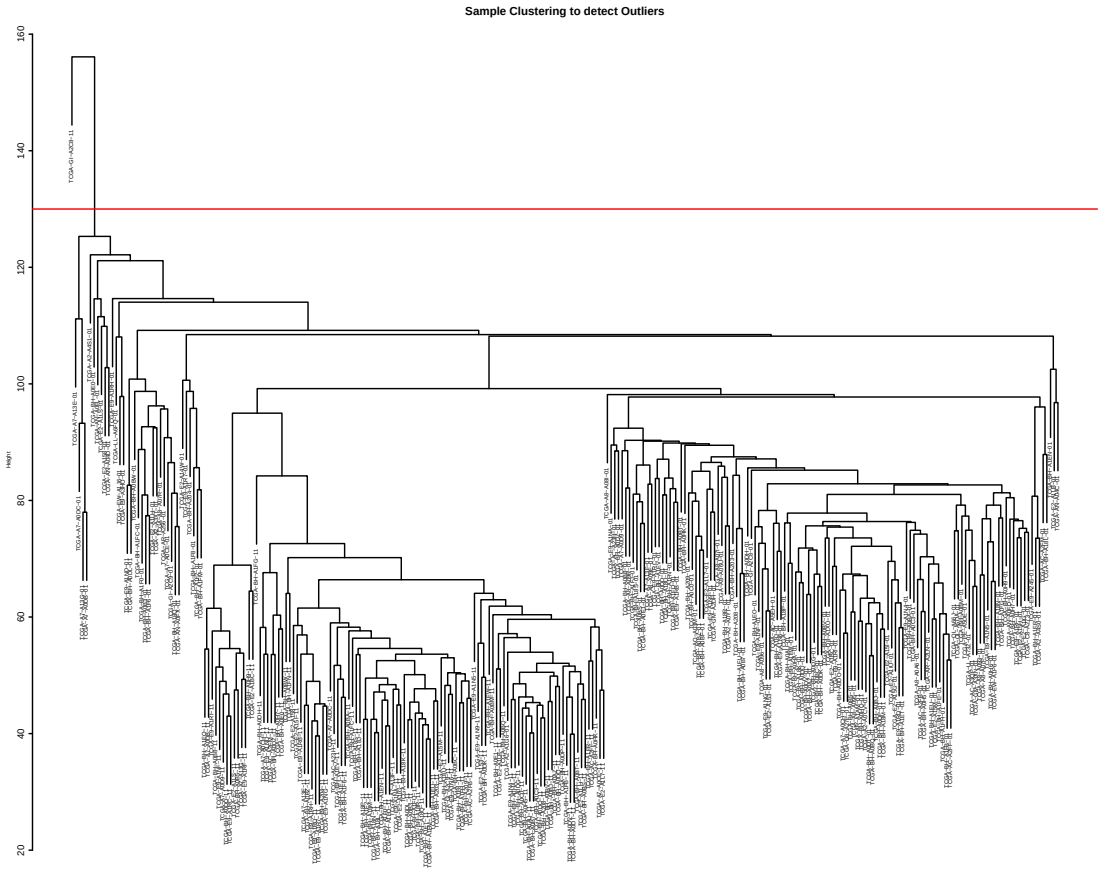

Figure A.20: Sample Clustering Dendrogram for BRCA

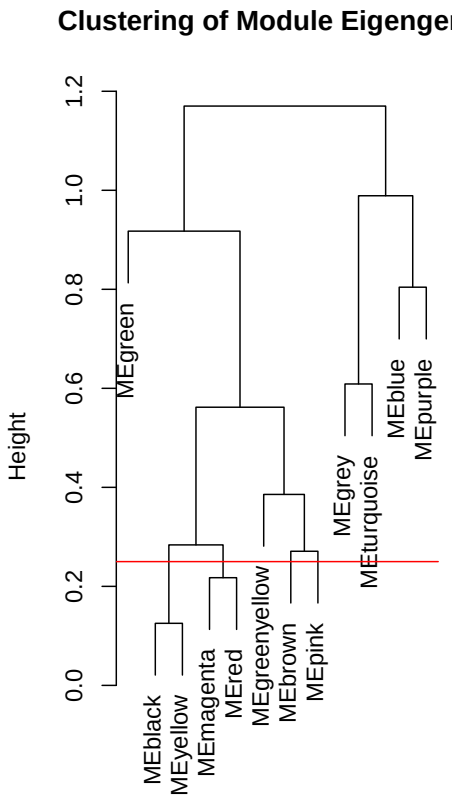

Figure A.21: Gene Dendrogram for BLCA

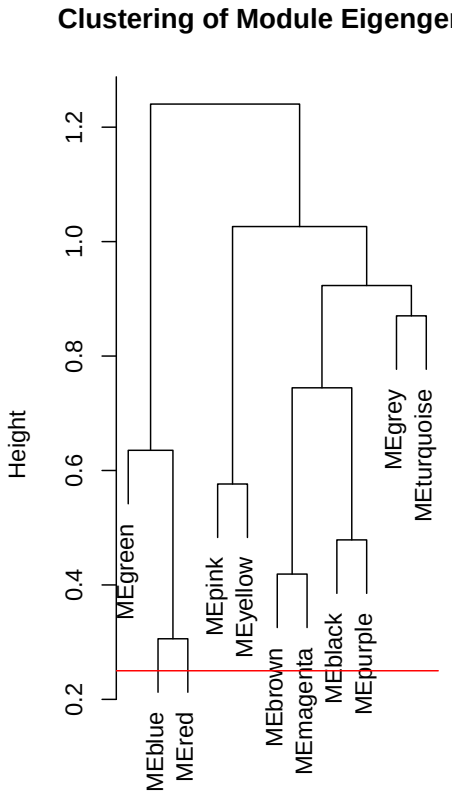

Figure A.22: Gene Dendrogram for LUAD

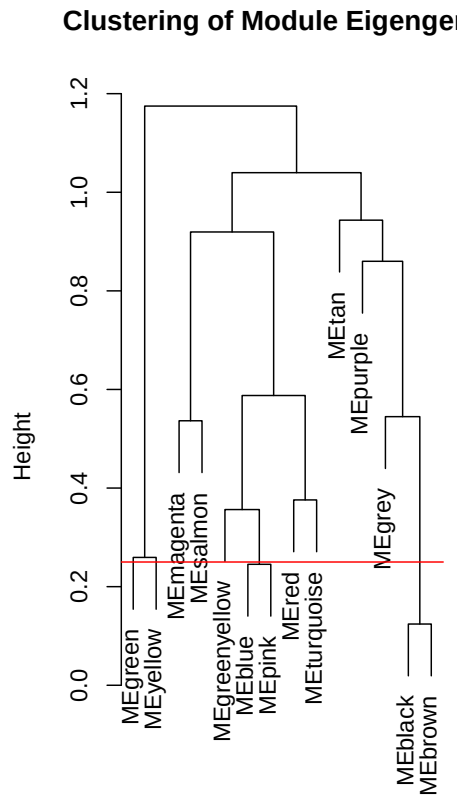

Figure A.23: Gene Dendrogram for LUSC

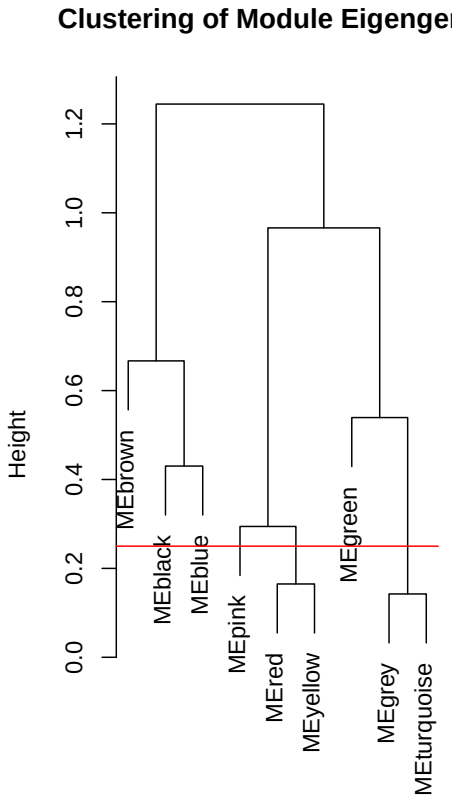

Figure A.24: Gene Dendrogram for LIHC

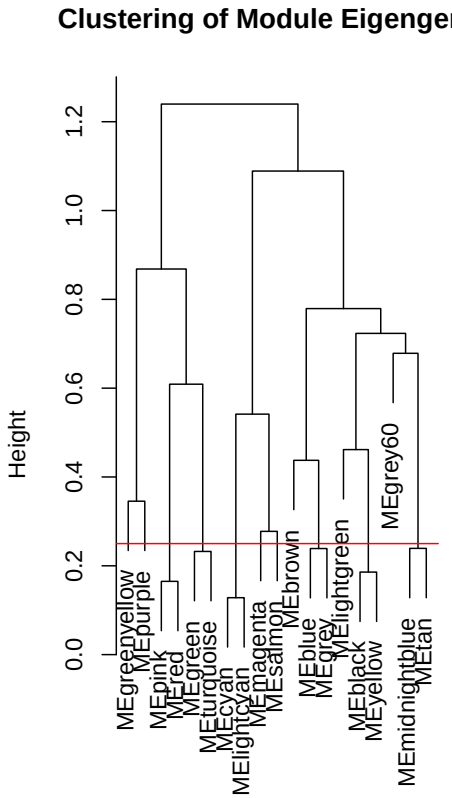

Figure A.25: Gene Dendrogram for KIRC

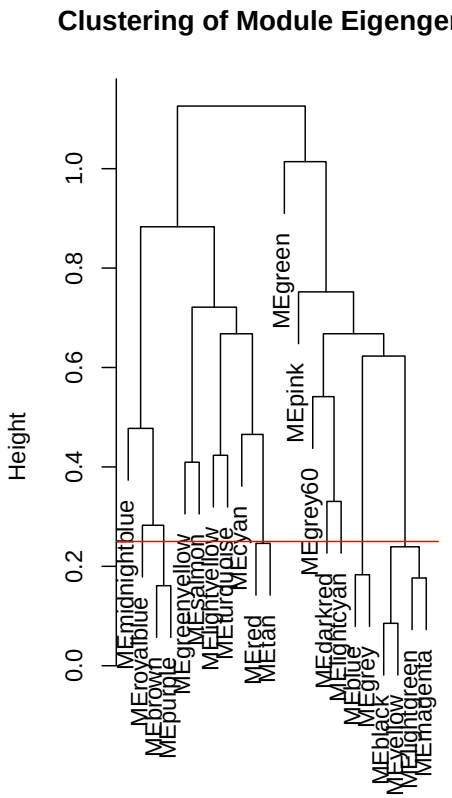

Figure A.26: Gene Dendrogram for STAD

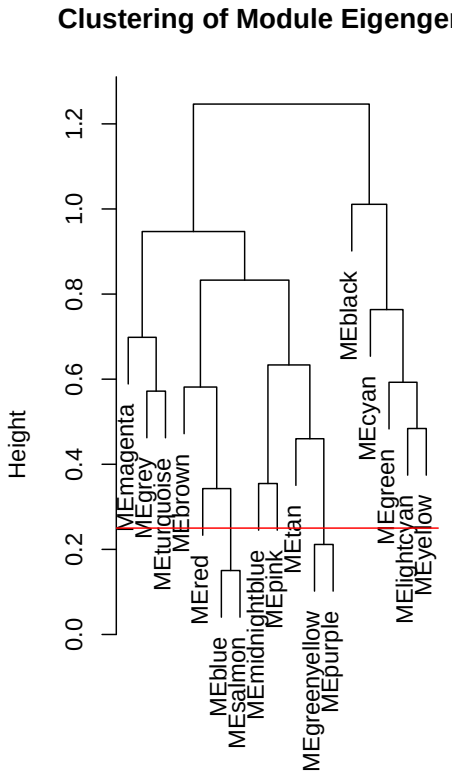

Figure A.27: Gene Dendrogram for ESCA

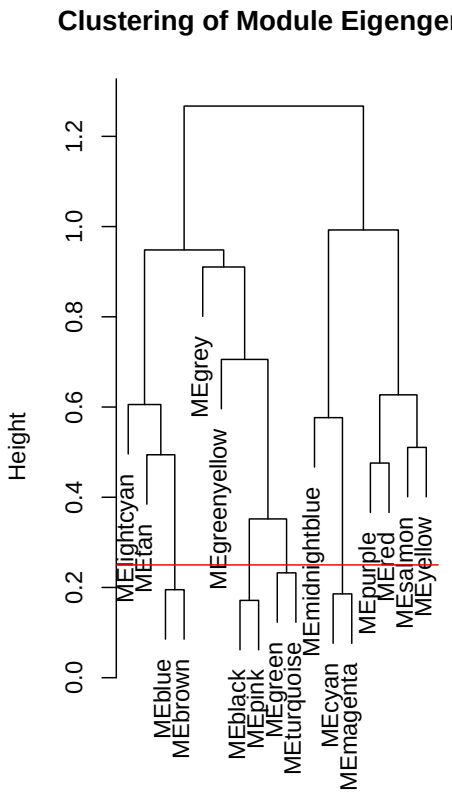

Figure A.28: Gene Dendrogram for COAD

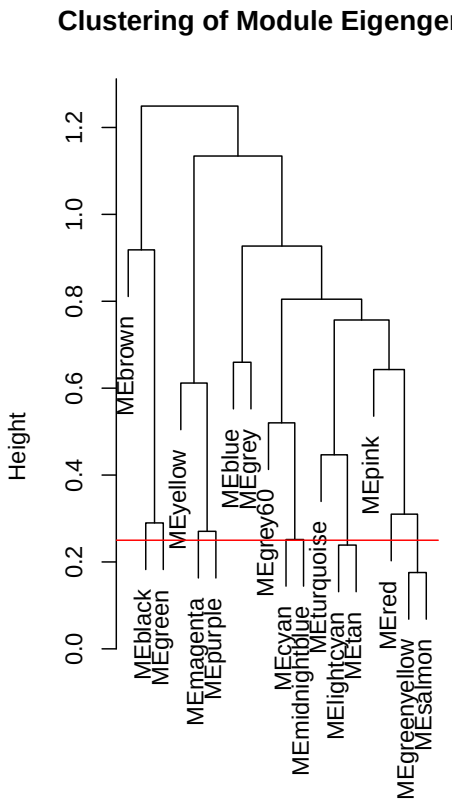

Figure A.29: Gene Dendrogram for HNSC

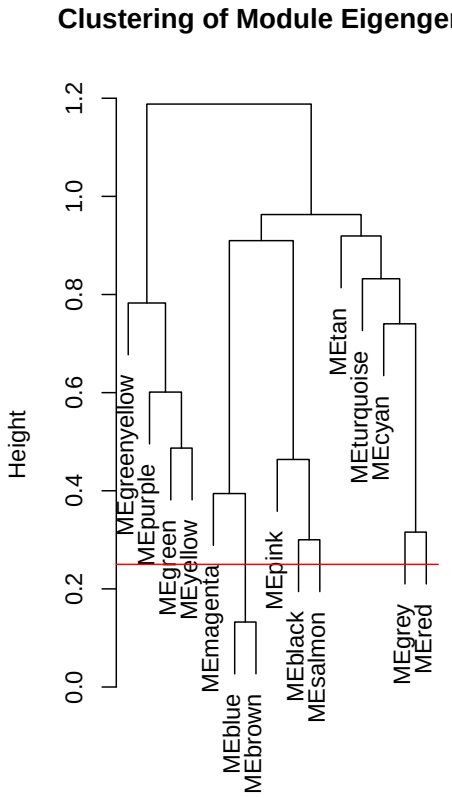

Figure A.30: Gene Dendrogram for BRCA

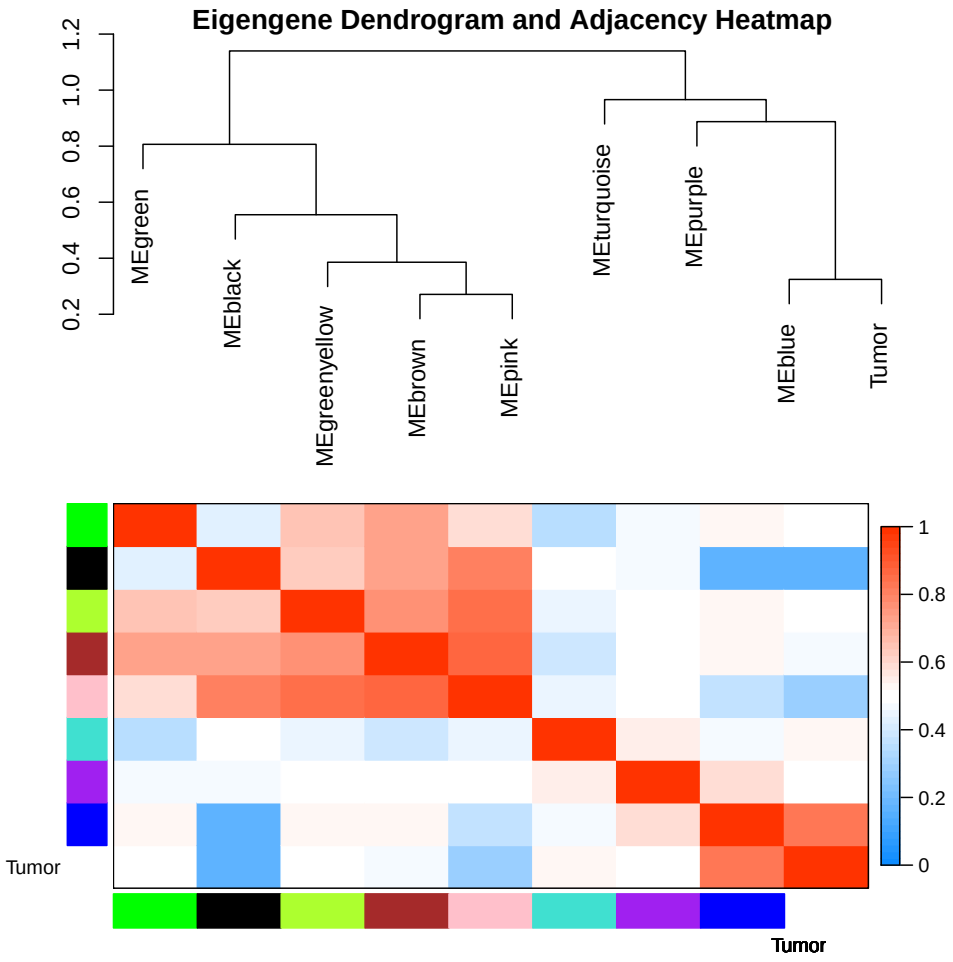

Figure A.31: Eigengene Network for BLCA

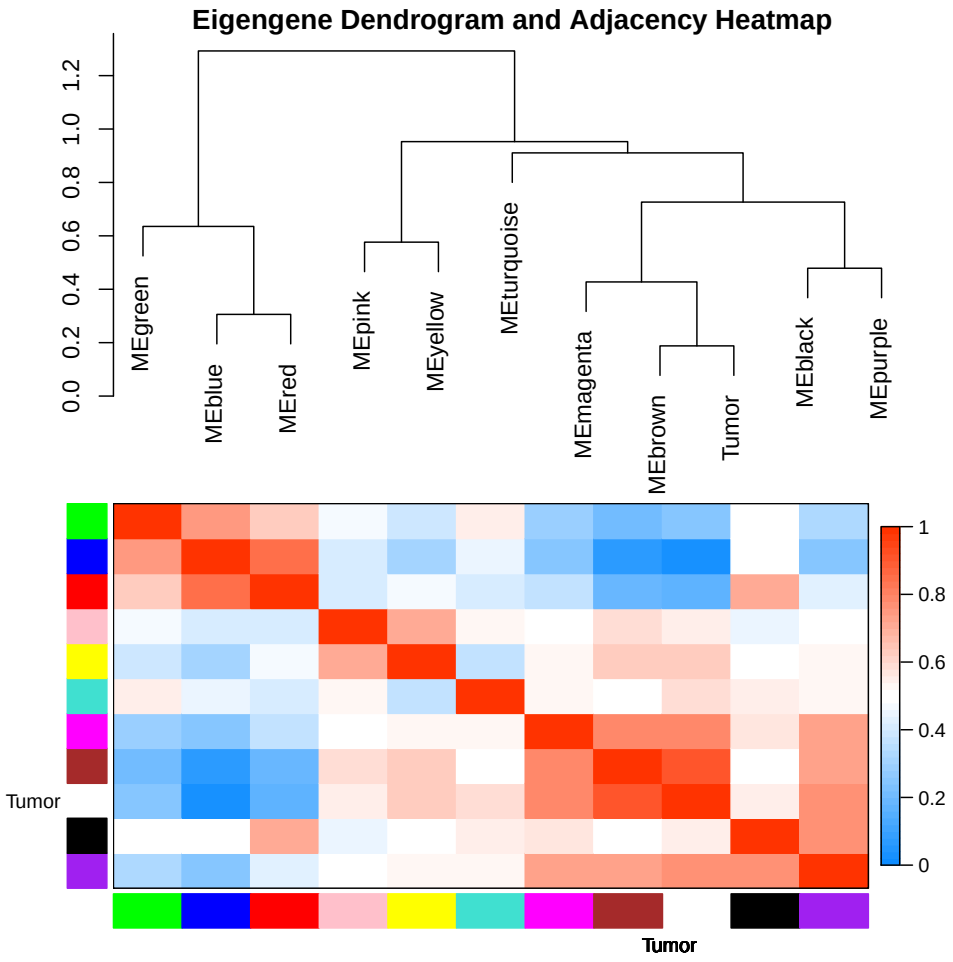

Figure A.32: Eigengene Network for LUAD

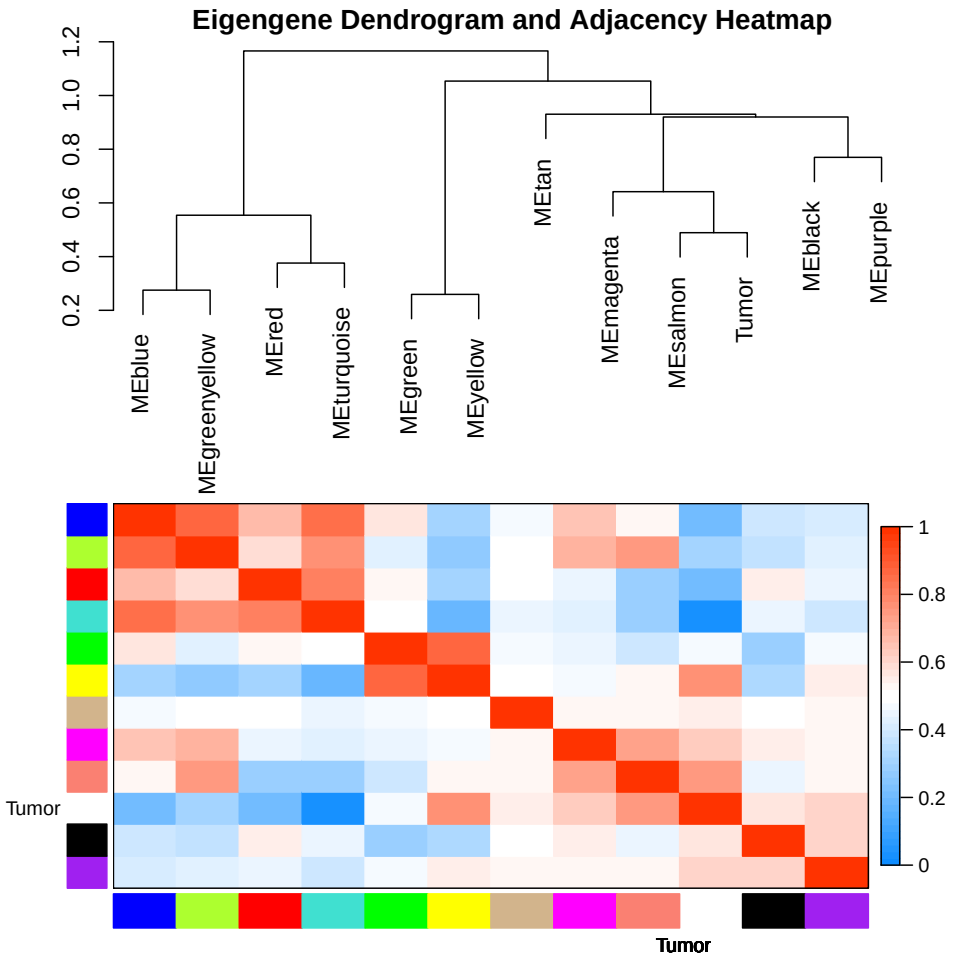

Figure A.33: Eigengene Network for LUSC

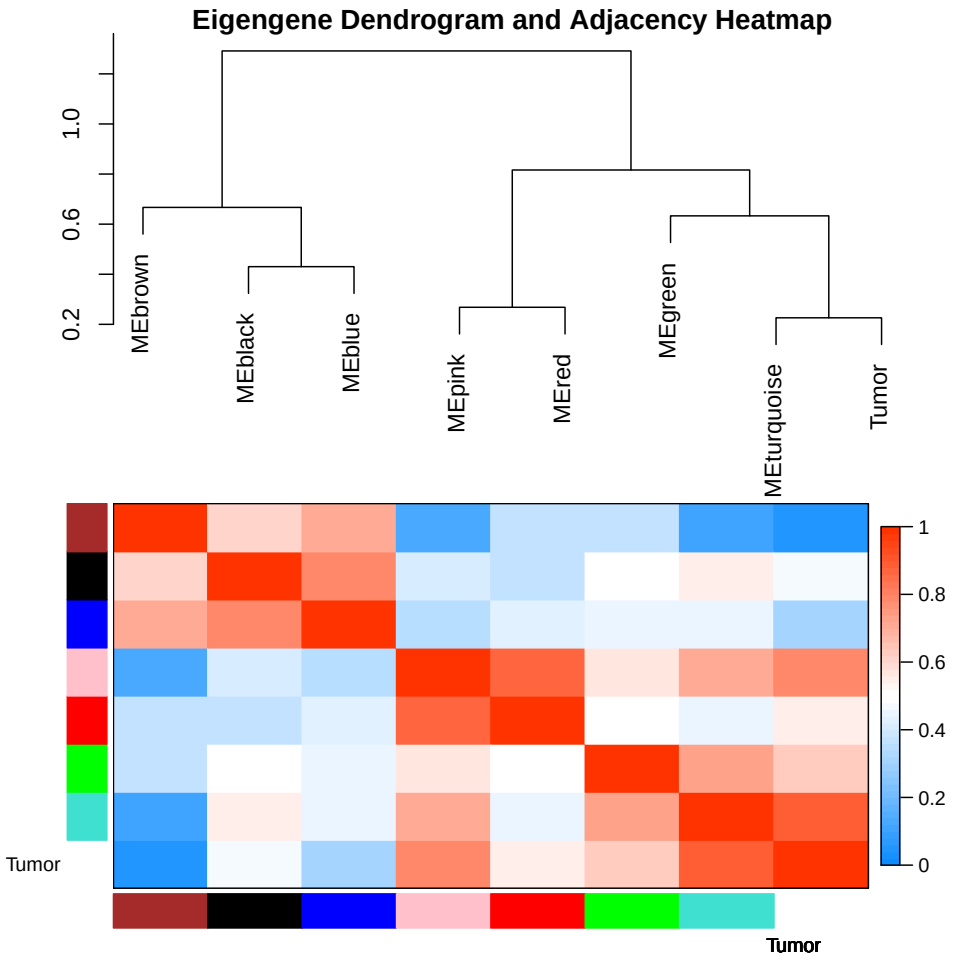

Figure A.34: Eigengene Network for LIHC

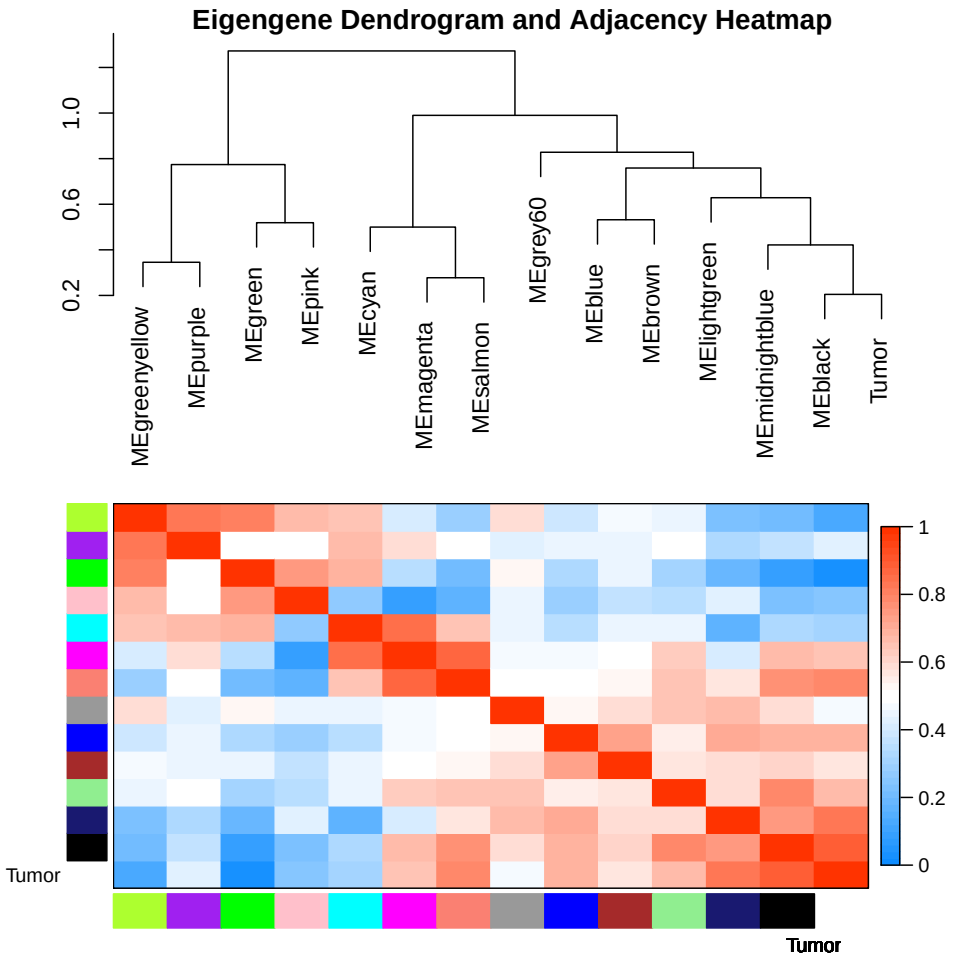

Figure A.35: Eigengene Network for KIRC

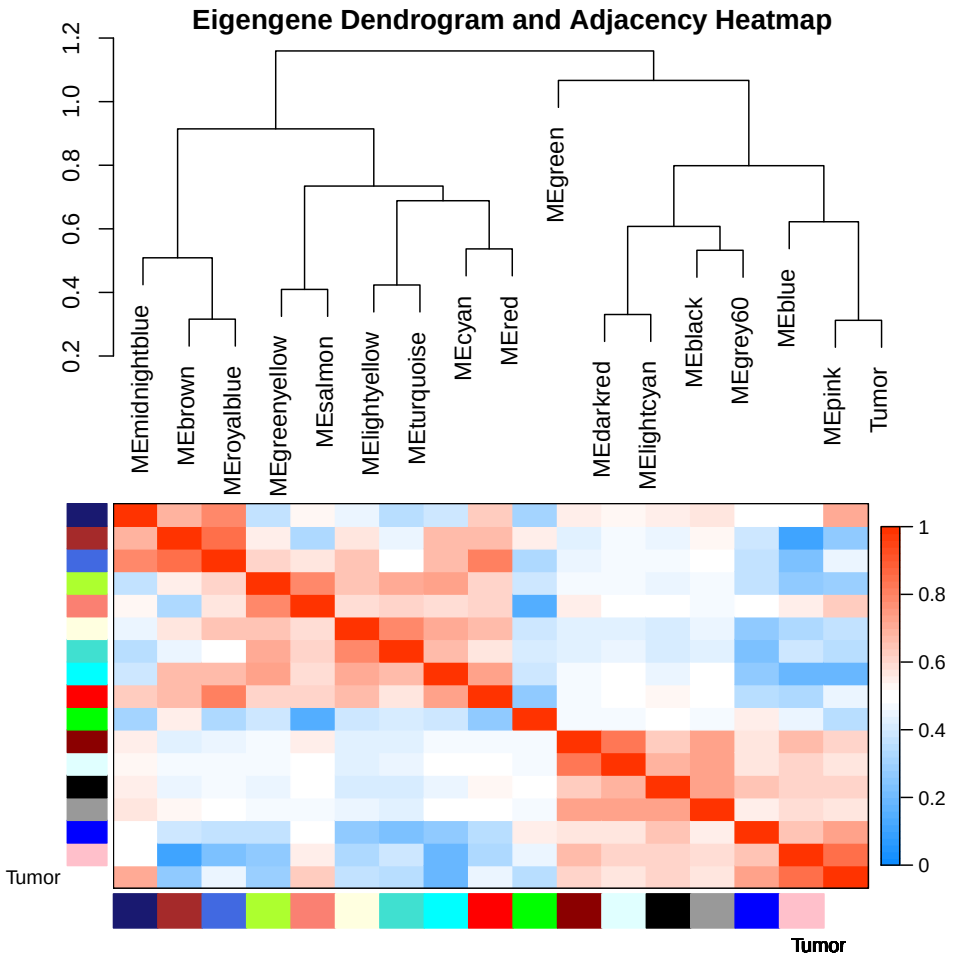

Figure A.36: Eigengene Network for STAD

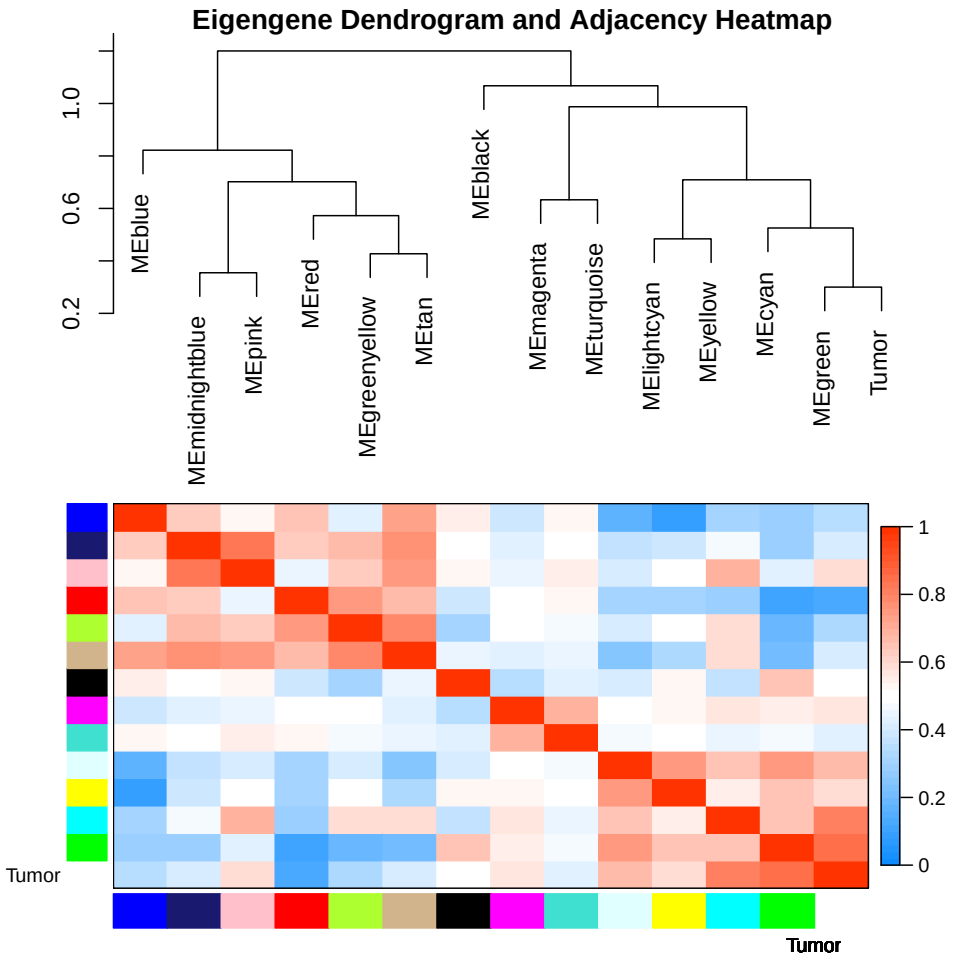

Figure A.37: Eigengene Network for ESCA

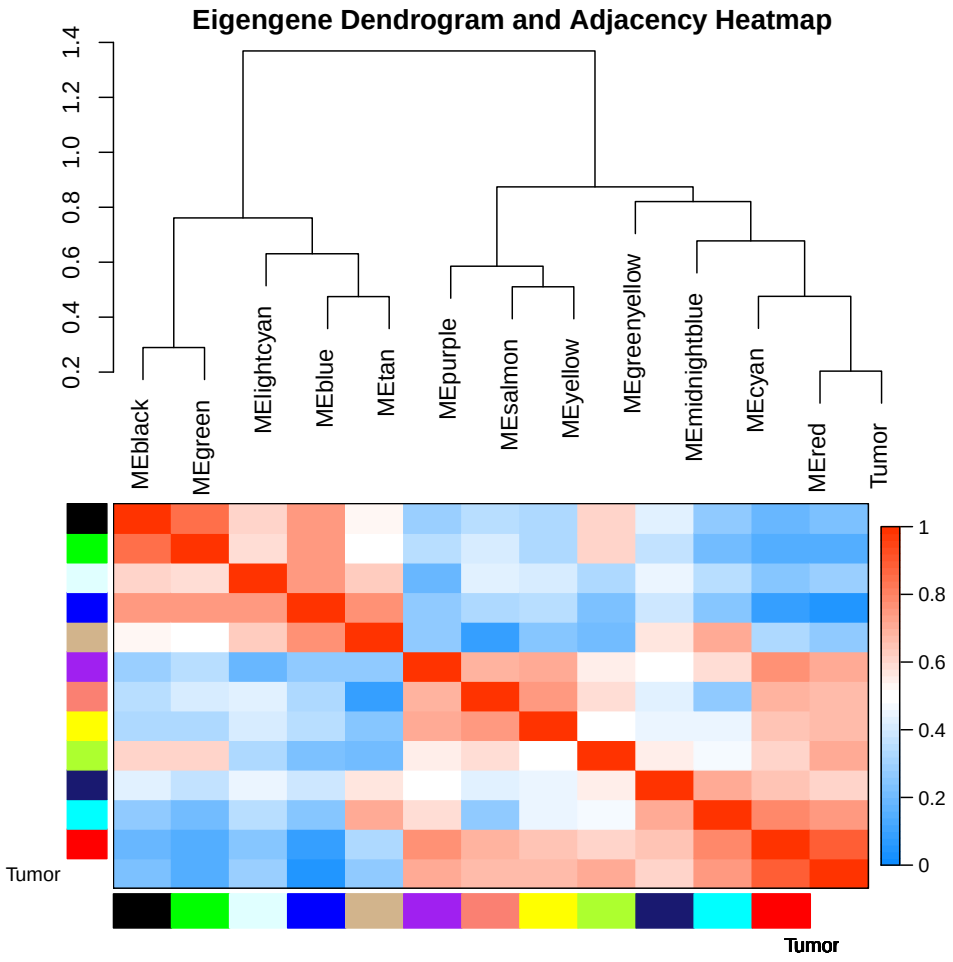

Figure A.38: Eigengene Network for COAD

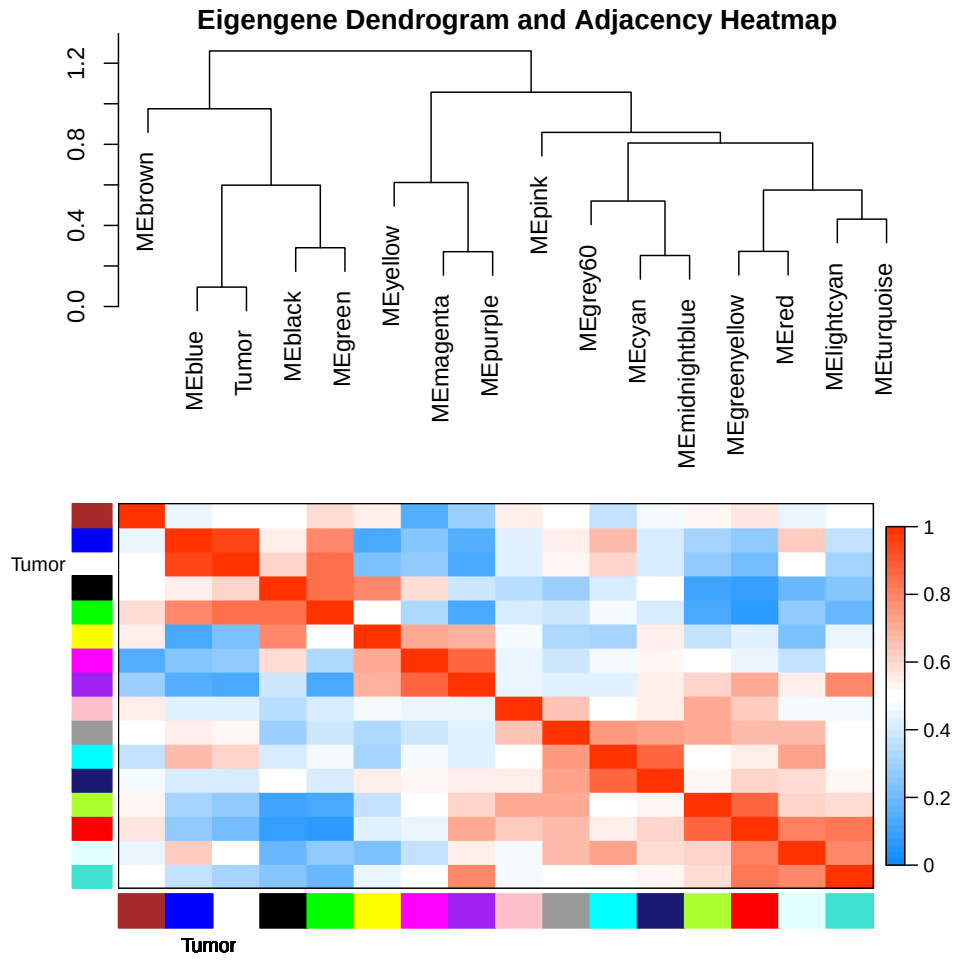

Figure A.39: Eigengene Network for HNSC

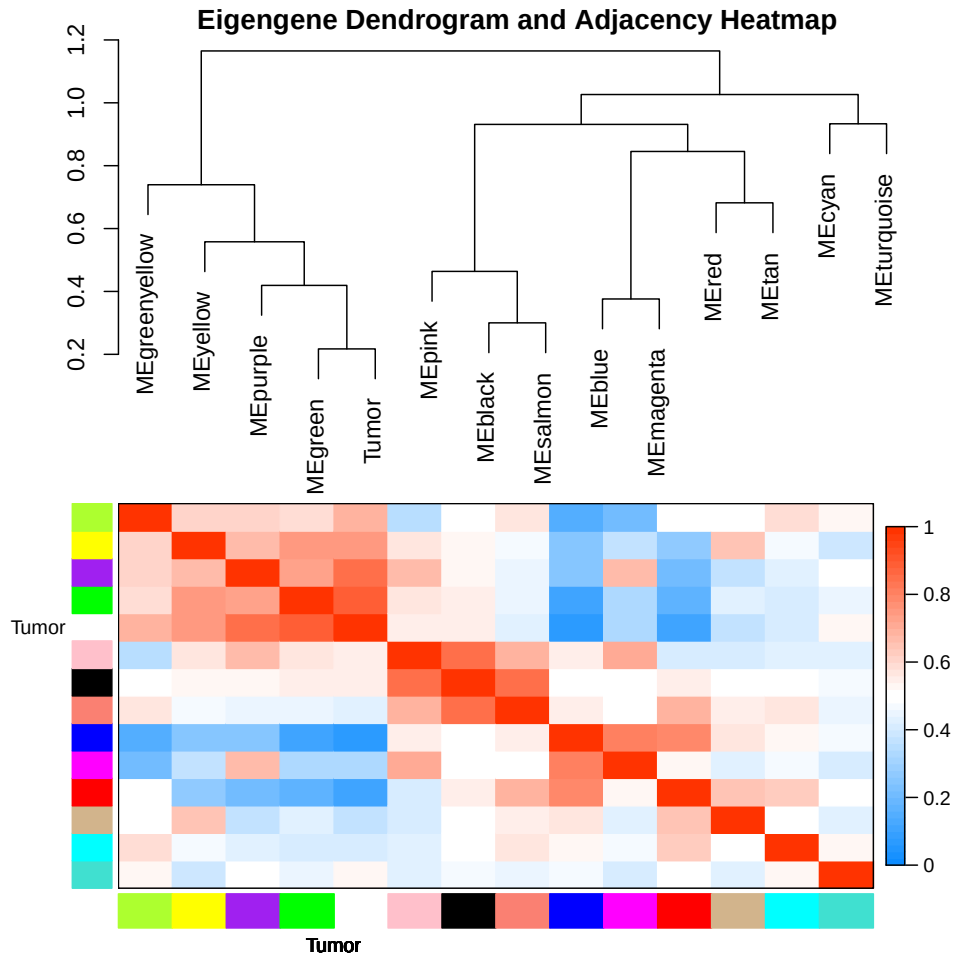

Figure A.40: Eigengene Network for BRCA