SURVIVAL ANALYSIS AND ITS APPLICATIONS IN IDENTIFYING GENES, SIGNATURES, AND PATHWAYS IN HUMAN CANCERS

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF BILKENT UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN MATERIALS SCIENCE AND NANOTECHNOLOGY

AYŞE ÖZHAN

SEPTEMBER 2021

Survival analysis and its applications in identifying genes, signatures, and pathways in human cancers

Ayşe Özhan
September 2021

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Özlen Konu Karakayalı (Advisor)

Michelle M. Adams

Ceren Sucularlı

Bala Gür Dedeoğlu

H. Hulusi Kafalıgönül

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

Survival analysis and its applications in identifying genes, signatures, and pathways in human cancers

Ayşe Özhan
Ph.D. in Materials Science and Nanotechnology
Özlen Konu Karakayalı
September 2021

Cancer literature makes use of survival analyses focused on gene expression based on univariable or multivariable regression. However, there is still a need to understand whether a) incorporating exon or isoform information on expression would improve estimation of survival in cancer patients; and b) applying multivariable regression to gene sets would allow to obtain cancer-specific independent gene signatures in cancer. Differential usage of individual exons, as well as transcripts, are phenomena common to cancerous tissue when compared to normal tissue. The glioblastoma, GBM; liver cancer LIHC; stomach adenocarcinoma, STAD; and breast carcinoma, BRCA datasets from The Cancer Genome Atlas (TCGA) were investigated to identify individual exons and transcripts with transcriptome-wide impact and significance on survival. Aggregation analyses of exons revealed the important genes for survival in each dataset, including *GNA12* in STAD, *AKAP13* in LIHC and *RBMXL1* and *CARS1* in BRCA. GSEA was applied on gene sets formed from the exon-based analysis, revealing distinct enrichment profiles for each dataset as well as overlaps for certain GO terms and KEGG pathways. In the second focus of this thesis, multivariable analyses on gene sets whose expressions were obtained from UCSC Xena were used to create two Shiny applications: one for dataset-specific analyses and one for analyses across TCGA-PANCAN. The dataset specific SmulTCan application incorporates Cox regression analyses with expressions of input genes of the user's choice. The SmulTCan application contains additional model validation, best subset selection and prognostic analyses. The ClusterHR application performs clustering analyses with Cox regression results, while it can also be used for bicluster identification and comparison. The axon-guidance ligand-receptor gene sets Slit-Robo, netrins-receptors and Semas-receptors were used for demonstrating the apps. Several hazard ratio signatures and best subsets that can differentiate between prognostic outcomes have been identified from the input gene sets, as well as ligand-receptor pairs with prognostic significance.

Keywords: RNA-Sequencing, exon, transcriptomics, cancer, TCGA-PANCAN, survival, Cox regression, hierarchical clustering, gene-set enrichment, Shiny, prognosis, machine learning, axon guidance, noncoding RNA

ÖZET

Gen, im ve yolak saptanmasında sağkalım analizi ve uygulamaları

Ayşe Özhan
Malzeme Bilimi ve Nanoteknoloji, Doktora
Özlen Konu Karakayalı
Eylül 2021

Sağkalım analizi kanser literatüründe yoğun olarak kullanılan bir yöntem olmakla birlikte, ekson ve transkript ifadeleriyle yapılmasının bu yönteme olan etkisi araştırılmaktadır. Ayrıca gen setlerinden kanser sağkalımında etkili olan çok değişkenli gen imleri oluşturma da aktif bir araştırma alanıdır. Ekson ve transkriptlerin ayrımsal ifadesi kanser hücrelerinde normal hücrelere göre sıklıkla gözlemlenen bir olgudur. Kanser sağkalımı için transkriptom boyu etki ve önemi olan ekson ve transkriptler, kanser genom atlası (TCGA) araştırma ağı üzerinden glioblastoma multiforme (GBM), hepatoselüler karsinom (LIHC), mide adenokarsinom (STAD) ve invazif meme karsinomu (BRCA) veri setleri incelenerek belirlenmiştir. Sonrasında ekson agregasyon analizi ile her bir kanser veri seti için önemli olan genler belirlenmiştir; örneğin STAD veri setinde *GNA12*, LIHC veri setinde *AKAP13*, BRCA veri setinde *RBMXL1* ve *CARS1* genleri, vb. Gen seti zenginleştirme analizi (GSEA) ile her bir veri setinden oluşan gen setlerinin farklı zenginleştirme profilleri incelenerek, veri setleri için ortak ve ayrık olmak üzere GO terimleri KEGG yolakları belirlenmiştir. Gen setleri ile çok değişkenli sağkalım analizi için Shiny kullanılarak 33 kanser veri setini içeren TCGA Pan-Kanser (TCGA-PANCAN) üzerinden biri seçilen veri setine özel, diğeri aynı anda birden fazla kanser veri setinde analiz yapabilen için iki farklı uygulama geliştirilmiştir. Bu uygulamalardan SmulTCan kullanıcının belirlediği bir gen seti ile Cox regresyonu yapabilmekte, ayrıca Cox modeli geçerliliği analizlerinin yanısıra en iyi alt küme belirlemek ve bu alt küme için prognoztik indis hesaplamak için kullanılabilmektedir. Diğer uygulama ClusterHR ise kullanıcı tercihine göre tek veya çok değişkenli Cox regresyonu sonuçları ile kümeleme analizi yapabildiği gibi, ikili kümeleme ve farklı gen setlerinden oluşturulmuş ikili kümeleme sonuçlarını karşılaştırabilmektedir. Uygulamaların nasıl kullanıldığının gösterimi için üç farklı akson sevk etme ligand ve reseptörlerinden oluşan gen seti incelenmiştir. Gösterimler sonucunda prognoz ayrışması sağlayabilen çeşitli risk oranı gen imleri ve prognoztik önemi olan ligand-reseptör çiftleri belirlenmiştir.

Anahtar sözcükler: RNA dizileme, ekson, transkriptomik, kanser, TCGA-PANCAN, sağkalım, Cox regresyonu, hiyerarşik kümeleme, gen seti zenginleştirme, Shiny, prognoz, makine öğrenmesi, akson sevk etme, kodlamayan RNA

ACKNOWLEDGEMENT

I would like to express my gratitude towards my advisor, Dr Özlen Konu, for her continuous support and mind-opening discussions about research ideas throughout the time I spent working in her research group. I am also thankful to my thesis committee members for their help in completing successful meetings on Zoom, a recent concept for all of us, spread out in a relatively short period. UNAM provided the logistics where most of the coding and writing for this thesis took place, in particular MSN director Dr Dönüş Tuncel's support should not be left unmentioned. Not forgetting various friends and colleagues at UNAM and Bilkent, for the conversations and venting sessions supposed to be natural to the PhD process. I would also like to thank Melike Tombaz and Farid Ahadli for their help in double checking the data and the codes used in this thesis. I should also mention the valuable contributions of the thesis defense committee members, whose comments helped shape the thesis into its final form. Finally, my family, for providing me with the comforting and hopefully still Covid-free home environment during the write-up.

Contents

1 Introduction 1

1.1 Survival analysis and methodologies 1

1.1.1 Kaplan-Meier analysis and Mantel–Cox test 2

1.1.2 Cox proportional hazards 3

1.2 Cancer dataset sources 3

1.2.1 The Cancer Genome Atlas 3

1.2.2 Genomic Data Commons 3

1.3 Aims and rationale of the thesis 4

1.3.1 Survival analysis in cancer 4

1.3.2 Background 4

1.4 Thesis structure 6

1.4.1 Chapters 2 and 3 6

1.4.2 Chapter 4 8

2 Univariable survival analysis of four TCGA datasets with exon expressions and enrichment analysis 9

2.1 Introduction 9

2.1.1 RNA-Sequencing 9

2.1.2 TCGA datasets used in the exon-based analyses 10

2.1.3 Long noncoding RNAs and cancer 10

2.1.4 Alternative exon usage in cancer 11

2.1.5 Aims 12

2.2 Methods 12

2.2.1 Data acquisition 12

2.2.2 Pre-processing 12

2.2.3 Data analysis and interpretation 12

2.2.4 Enrichment analysis 13

2.3 Results 14

2.3.1 Stomach adenocarcinoma 14

2.3.2 Glioblastoma multiforme 16

2.3.3 Liver hepatocellular carcinoma 19

2.3.4 Breast invasive carcinoma 22

2.3.5 Analysis of gene and lincRNA sets for the four datasets 25

2.3.6 Enrichment analysis of gene sets 27

2.4 Discussion 30

3 Univariable survival analysis of four TCGA datasets with isoform expressions 37

3.1 Introduction 37
3.1.1 Alternative splicing 37
3.1.2 TCGA datasets used in the transcript-based analyses 38
3.2 Methods 38
3.2.1 Data acquisition 38
3.2.2 Pre-processing 39
3.2.3 Data analysis and interpretation 39
3.2.4 Aim 39
3.3 Results 39
3.3.1 Stomach adenocarcinoma 39
3.3.2 Glioblastoma multiforme 43
3.3.3 Liver hepatocellular carcinoma 44
3.3.4 Breast invasive carcinoma 46
3.4 Discussion 48
4 Multivariable analyses and application development with Shiny 50
4.1 Introduction 50
4.1.1 University of California Santa Cruz Xena 50
4.1.2 Best subset selection 52
4.1.3 Prognostic outcome analysis 52
4.1.4 Clustering, biclustering and cluster comparison 53
4.1.5 MicroRNAs 54
4.1.6 Axon guidance molecules in cancer 55
4.1.7 TCGA datasets, sample sizes and survival types 60
4.1.8 Shiny 62
4.1.9 Aims 62
4.2 Methods 63
4.2.1 Data acquisition 63
4.2.2 Data analysis and interpretation 63
4.2.3 Warnings 65
4.2.4 Downloads 65
4.3 Results 65
4.3.1 SmulTCan Shiny application 65
4.3.2 ClusterHR Shiny application 78
4.4 Discussion 83
4.4.1 SmulTCan 84

4.4.2 ClusterHR 84
4.4.3 Slit-Robo 84
4.4.4 Netrins-receptors 85
4.4.5 Semas-receptors 86
5 Conclusion and future perspectives 86

List of Figures

Figure 1: Survival function representing survival probability for a two-year period. 1
Figure 2: K-M analysis of the (A) "colon" and (B) "lung" datasets divided into the two sexes. Males and 2
Figure 3: Representation of the structure of a human gene, with exons interrupted by introns. 4
Figure 4: Effect of TMM normalization on STAD samples. (A) Distribution of logCPM expressions of raw or unnormalized samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 14
Figure 5: Effect of TMM normalization on GBM samples. (A) Distribution of logCPM expressions of raw or unnormalized samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 17
Figure 6: Effect of TMM normalization on LIHC samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 20
Figure 7: Effect of TMM normalization on BRCA samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 23
Figure 8: Venn diagrams showing the numbers of genes in percentage that were found to be overlapping or distinct between the four datasets after the filtering and aggregation steps. (A) Genes that were designated hazardous with mean HRs > 1. (B) Protective genes with mean HRs < 1. (C) Hazardous lincRNAs with mean HRs >1. (D) Protective lincRNAs with mean HRs < 1. 26
Figure 9: GSEA plot for STAD. The top three GO terms with the most significant enrichment scores are indicated with colors red, green, and blue from first to third. These terms were negatively enriched for the protective genes of this dataset. 27
Figure 10: GSEA plot for LIHC indicates that mainly the hazardous genes in the gene set have been enriched for terms related to the mitotic cell cycle and RNA splicing. 27
Figure 11: GSEA plot for BRCA indicates enrichment for GO terms related to vesicle-mediated transport. 28
Figure 12: GSEA plot for GBM indicated negative enrichment of protective genes for the listed GO terms. 28
Figure 13: Emap plot for KEGG pathway enrichment of each dataset revealed several common pathways for the hazardous genes of these datasets; such as nucleocytoplasmic transport and aminoacyl-tRNA biosynthesis for LIHC-BRCA, i.e. genes with mean HR > 1 based on exon expressions. 29
Figure 14: Emap plot for KEGG pathway enrichment of each dataset indicates that the spliceosome pathway as common for datasets GBM, STAD and BRCA, for their protective genes, i.e. genes with mean HR < 1 based on exon expressions. 30
Figure 15: K-M plot with *GNA12* expressions for survival rates in STAD. The blue curve indicates the survival rate for samples with less than the median ($= 10.76$) expression for *GNA12* ($n = 219$), while the red curve indicates survival for samples with more than or equal to the median expression of the gene ($n = 224$). 31
Figure 16: K-M plot with *CSNK1D* expressions for survival rates in GBM. The blue curve indicates the survival rate for samples with less than the median ($= 11.91$) expression for *CSNK1D* ($n = 85$), while the red curve indicates survival for samples with more than or equal to the median expression of the gene ($n = 81$). 32
Figure 17: Comparison of normalized expressions of *AKAP13* in TCGA-LIHC for primary tumors vs. normal tissue. 33
Figure 18: Comparison of normalized expressions of *CARS1* in TCGA-BRCA between primary tumor and normal samples. 34

Figure 19: Different transcripts of the same gene, produced by alternative splicing. In the left transcript, the fourth exon is excluded in the mRNA; in the transcript in the middle, the second is excluded, while the third and fourth exons have switched orders in the transcript on the right. 37
Figure 20: Effect of TMM normalization on STAD isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.....40
Figure 21: Genes with transcripts that have opposite effects on survival in STAD. (A) *CLK1* with two transcripts of opposite impact on survival rates. (B) One hazardous and one protective transcript for *BICD2* were identified. (C) *ESYT2* has a hazardous and protective transcript with transcriptome-wide significance. Univariable significance levels for the isoforms were indicated on each bar with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. 42
Figure 22: Effect of TMM normalization on GBM isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 43
Figure 23: Effect of TMM normalization on LIHC isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples. 44
Figure 24: Two different transcripts of *AKAP8L* with opposing effects on survival in TCGA-LIHC. Univariable significance levels for the isoforms were indicated on each bar with ** for $p < 0.01$. 45
Figure 25: Effect of TMM normalization on BRCA isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.....46
Figure 26: Genes with transcripts that were found to have opposite effects on survival. (A) *DDX6* had two significantly associated transcripts with survival in BRCA. (B) *EEF1B2* also had two significantly associated transcripts with survival in BRCA. Univariable significance levels for the isoforms were indicated on each bar with * for $p < 0.05$ and ** for $p < 0.01$. 47
Figure 27: UCSC Xena user interface after the selection of genes. In Panel A on the left, the TCGA-PANCAN study was selected, indicating the total number of samples in the study, which is 12,839. In Panel B in the middle, the log2$(norm_value + 1)$ expression values of the selected Slit-Robo genes are listed.....50
Figure 28: Screenshot from Xena depicting log2$(norm_value + 1)$ distributions of each Slit-Robo gene across TCGA-PANCAN. 51
Figure 29: Diagram of Sema signaling, where plexinAs (blue) receive signals from Sema3s and Sema6s (red). Sema3s also require neuropilins (green) to bind plexins. (A) Sema3-neuropilin-plexin complex requires IgCAMs for axon guidance. This complex also regulates actin and microtubule dynamics, as well as endocytosis. (B) When Sema3E binds directly to PlexinD1, it results in endothelial cell and axon repulsion. Involvement of neuropilin in this interaction results in axon attraction. (C) Sema6s bind plexinAs during neural development, eliciting axon repulsion. (D) Sema6D can bind plexinA1 to elicit endothelial cell attraction through OTK interaction and repulsion through VEGFR2 interaction. Sema6D can also regulate immune cell and osteoclast function through Trem2. Figure taken from Zhou et al. (2008), see Permissions at the end of this document. 56
Figure 30: Diagram of signaling through Sema4s and Sema5s (red) through common and distinct receptors. (A, B) Sema4D binds to CD72 and Sema4A binds to Tim-2 receptors located on immune cells. (C) Sema4A binding to PlexinD1 inhibits angiogenesis. (D) The Sema4D-PlexinB1 complex mediates cell migration and invasive growth through interaction with Met and cell migration and growth cone collapse through interaction with ErbB2. (E) Sema5A binding to plexinB3 promotes cell migration through repulsion, while invasive growth is regulated through Met. Axon guidance through Sema5A requires proteoglycans HSPG and CSPG. Figure taken from Zhou et al. (2008), see Permissions at the end of this document. 57
Figure 31: Slit-Robo signaling pathway and downstream targets. Slits also contain a Cystein-rich C-terminal domain shown as a green circle. Fibronectin type-III (FN3) domains on Robos are indicated with purple rectangles. Figure taken from Tong et al. (2019) , see Permissions at the end of this document. 58
Figure 32: (A) Cross-sectional view of the early embryonic neural tube, from which the spinal cord and brain are developed. The ventricular zone (VZ) and floor plate (FP) express netrin (blue) and the roof plate (RP) expresses draxin (red). (B) Structural domains of the guidance cues netrin and draxin. (C) Binding sites for

netrin, draxin and slit are indicated on each receptor. Figure taken from Meijers et al. (2020) , see Permissions at the end of this document.....59
Figure 33: App architecture for SmulTCan indicates the two main parts of the app, model analysis and best subset selection. Reactivity in the app is shown with double-headed arrows. Functions required for the computations in each main tab, imported from existing packages for CPH model analysis are indicated on the arrows. Figure taken from Ozhan et al. (2021), see Permissions at the end of this document.....66
Figure 34: The SmulTCan app interface includes a side-bar panel where users can upload the file containing expressions of genes they would like to analyze, which they obtained from UCSC Xena earlier. Distribution of Slit-Robo expressions in TCGA-MESO.....67
Figure 35: Screenshot from SmulTCan showing the forest plot of ESCA for DSS. The "Genes" menu in the side-bar panel indicates that all seven members of Slit-Robo have been selected.....67
Figure 36: Results of ANOVA analysis of multivariable CPH models built with input genes sets can be visualized from the "ANOVA" tab's "Plot" sub-tab. 68
Figure 37: Forest plot of netrins-receptors in COAD indicates positive association of *UNC5C* and negative association of *NTN3* with OS. 69
Figure 38: ROC plot for OS in COAD, whereby the CPH model includes all twelve netrins-receptors. The plot indicates the CPH model is a reasonably good predictor of prognosis. 70
Figure 39: ANOVA results of CHOL for DSS indicate *DCC* as the best predictor among netrins-receptors. 71
Figure 40: Plot of the coefficients of the best subset of netrins-receptors determined with the default parameters of the BeSS algorithm contain eleven genes in THYM for OS. 72
Figure 41: Plot of ANOVA results in THYM for OS with netrins-receptors. The best predictor of the gene set in this case is *DSCAM*. 73
Figure 42: Forest plot of ACC for OS indicates a strongly protective role for *NRP2* and *SEMA4D* among Semas-receptors.....74
Figure 43: K-M plots of KIRC for (A) OS, (B) DSS, and (C) PFI obtained from SmulTCan for the Slit-Robo genes. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type with respect to the log-rank test. 75
Figure 44: K-M plots of KIRP obtained from SmulTCan for (A) OS, (B) DSS, and (C) PFI with Slit-Robo. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type with respect to the log-rank test's results.....75
Figure 45: K-M plots of LGG for A) OS, (B) DSS, (C) DFI, and (D) PFI. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type. 76
Figure 46: K-M plot of DSS in LGG for the best subset of Semas-receptors gene set identified with lasso.....77
Figure 47: Cumulative hazards plot of DSS in LGG for the best subset of Semas-receptors identified with lasso has differentiated between prognostic outcomes. 77
Figure 48: ROC plot of the PI covariate of the best subset consisting of *SEMA4G* and *SEMA6B* in LGG.....78
Figure 49: Screenshot of the ClusterHR app indicating the selected thirteen TCGA datasets on the side-bar panel. The "Multivariable matrix" sub-tab of the OS sub-tab of the "Heatmaps" main tab was selected. The user can select between clustering options, the distance methods in the clustering and the hierarchical clustering methods using the internal side-bar panel.....79
Figure 50: Slit-Robo multivariable $\log_2$HR values across thirty-one TCGA datasets displayed in a clustering heatmap. 80
Figure 51: Multivariable analysis $\log_2$HR values computed from netrins-receptors across twenty-six TCGA datasets are displayed in the clustering heatmap. 81
Figure 52: Multivariable analysis $\log_2$HR values calculated from Semas-receptors across nineteen TCGA datasets are displayed in the clustering heatmap. 82
Figure 53: Bicluster generated from the Semas-receptors gene set with Cheng and Church's algorithm, where all members of the Semas and their receptors were included. There are seventeen TCGA datasets in this bicluster. 82

Figure 54: Another bicluster from Semas-receptors; this one contains nine genes from the gene set. We can observe that this subset of genes behaves differently in terms of HRs in PAAD and KIRC. For these two TCGA datasets, the nine members of Semas-receptors have similar $\log_2$HR profiles. 83

List of Tables

Table 1: Sample sizes of TCGA datasets used in the exon-based analyses, consisting of primary tumors for each dataset. 10
Table 2: Exons with the highest and lowest HR values in the coding hg19 transcriptome for TCGA-STAD with individual transcriptome-wide significance. 15
Table 3: Exons with the five highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-STAD. 15
Table 4: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation with significant exons for TCGA-STAD. 16
Table 5: Exons with the highest and lowest HR values in the coding hg19 transcriptome for TCGA-GBM, which have passed the $\log CPM > 1$ count filter of 108 for this dataset. 17
Table 6: Exons with five of the highest and lowest HR values in the noncoding hg19 transcriptome in TCGA-GBM. The listed exons have been obtained after filtering out those with insufficient $\log CPM > 1$ counts in the GBM dataset. 18
Table 7: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-GBM. 19
Table 8: Exons with the highest and lowest HR values in the coding hg19 transcriptome in TCGA-LIHC. Exons with the ten highest and ten lowest HRs with transcriptome-wide significance that have passed the $\log CPM > 1$ count filter were listed. 20
Table 9: Exons with the highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-LIHC. 21
Table 10: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-LIHC. 22
Table 11: Exons with the highest and lowest HR values in the coding hg19 transcriptome in TCGA-BRCA. Ten highest and ten lowest HRs with transcriptome-wide significance were indicated. 23
Table 12: Exons with the highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-BRCA. 24
Table 13: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-BRCA. 24
Table 14: Number of genes from the exon-based analysis after the filtering and aggregation steps for each cancer dataset. 25
Table 15: Number of lincRNAs from the exon-based analysis of the noncoding transcriptome after the filtering and aggregation steps for each cancer dataset. 26
Table 16: Sample sizes of TCGA datasets used in the transcript-based analyses, consisting of primary tumors for each dataset. 38
Table 17: Transcripts with the ten highest and lowest HR values whose expressions have transcriptome-wide significance for survival in TCGA-STAD. 40
Table 18: Transcripts identified in GBM with the highest and lowest HRs whose expressions associated either positively or negatively with survival. 43
Table 19: Transcripts with the highest and lowest HRs with transcriptome-wide significance in TCGA-LIHC. 45
Table 20: Transcripts associated with survival univariably in TCGCA-BRCA with the ten highest and lowest HR values. 46

Table 21: Ten rows of a downloaded example Xena TSV file containing TCGA samples and their expression values of netrins, as well as TCGA study names and tumor type information. This column structure is required as input for the apps in this study. 51
Table 22: The column format of a survival TXT file from GDC. The table shows ten TCGA-LIHC samples and their survival information, including event and time-to-event, for four different types of survival. 60
Table 23: TCGA-PANCAN total sample sizes for each survival type incorporated in the Shiny applications...61
Table 24: ANOVA results table of CHOL for DSS, generated from the downloaded output TSV file from SmulTCan.....71

1 INTRODUCTION

Cancer constitutes a group of diseases characterized by the uncontrolled and unwanted growth of abnormal cells that can potentially invade or spread to other parts of the body. It is a serious threat to human life and has become a leading cause of death in humans in recent years (Si et al., 2019). Survival analysis and its applications constitute a group of statistical approaches commonly used in understanding the underlying causes and identifying better treatments for different types of cancer. In this chapter, I will provide a general introduction to survival analyses methods, the survival types, as well as pan-cancer expression datasets that contain survival information and tools to process them.

1.1 SURVIVAL ANALYSIS AND METHODOLOGIES

Survival analysis, or time-to-event analysis, is defined as a set of statistical approaches dealing with the duration of time until one or more events happen. The survival times of members of a group can be described by the survival function, denoted S, also known as the reliability function, which represents the probability of an individual (usually a patient) surviving beyond time t. The S function is also known as the complementary distribution function $F(t)$, and hence represented by Equation 1 (Rodríguez, 2007) on the interval $[0, \infty)$:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t) \quad (1)$$

The S function usually takes the form of the plot in Figure 1; such that point P represents the probability of surviving beyond 4.8 months, which in this case is 88%.

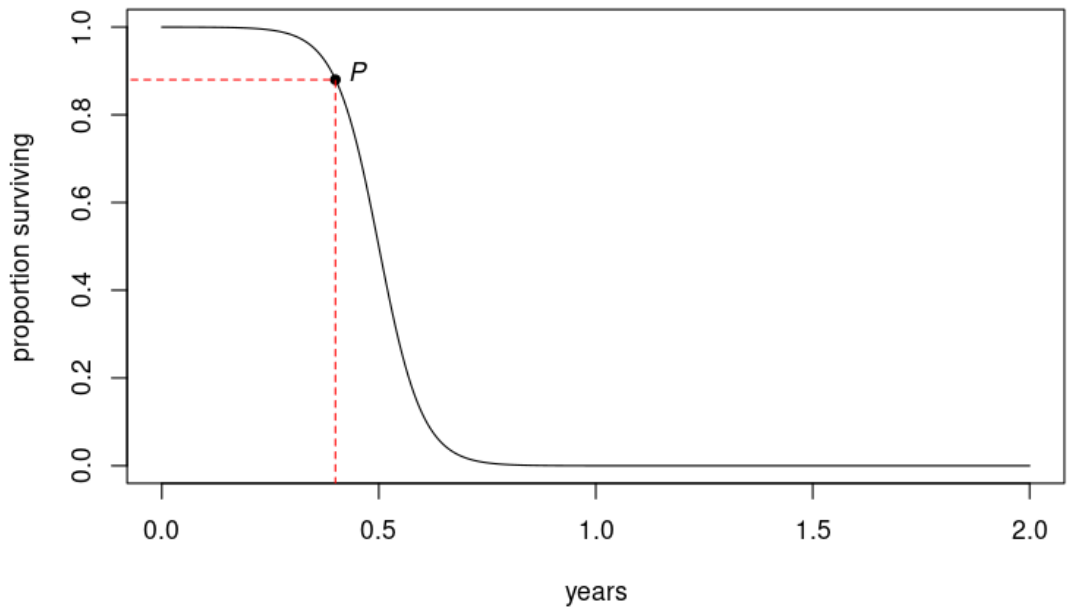

Figure 1: Survival function representing survival probability for a two-year period.

1.1.1 KAPLAN-MEIER ANALYSIS AND MANTEL–COX TEST

Survival analysis is a method commonly used in cancer research, usually by estimating the survival function with the Kaplan-Meier (also called product limit) method, where the estimator is given by Equation 2, where t_i represents a time when at least one event happened, d_i is the number of deaths that happened at time t_i and n_i represents the number of individuals to have survived up to time t_i:

$$S(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \quad (2)$$

The Kaplan-Meier (K-M) estimator can be used to group patients into categories, such as carrier status of a gene variant or smoker status, and to understand the differences in survival rates of two categories using the Mantel–Cox (or log-rank) test. Let $1, \ldots, J$ be the distinct times of the events observed in the groups, $N_{1,j}$ and $N_{2,j}$ the number of samples in each group that the event had not yet happened to, $O1,j$ and $O_{2,j}$ the observed number of events in the groups at time j, and $N_j = N_{1,j} + N_{2,j}$ and $O_j = O_{1,j} + O_{2,j}$. Then, for each group $i = 1, 2$, $E_{i,j} = N_{i,j}\frac{O_j}{N_j}$ represents the expected value and $V_{i,j} = E_{i,j}(\frac{N_j - O_j}{N_j})(\frac{N_j - N_{i,j}}{N_j - 1})$ represents the variance. The log-rank statistic is approximated by the χ^2 distribution and given by $\frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}}$.

Example K-M plots using the "survival" package's "colon" and "lung" datasets (Therneau & Grambsch, 2000) is given in Figure 2, where patients are divided into their respective sexes as "Male" and "Female". With the resulting p-value from the K-M analysis being > 0.05 for the "colon" dataset and < 0.05 for the "lung" dataset, we can conclude that sex is not a determining factor for survival time in colon cancer, while it is a determining for the input dataset "colon".

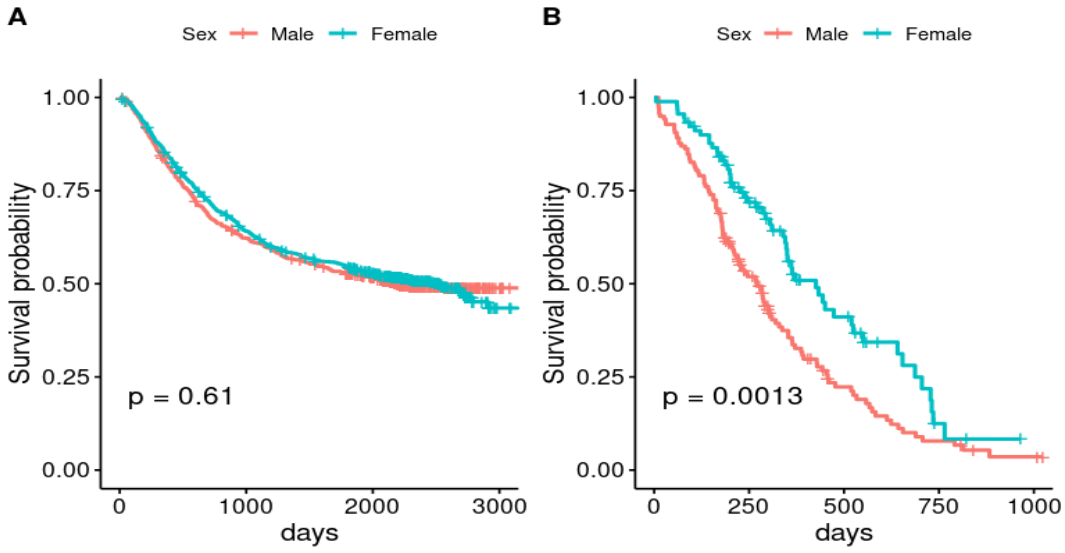

Figure 2: K-M analysis of the (A) "colon" and (B) "lung" datasets divided into the two sexes. Males and females have similar survival rates in colon cancer, while prognosis is significantly poorer for males in lung cancer.

1.1.2 COX PROPORTIONAL HAZARDS

The Cox proportional hazards (CPH) model is a survival model which assumes the proportional hazards condition, whereby covariates are multiplicatively related to the hazard. The regression model was developed in 1972 by Sir David Cox (Brembilla et al., 2018), where the purpose is to evaluate the simultaneous effect of several factors on survival. The CPH model can be applied to gene expressions. Let n represent sample size and p the number of genes in the multivariable model. Then, $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]$ is the vector of the expression level of p genes. Then the hazard function $\lambda(t)$ of p genes is given by Equation 3:

$$\lambda(t) = \lambda_0 e^{(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)} = \lambda_0 e^{(\beta^T \mathbf{X})} \quad (3)$$

where $\lambda_0(t)$ is the baseline function (unspecified), $\beta = [\beta_1, \beta_2, \ldots, \beta_p]$ is the vector of regression coefficients and $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ is the vector of gene expressions (Abdul Aziz et al., 2016). Therefore, the HR of the i^{th} gene is given as e^{β_i}. Accordingly, a gene with an HR > 1 is positively associated with event probability, and thus negatively associated with the length of survival; whereas a gene with an HR < 1 is negatively associated with event probability and positively associated with the length of survival. An HR $= 1$ is neither negatively nor positively associated with event probability or survival.

1.2 CANCER DATASET SOURCES

1.2.1 THE CANCER GENOME ATLAS

The Cancer Genome Atlas (TCGA) is a milestone cancer genomics program (https://www.cancer.gov/tcga/) provided by the National Cancer Institute (NCI), a United States government funded agency which aims to advance cancer detection and diagnosis, resulting in longer and healthier lives for cancer patients. TCGA was initiated in 2006 as a pilot project and was authorized to full production phase in 2009 (NCI, 2021). Together with the efforts of the National Human Genome Research Institute (NHGRI), TCGA collected samples from more than 12,500 patients across 33 tumor types to generate an extensive and comprehensive dataset consisting of up to 2 petabytes (PBs) of data, for better understanding the molecular changes that occur in cancer. TCGA consists of five molecular data platforms: mRNA expression, DNA methylation, microRNA (miRNA) expression, copy number variations/alterations, and single nucleotide variants, which are all trusted and publicly available references. TCGA efforts were culminated in the Pan-Cancer Atlas project (TCGA-PANCAN), where samples from all tumor types were uniformly analyzed (Bailey et al., 2018).

1.2.2 GENOMIC DATA COMMONS

The NCI's Genomic Data Commons (GDC) (https://portal.gdc.cancer.gov/) is a data sharing platform (Langmead & Nellore, 2018) aiming to unify the cancer genomics research community by creating a knowledge network encompassing standardized genomic and clinical data from various cancer research programs including TCGA and

Therapeutically Applicable Research to Generate Effective Therapies (TARGET). The vast archive of GDC's legacy site can be accessed through https://portal.gdc.cancer.gov/legacy-archive/ where it is possible to search for datasets of different types of cancer and obtain raw RNA-Seq counts information for samples of specific cancer types, which could then be subjected to pre-processing steps through a workflow of choice. RNA-Seq samples in the GDC legacy archive have been aligned against genome version hg19 (GRCh37).

1.3 Aims and rationale of the thesis

1.3.1 Survival analysis in cancer

Survival analysis and its applications constitute a group of statistical approaches commonly used in understanding the underlying causes and identifying better treatments for different types of cancer. The overall aim of this thesis was to use univariable and multivariable survival analysis methods with exon, isoform, and gene expressions to identify important exons, isoforms, genes, and gene signatures in different cancers. The rationale of these aims is further explained in the following sections.

1.3.2 Background

1.3.2.1 Exon-based survival analysis

The structure of a typical human gene is given in Figure 3, which consists of a set of exons interrupted by introns. While all the elements in the below figure are transcribed into RNA, only the exons are translated and appear in the final protein product of the gene. Therefore, expression levels of exons are an important determinant on the functionality of a gene and could be used to understand the importance of the gene that they belong to on the viability of the cell. The idea behind the analyses in Chapter 2 was to perform univariable survival analyses with exon expressions which would help identify important exons in the transcriptome by ranking the resulting HR values that were filtered for transcriptome-wide significance ($p < 0.05$).

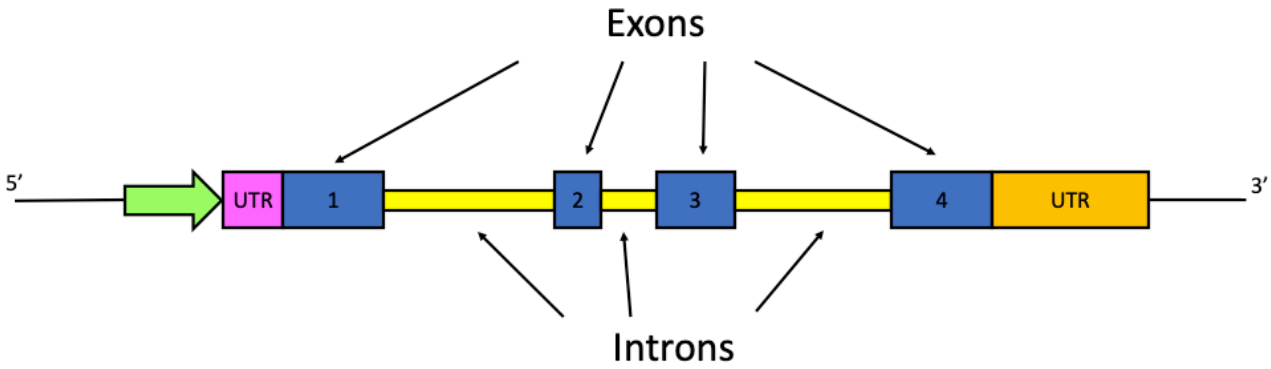

Figure 3: Representation of the structure of a human gene, with exons interrupted by introns.

1.3.2.2 ISOFORM-BASED SURVIVAL ANALYSIS

Cancerous tissues are known to have different expression levels of exons and transcripts of specific genes compared to normal tissue, which affects tumor formation and proliferation (Bjorklund et al., 2017). Known as alternative splicing, the process also accounts for the plasticity of the human genome, whereby almost all genes with multiple exons go through (Marden, 2008). Cancer cells exploit alternative splicing to switch to producing isoforms that would benefit their proliferation, metastasis, and survival (Chen & Weiss, 2015). Due to its role in oncogenesis, alternative splicing has been the focus of previous studies such as in the SURVIV statistical model, whereby the hazard function was associated with exon-inclusion levels, which was modeled by the binomial function to incorporate the estimation uncertainty (Shen et al., 2016). Authors of SURVIV have argued that their model better predicts survival rates than conventional Cox regression with isoform ratios formed with expression levels.

Given the importance of alternative splicing in cancer and its role in the viability of tumor cells, identifying isoforms (or transcripts) of impact in the coding transcriptome is crucial. Therefore, in Chapter 3, univariable survival analyses were performed with isoform expressions in selected cancer datasets and the results were ranked based on their HR values filtered for transcriptome-wide significance ($p < 0.05$). The identified isoforms with high impact in each cancer dataset could be studied further experimentally to determine their role in the dataset they were identified from.

1.3.2.3 GENE SET-BASED MULTIVARIABLE SURVIVAL ANALYSIS

Cancer develops through the interaction of multiple genes and researchers may want to understand the comparative role of each gene within a set of genes with respect to each other; as opposed to trying to understand the effect a single gene against all other genes. Therefore, developing applications that can be used by researchers from a wide range of backgrounds with different interests for analyzing the effects of gene sets on cancer survival rates is a crucial task. Most of the available online survival analysis tools carry out univariable analyses with expressions, including GEPIA (http://gepia.cancer-pku.cn/), built with HTML5 and JavaScript, for univariable K-M plots along with exploratory graphics (Tang et al., 2017). There is also PROGgeneV2 (http://www.progtools.net/gene/), which accepts genes, gene ratios or gene signatures (GSs) as input to generate expression-based K-M plots (Goswami & Nakshatri, 2014), is another tool for analyzing TCGA and the Gene Expression Omnibus (GEO) (Barrett et al., 2013) datasets.

The online KM plotter (https://kmplot.com/analysis/) tool (Gyorffy, 2021) is used extensively in cancer research for generating K-M plots with gene and microRNA (miRNA) expression levels in TCGA datasets. An interactive tool for univariable protein-level based survival analysis is TRGAted (https://nborcherding.shinyapps.io/TRGAted/), useful for identifying biomarkers across TCGA-PANCAN (Borcherding et al., 2018). There also exist online tools for multivariable survival analysis such as SurvExpress (http://bioinformatica.mty.itesm.mx/SurvExpress/; Aguirre-Gamboa et al. (2013)) and

SurvMicro (http://bioinformatica.mty.itesm.mx/SurvMicro/, Aguirre-Gamboa and Trevino (2014)) for gene and miRNA sets, respectively. However, a highly interactive tool with publication quality visualization tools that also allows for multivariable regression of gene sets with abilities for best subset selection is missing in the literature. Moreover, when such regression is applied to all possible cancer datasets at once it would be possible to visualize the cancer types according to their survival qualities and quantities using clustering methodologies.

1.4 THESIS STRUCTURE

1.4.1 CHAPTERS 2 AND 3

In these two chapters, univariable survival analyses were used to address the potential of using exon and isoform data from the GDC legacy archive. For this, four TCGA datasets GBM for glioblastoma multiforme, STAD for stomach adenocarcinoma, LIHC for liver hepatocellular carcinoma, and BRCA for breast invasive carcinoma have been selected from TCGA, representative of each disorder due to their importance as disorders having a detrimental impact on human lives. In addition to the identification of individual exons with coding and noncoding transcriptome-wide impact and significance, an aggregation pipeline was developed in Chapter 2 to identify important genes and lincRNAs for survival in each dataset. From thereon, enrichment analyses for gene ontology (GO) terms and pathways were applied to each dataset's gene sets and their enrichment profiles were compared for (Kyoto Encyclopedia of Genes and Genomes) KEGG pathways. In Chapter 3, individual gene transcripts with transcriptome-wide impact emerged from analyses with the same four cancer datasets. The exon-based analyses in Chapter 2 and the isoform-based analyses in Chapter 3 were carried out with these four selected datasets. These cancer types were selected because they presented difficult to treat illnesses, had poor survival rates and/or was very commonly observed and therefore presented importance.

1.4.1.1 STOMACH CANCER

Stomach, or gastric cancer, is one of the leading causes of cancer-related deaths worldwide and identification of effective biomarkers for prognosis remains a crucial task (Zhang et al., 2018). Smoking and infection with *Helicobacter pylori* or the Epstein-Barr virus (EBV) have been given as environmental contributors to stomach cancer. Gastric adenocarcinoma (GAC), which develops from the innermost lining of the stomach (mucosa), makes up the most of stomach cancer cases. Hereditary forms of GAC exist, caused by germline mutations in the cadherin 1 gene *CDH1* (Oliveira et al., 2004) and the mismatch repair gene *STK11* (Shinmura et al., 2005). GAC is also known to arise from the deregulation of stemness by mutations/alterations in the Hippo, WNT/β-catenin, and Hedgehog pathways (Ajani et al., 2017). GAC was represented by the stomach adenocarcinoma (STAD) dataset of TCGA in this study.

1.4.1.2 LIVER CANCER

Liver cancer, of which the primary type is known as hepatocellular carcinoma (HCC), is also a leading cause of death worldwide with increasing frequency. About half of HCC cases are known to be caused by Hepatitis B virus (HBV) infection, while hepatitis C or D virus infection and excessive alcohol consumption are also listed as risk factors (Fujiwara et al., 2019). Driver mutations of HCC include promoter mutations in the *TERT* gene, which causes telomerase activation and mutations in *AXIN1*, *APC* or *CTNNB1* causing WNT/β-catenin signaling pathway activation is observed in ~40% of cases (Llovet et al., 2021). HCC is represented by the liver hepatocellular carcinoma (LIHC) dataset of TCGA in this study.

1.4.1.3 GLIOBLASTOMA

Glioblastoma multiforme, or grade IV astrocytoma is a highly aggressive brain tumor with causes many of which remain unclear. It constitutes ~15% of all brain tumors and affects astrocytes, which make up the main connective tissue of the brain (Weller et al., 2015). This cancer has a high recurrence rate and a poor survival rate, with a typical survival duration of 12-15 months following diagnosis (Krex et al., 2007). In the cytosine-phosphate-guanine (CpG) island methylator phenotype (G-CIMP) subtype of secondary glioblastoma, mutations in isocitrate dehydrogenase (IDH) enzymes are common, together with proneural gene expression (Han et al., 2020). The activation of oncogenic pathways including MAPK (Riddick et al., 2017) and AKT (Zhao et al., 2017), as well as impairment of p53 and retinoblastoma protein 1 (RB1)-regulated tumor-suppressive pathways such as apoptosis and the cell cycle are frequently observed across glioblastoma subgroups (Backlund et al., 2005; Lee et al., 2020). Hypoxia caused upregulation of hypoxia-inducible factors (HIFs) are known to promote aberrant angiogenesis through vascular endothelial growth factor (VEGF) in glioblastoma (Weller et al., 2015). Glioblastoma multiforme is represented by the GBM dataset of TCGA in this study.

1.4.1.4 BREAST CANCER

Breast cancer develops from the breast tissue, is a leading type of cancer in women accounting for about a quarter of all cancer cases. While the early-stage disorder is curable in most patients, advanced-stage breast cancer that has metastasized to other organs is detrimental to the patient and is known to be incurable (Mapes, 2014). The invasive subtype of breast cancer is known as ductal carcinoma. Mutations in tumor suppressor genes *BRCA1* (Hall et al., 1990) and *BRCA2* (Wooster et al., 1994) that are involved in homologous DNA repair, are the most common causes of familial breast cancer. The *TP53* gene encoding for p53 is known to be suppressed in breast cancers, while *CCND1* encoding for cyclinD1 is activated (Ortiz et al., 2017). The HER2 pathway is activated in ~14% of breast cancers caused by the amplification of the *ERBB2* gene (Harbeck et al., 2019). Ductal carcinoma is represented by the breast invasive adenocarcinoma (BRCA) dataset from TCGA in this study. The importance of

multidisciplinary research in describing the molecular mechanism of breast cancer and advancing targeted therapy approaches against the disease has been emphasized in a study by (Harbeck, 2020).

1.4.2 CHAPTER 4

This chapter focuses on use of multivariable regression with gene sets that contain coexpressed, coregulated, or coacting members. It differs from the previous chapters in that it uses multivariable Cox regression to reveal novel, cancer-specific and independent survival signatures using transcriptomics. Although the online tools SurvExpress (Aguirre-Gamboa et al., 2013) and SurvMicro (Aguirre-Gamboa & Trevino, 2014) can perform multivariable Cox regression, Chapter 4 performs this across TCGA-PANCAN datasets with two novel Shiny applications (apps). The first app, SmulTCan, can be used for building and analyzing CPH models with individual cancer datasets, while the second app ClusterHR allows for visualization of CPH models over multiple cancer datasets. Both apps accept the same type of file downloaded from University of California Santa Cruz (UCSC) Xena. Three different sets of axon guidance ligand-receptors, Slit-Robo, netrins-receptors, and Semas-receptors have been used to demonstrate the apps; from which resulting interesting findings have been presented.

2 UNIVARIABLE SURVIVAL ANALYSIS OF FOUR TCGA DATASETS WITH EXON EXPRESSIONS AND ENRICHMENT ANALYSIS

2.1 INTRODUCTION

2.1.1 RNA-SEQUENCING

RNA-Sequencing (RNA-Seq) is a method used to identify and quantify RNA using next generating sequencing (NGS). RNA-Seq has been replacing microarrays in recent years as the main method of analyzing many genes in an organism under different biological conditions, due to it being a more affordable and straightforward method (Evans et al., 2018). RNA-Seq allows for investigation and discovery of the transcriptome, which is key in connecting genomic information with its functional protein expression. RNA-Seq has increasingly become the method of choice in studies involving the detection of alternative splicing isoforms, transcript fusions and strand-specific expression, and the creation of transcriptome assemblies (Dillies et al., 2013).

Most of the publicly available mRNA-Sequencing data is comprised of short-read sequenced cDNA by Illumina technology (Lachmann et al., 2018). Short-read cDNA sequencing, which could either be ambiguous or unambiguous to exons as well as isoforms, has become the foremost RNA-Seq method to detect and quantify transcriptome-wide gene expression (Stark et al., 2019). The Illumina HiSeq 2000 sequencing platform of the short-read technology was used to create the RNA-Seq exon and isoform expression profiles analyzed in this study. Short-read cDNA sequencing is a robust method which results in cDNA fragments that are shorter than 200 base pairs (bp). The Illumina workflow employs sequencing by synthesis using 3'-blocked fluorescently labeled nucleotides on clustered cDNA molecules on a flowcell.

The main stages of RNA-Seq analysis can be given as: alignment and quantification; followed by the filtering and normalization of raw RNA-Seq reads (Stark et al., 2019). The most relevant and preferred methods of raw RNA-Seq counts pre-processing include the limma-trend when sequencing depth is reasonably consistent across samples and the voom approach for variable library sizes across samples (Law et al., 2014; Liu et al., 2015). In both methods, RNA-Seq read counts must be converted to $\log_2$-counts-per-million ($\log_2 CPM$).

Scale normalization is applied to RNA-Seq read counts usually after a filtration step to remove rows that have consistently zero or very low counts across samples (Robinson & Oshlack, 2010). Filtering is mainly used to remove uniformly low-expressed genes as in the case of all expression data analyses (Sha et al., 2015). For the normalization step, several methods exist, such as reads per kilobase of transcript, per million mapped reads ($RPKM$), the trimmed mean of M-values (TMM), total count (TC), upper quartile (UQ), quantile (Q) and normalization using the DESeq method (Dillies et al., 2013).

The $RPKM$ method (Mortazavi et al., 2008), which attempts to scale by transcript length, is one of the most used normalization methods, though it is often criticized for introducing per-gene bias in lowly-expressed genes (Dillies et al., 2013). In TC

normalization, gene counts are divided by their library size and multiplied by the mean total count of all libraries. UQ normalization is like TC in principle, except the mean total counts in the upper quartile are used in multiplication (Bullard et al., 2010). The counts of a gene for a single sample divided by the geometric mean of its counts across the library is used to compute a median value for the gene, i.e., the scaling factor of the gene in the DESeq method, implemented in the Bioconductor package "DESeq2" (Love et al., 2014).

Another commonly used and robust method for scale normalization is TMM (Robinson & Oshlack, 2010); whereby a reference sample is chosen for each gene and remaining samples are considered test samples. The TMM for a gene is the weighted mean of the log ratios between each test sample and the reference, after the exclusion of the most expressed genes and genes with the largest log ratios, with the assumption that most genes are not differentially expressed (DE). TMM was chosen as the normalization method for exon and transcript expressions in Chapters 2 and 3, with the "edgeR" package of Bioconductor (Robinson et al., 2010) as recommended by the package manual.

2.1.2 TCGA DATASETS USED IN THE EXON-BASED ANALYSES

Datasets of four different cancer types were selected from TCGA for analysis: STAD for stomach adenocarcinoma, LIHC for liver hepatocellular carcinoma, BRCA for breast invasive carcinoma, and GBM for glioblastoma multiforme. Sample sizes of the selected datasets for the exon-based analyses are given in Table 1 below.

Table 1: Sample sizes of TCGA datasets used in the exon-based analyses, consisting of primary tumors for each dataset.

STAD	GBM	BRCA	LIHC
386	155	1073	365

2.1.3 LONG NONCODING RNAS AND CANCER

Most of the mammalian genome produces long noncoding RNAs (long ncRNAs or lncRNAs), while only a small portion is transcribed into mRNAs that encode for proteins, creating a ratio of ≃ 3∶1 (Zheng et al., 2019). Long ncRNA molecules are longer than 200 nucleotides and consist of the three main types: mRNA-like intergenic transcripts (lincRNAs), antisense transcripts of protein-coding genes and unconventional lncRNAs derived from RNA polymerase II (Pol II) primary transcripts that are subject to different types of post-transcriptional modifications required for stabilization (Yao et al., 2019).

Long ncRNAs have been identified to be involved in various cellular functions, particularly inside the nucleus, including regulating chromosome architecture, modulating inter- and intrachromosomal interactions and regulating the recruitment of chromatin modifiers (Statello et al., 2021). They are also involved in transcriptional regulation by forming R-loops, interfering with Pol II machineries, and through their DNA loci locally.

Long ncRNAs are also involved in regulating nuclear bodies; while in the cytoplasm they regulate mRNA turnover, translation, and post-translational modifications (Yao et al., 2019).

Given the importance of lncRNAs in various cellular functions, the involvement of their aberrant expression and dysfunction in disorders is inevitable. Accordingly, the lncRNA activated by *TGF-β* (*LNCRNA-ATB*) promotes the epithelial-mesenchymal transition (EMT) in liver cancer by upregulating levels of *ZEB1* and *ZEB2* (Yuan et al., 2014). Stemness, an important property of metastasis-initiating tumor cells, was found to be inhibited by the lncRNA molecule *TSLNC8* through the STAT3 signaling pathway (Jiang et al., 2019). On the other hand, *HOXC-AS3* was found to promote tumorigenesis in gastric cancer (Zhang et al., 2018). Additionally, the *H19* molecule is upregulated in various cancers including colorectal, breast and ovarian, and promotes oncogenesis and drug resistance during therapy (Liu et al., 2020). Identification and functional annotation of lncRNAs constitutes a current important biological problem (Gudenas & Wang, 2018).

The type of lncRNA investigated in this study was the intervening/intergenic type, i.e., lincRNA, that constitutes more than half of lncRNA transcripts in humans. Despite resembling mRNAs, lincRNAs contain fewer exons and are more prone to degradation due to inefficient splicing and polyadenylation mechanisms (Ransohoff et al., 2018). LincRNAs are abundant in the nucleus, exhibit tissue specificity and are involved in developmental patterning (He et al., 2014; H. Wang et al., 2017). They stabilize chromatin topology, modulate protein scaffolding, act as protein and RNA decoys, regulate neighboring transcription, and in certain cases encode functional micropeptides (Flower et al., 2020). For example, the *MALAT1* lincRNA is upregulated in non-small cell lung and colorectal cancers, while *LINK-A* promotes resistance to AKT inhibitors in breast cancer (Ransohoff et al., 2018). Another lincRNA molecule HOX antisense intergenic RNA (*HOTAIR*) promotes cancer metastasis and is upregulated in epithelial cancers, such as breast, gastric, squamous cell carcinomas and GBM (Liu et al., 2020).

2.1.4 ALTERNATIVE EXON USAGE IN CANCER

In a study involving samples from 113 neuroblastoma patients, *MYCN* and *NRTK1* were identified as the genes with the highest alternative exon usage (AEU), and authors of this study concluded that signatures formed with exon expressions had better predictive power over gene expression signatures on survival (Schramm et al., 2012).

Another study of AEU with 250 glioblastoma patients identified single exon genes of the Snord family, which are small nucleolar RNAs, to be associated with survival, with *SNORD28* and *SNORD104* negatively and *SNORD123* positively associated (Sadeque et al., 2012). In another study by Tian et al. (2020) with TCGA's head and neck squamous cell carcinoma (HNSC) dataset and semaphorins (Semas), expression of *SEMA3A*'s ninth exon was associated with poor disease-free survival. While previous studies exist with associating individual exons with survival in different types of cancer, there lacks a transcriptome-wide study involving several cancer datasets in which exons are ranked based on an HR value of their expressions.

2.1.5 AIMS

- Identify individual exons with transcriptome-wide significance and impact with respect to the four TCGA datasets: GBM, STAD, LIHC, and BRCA.
- Use aggregation methods to form gene HR sets for each cancer dataset from their identified significant exons.
- Apply the methods above to the lincRNA transcriptome for the selected TCGA datasets.

2.2 METHODS

2.2.1 DATA ACQUISITION

For each of the four cancer datasets, the GDCdownload() function of the "TCGAbiolinks" Bioconductor package (Colaprico et al., 2016) was used to download the folders containing the TXT files with raw exon quantification RNA-seq counts from the GDC legacy archive. The GDCquery() function of the same package was then used to access cancerous tissue sample names and information for each of the downloaded TXT files. This information was then used to construct a raw counts matrix for each of the four cancer datasets from primary tumor samples sequenced with the Illumina HiSeq 2000 v2 platform submitted from the University of North Carolina (UNC). The "UCSCXenaTools" package (Wang, 2019) was used to download the survival TXT files for each of the four TCGA datasets from GDC after which sample names in the raw counts matrix were matched and confirmed with those in the survival data file.

2.2.2 PRE-PROCESSING

Pre-processing of the raw counts matrices for all four TCGA datasets was carried out using default parameters of the calcNormFactors() function of the package "limma" (Ritchie et al., 2015) and cpm() function of "edgeR" (Robinson et al., 2010) to obtain their TMM-normalized $\log_2(CPM + \frac{2}{L})$ expression values, where CPM is mapped read counts-per-million and L denotes the library size of each exon in the panel. The $\log_2(CPM + \frac{2}{L})$ expression values will be referred to as logCPM values from now on in the text for convenience.

2.2.3 DATA ANALYSIS AND INTERPRETATION

The coding and noncoding transcriptomes were accessed through UCSC's "knownGene" (2015a) and "lincRNATranscripts" (2015b) tables for genome version hg19. For the "knownGene" library, transcripts of the associated Entrez GENEID, as well as the exons associated with each transcript were searched for in the coordinates of the exon expression panel. For the "lincRNATranscripts" library, exons of each lincRNA transcript in the library were searched for in the expression panel. Exons for which there exists more than one associated probe in the platform, the mean read value for the probes

was calculated. The read values of exons for which there are no probes on the platform were designated as "not in panel" and regression was not performed on these regions.

The coxph() function of the "survival" package of R (Therneau & Grambsch, 2000) with the $\log CPM$ expression of RNA extracted from cancerous tissue samples was used to calculate a hazard ratio (HR) and an associated p-value for each exon in the hg19 coding and noncoding transcriptomes for all four TCGA datasets in the study. Specifically, for TCGA-LIHC, the exons with IDs: 270473, 258314, 68613, 174774, 27437, 221500, 115393, 31246, 89278, 12353, 12352, 93301, 111485, 155147, 252188, 101018, 89474, 117759, 33416, 27138, 27132, 215074, 42957, 175895, 175901, 163327, and 163319 were excluded from the exon-based analysis. This was due to highly repetitive expression values of these exons and the resulting coxph() singularity error.

The coding transcriptome results tables were annotated by the addition of "GENENAME" and "SYMBOL" columns using the "org.Hs.eg.db" annotation package version 3.11.4 (2020). Individual exons with transcriptome-wide impact were identified by filtering the annotated files for exons with $p < 0.05$ and those with a $\log CPM > 1$ count in ≃70% of the total number of samples for that dataset. Thresholds for the $\log CPM > 1$ counts of individual exons in the coding transcriptome were set at 270, 108, 255, and 750: for datasets STAD, GBM, LIHC, and BRCA, respectively. For the noncoding transcriptome, exons that have a ≥ 100 $\log CPM > 1$ count across samples with $p < 0.05$ were kept for TCGA-STAD and TCGA-LIHC, whereas the threshold for total $\log CPM > 1$ values was 50 for TCGA-GBM and 330 for TCGA-BRCA. The total counts filter was loosened to ≃23-40% of the total number of samples in the datasets for the noncoding transcriptome since its overall expression levels are lower than the coding transcriptome.

The filtered exon tables were then aggregated based on the "GENEID" and "tx_id" columns for the "knownGene" and "lincRNATranscripts" libraries, respectively, to generate the gene and lincRNA lists for each dataset. A mean HR was calculated for each gene and lincRNA from its filtered exons, and a new p-value and adjusted p-value was calculated for each using their significant exons' p-values with Fisher's method (Cinar & Viechtbauer, 2021).

2.2.4 ENRICHMENT ANALYSIS

The "clusterProfiler" (Wu et al., 2021) and "enrichplot" (Yu, 2021) packages of Bioconductor were used in the enrichment analyses for the gene lists of each dataset that came through the aggregation step. Input gene lists of each dataset to the gseGO() function were ranked based on the mean HR values of their filtered exons calculated in the previous step, while the "org.Hs.eg.db" annotation package version 3.11.4 was also used inside the function.

Two sets of genes were formed for the four TCGA datasets in this study, from genes that were found to be either positively ($HR < 1$) or negatively ($HR > 1$) associated with cancer survival rates by their mean HR values. Genes that contained both positively and negatively associated exons were excluded from the HR direction-based analyses. Genes deemed neither hazardous nor beneficial and were excluded were: *JUP* and *MED26* from

BRCA, *HNRNPL* from STAD and *MTFP1* from LIHC. For the KEGG pathway enrichment, "enrichKEGG" was used for the "fun" parameter of the compareCluster() function of "clusterProfiler", which in turn was used inside the pairwise_termsim() function of "enrichplot". The hazardous and protective gene sets of each dataset were enriched separately for the KEGG pathways.

2.3 Results

2.3.1 Stomach adenocarcinoma

The effect of TMM normalization on logCPM expression values for several samples from the stomach adenocarcinoma dataset TCGA-STAD is shown in Figure 4. The normalization step was performed prior to downstream survival calculations, which resulted in exon expressions to be more uniformly distributed across samples.

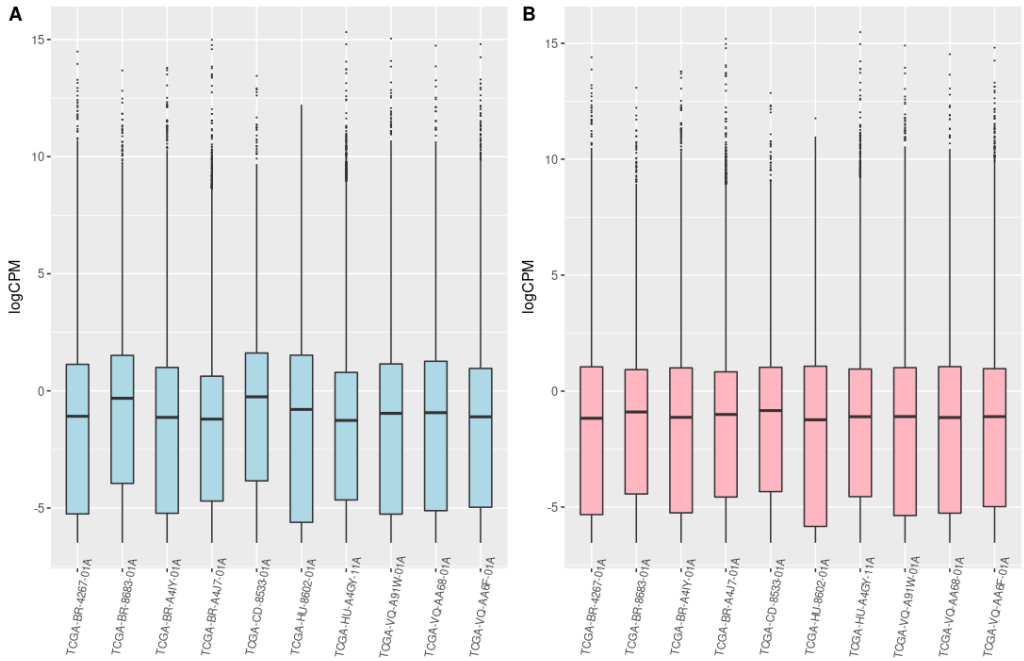

Figure 4: Effect of TMM normalization on STAD samples. (A) Distribution of logCPM expressions of raw or unnormalized samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Individual exons of transcriptome-wide significance and impact that have passed the log$CPM > 1$ count filter for the STAD dataset are listed in Table 2 for those exons with the ten highest and ten lowest HRs. In total, 16,579 exons that have passed the log$CPM >$ 1 count filter were identified with transcriptome-wide significance ($p < 0.05$) in STAD. The exon with the highest HR, 207586, belongs to the *FTO* gene and Exon 237010, with the lowest HR, belongs to *ILF3*.

Table 2: Exons with the highest and lowest HR values in the coding hg19 transcriptome for TCGA-STAD with individual transcriptome-wide significance.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
207586	chr16	54145674	54148379	2706	+	1.7515963	0.0003058
214634	chr16	74490547	74490653	107	-	1.6755667	0.0055029
4518	chr1	52250114	52250254	141	+	1.6598559	0.0028690
183381	chr14	102514154	102514365	212	+	1.6207058	0.0063166
183380	chr14	102510932	102511035	104	+	1.6153901	0.0092129
213019	chr16	46696895	46697074	180	-	1.5987201	0.0098095
106701	chr7	2767741	2771384	3644	-	1.5897786	0.0137978
170198	chr12	53905843	53911138	5296	-	1.5839411	0.0168278
170199	chr12	53910799	53911138	340	-	1.5839411	0.0168278
183378	chr14	102510212	102510382	171	+	1.5751663	0.0144610
244156	chr19	5611131	5611641	511	-	0.5417783	0.0095328
155633	chr11	62365496	62365613	118	-	0.5406838	0.0004894
148652	chr11	65825785	65825876	92	+	0.5327616	0.0054166
148651	chr11	65825525	65825621	97	+	0.5325170	0.0013859
127631	chr9	140705913	140706067	155	+	0.5318675	0.0004293
127612	chr9	140622801	140622981	181	+	0.5318592	0.0007495
236248	chr19	5649000	5649510	511	+	0.5267369	0.0074154
148653	chr11	65826301	65826516	216	+	0.5218674	0.0033903
237011	chr19	10791993	10792005	13	+	0.5150928	0.0005898
237010	chr19	10791895	10792005	111	+	0.5150928	0.0005898

Exons with transcriptome-wide significance and impact for TCGA-STAD identified in the noncoding lincRNA transcriptome are indicated in Table 3. The exons identified with the highest HRs ($\simeq 1.46$, $p < 0.05$) were within the region chr7:66018453-66019824 on the negative strand. On the other hand, the bottom five exons in the list were within the chr12:112279131-112279765 region's negative strand, which maps to the *MAPKAPK5-AS1* RNA gene. A total of 101 exons were identified with sufficient $\log CPM > 1$ counts from the noncoding transcriptome in STAD.

Table 3: Exons with the five highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-STAD.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
24725	chr7	66018553	66019824	1272	-	1.4567186	0.0275127
24726	chr7	66018620	66019824	1205	-	1.4567186	0.0275127
24724	chr7	66018453	66019824	1372	-	1.4567186	0.0275127
24727	chr7	66019572	66019824	253	-	1.4567186	0.0275127
30770	chr10	99476951	99477903	953	+	1.3291021	0.0007100
35588	chr12	112279131	112279274	144	-	0.6393188	0.0212597
35593	chr12	112279488	112279749	262	-	0.6393188	0.0212597
35591	chr12	112279173	112279731	559	-	0.6393188	0.0212597
35589	chr12	112279131	112279765	635	-	0.6393188	0.0212597
35592	chr12	112279488	112279710	223	-	0.6393188	0.0212597

Genes and lincRNAs identified from aggregation of filtered exons, with the highest and lowest mean HR values are given in Table 4. The number of exons that passed the $\log CPM > 1$ count filter with $p < 0.05$ used in the adjusted p-value calculations for each listed gene is indicated in the "# of sig. exons" column. On the top of the list is *GNA12* with one exon of transcriptome-wide significance; on the bottom of the list is *YTHDC1*, also with one significant exon.

Table 4: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation with significant exons for TCGA-STAD.

Gene symbol	Mean HR of sig. exons	# of sig. exons	Adj. p-value
GNA12	1.5897786	1	0.0137978
ATF7	1.5839411	2	0.0025966
SOS1	1.5570613	1	0.0188534
MFAP3	1.5506783	1	0.0150916
ABL2	1.5465195	3	0.0000884
SRGAP2	1.5406621	2	0.0008190
VPS35	1.5293473	4	0.0000737
GLG1	1.5284859	4	0.0000981
DYNC1H1	1.5156677	9	0.0000000
FTO	1.5141765	2	0.0001506
TCONS_l2_00027403	1.4567186	1	0.0275127
TCONS_l2_00027401	1.4567186	1	0.0275127
TCONS_l2_00027400	1.4567186	1	0.0275127
TCONS_l2_00027398	1.4567186	1	0.0275127
TCONS_l2_00027397	1.4567186	1	0.0275127
TCONS_00020926	0.6393188	1	0.0212597
TCONS_00020925	0.6393188	1	0.0212597
TCONS_00020924	0.6393188	1	0.0212597
TCONS_00020221	0.6393188	1	0.0212597
TCONS_00020220	0.6393188	1	0.0212597
AP4B1-AS1	0.5977880	1	0.0004046
ZNF142	0.5890059	1	0.0111163
AKAP8	0.5881798	1	0.0006403
SMCR8	0.5809372	1	0.0058611
SP2	0.5777289	1	0.0113160
THRAP3	0.5742799	3	0.0005781
MIR3618	0.5578095	1	0.0027721
MIR1306	0.5578095	1	0.0027721
DGCR8	0.5578095	1	0.0027721
YTHDC1	0.5549425	1	0.0034225

2.3.2 GLIOBLASTOMA MULTIFORME

The effect of TMM normalization for ten randomly selected samples of the glioblastoma multiforme dataset TCGA-GBM are given in Figure 5. Medians of the exon expression samples were more uniformly distributed after the normalization step as expected.

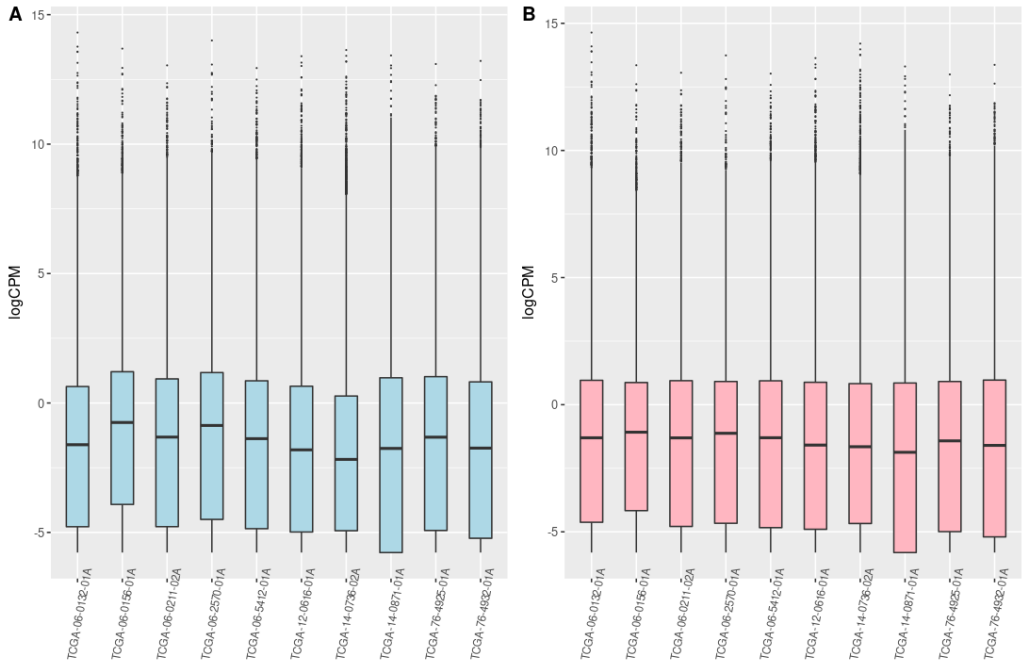

Figure 5: Effect of TMM normalization on GBM samples. (A) Distribution of logCPM expressions of raw or unnormalized samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Specific exons of transcriptome-wide significance and impact that have passed the log$CPM > 1$ count filter for the GBM dataset are listed in Table 5 for those with the ten highest and ten lowest HRs. The top six exons belong to the *CSNK1D* gene. The exon that was identified as the most protective with ID 141821 (HR $\simeq 0.36$, $p < 0.001$) belongs to *AP3M1*. In total, 13,714 exons have passed the counts filter for GBM with transcriptome-wide significance ($p < 0.05$).

Table 5: Exons with the highest and lowest HR values in the coding hg19 transcriptome for TCGA-GBM, which have passed the log$CPM > 1$ count filter of 108 for this dataset.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
231067	chr17	80210310	80210480	171	-	3.1376097	0.0025692
231066	chr17	80209255	80209403	149	-	2.8859359	0.0049942
231062	chr17	80206751	80206890	140	-	2.8611829	0.0016247
231068	chr17	80210892	80211120	229	-	2.6688082	0.0064105
231069	chr17	80210892	80212834	1943	-	2.6688082	0.0064105
231070	chr17	80213305	80213453	149	-	2.4017941	0.0135448
38132	chr2	25043593	25043717	125	-	2.3636808	0.0005888
112127	chr7	140155983	140156666	684	-	2.2767158	0.0012633
112128	chr7	140156452	140156666	215	-	2.2767158	0.0012633
216725	chr17	8363327	8363478	152	+	2.2523502	0.0027079
129858	chr9	86590377	86590420	44	-	0.4503470	0.0084626
129857	chr9	86589432	86590713	1282	-	0.4503470	0.0084626
45583	chr2	198267280	198267550	271	-	0.4246570	0.0067110
94535	chr6	33235642	33235774	133	-	0.4205400	0.0019951

94534	chr6	33235459	33235516	58	-	0.4108881	0.0007090
45582	chr2	198266709	198266854	146	-	0.3883159	0.0053012
54760	chr3	183855408	183855863	456	+	0.3770775	0.0018374
54761	chr3	183855686	183855863	178	+	0.3770775	0.0018374
54759	chr3	183855408	183855593	186	+	0.3770775	0.0018374
141821	chr10	75885906	75886113	208	-	0.3582646	0.0004534

The noncoding lincRNA transcriptome was scanned univariably in TCGA-GBM and individual exons with transcriptome-wide significance and impact were identified, as indicated in Table 6. The top five most hazardous exons in this list were from the positive strand of region chr7:72299944-72300362, which contains the upstream region of the *SBDSP1* gene. This is a long noncoding RNA (lncRNA) associated with colorectal cancer (Shi et al., 2017). The bottom three exons with the lowest HRs, exons 36293-5, map to the 5'UTR region of 40S ribosomal protein S4X. The number exons that passed the $\log CPM > 1$ count filter for GBM was 184.

Table 6: Exons with five of the highest and lowest HR values in the noncoding hg19 transcriptome in TCGA-GBM. The listed exons have been obtained after filtering out those with insufficient $\log CPM > 1$ counts in the GBM dataset.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
23149	chr7	72299952	72300362	411	+	1.4860461	0.0158498
23152	chr7	72300538	72300555	18	+	1.4860461	0.0158498
23151	chr7	72299977	72300362	386	+	1.4860461	0.0158498
23150	chr7	72299972	72300362	391	+	1.4860461	0.0158498
23148	chr7	72299944	72300362	419	+	1.4860461	0.0158498
46342	chr19	28283908	28284285	378	-	0.7380067	0.0290904
46337	chr19	28283908	28284140	233	-	0.7380067	0.0290904
36295	chr13	52034808	52035522	715	+	0.7116053	0.0228919
36294	chr13	52034637	52034692	56	+	0.7116053	0.0228919
36293	chr13	52034621	52036829	2209	+	0.7116053	0.0228919

Genes and lincRNAs identified from aggregation of filtered exons, with the highest and lowest mean HR values are given in Table 7. The number exons with univariable HR values with $p < 0.05$ used in the adjusted p-value computations are indicated in the "# of sig. exons" column. According to the table, the *CSNK1D* gene was found to have the highest mean HR with the contribution of twelve exons: each with HR values > 1 with univariable transcriptome-wide significance. *AP3M1* was found to have the lowest mean HR with one contributing exon.

Table 7: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-GBM.

Gene symbol	Mean HR of sig. exons	# of sig. exons	Adj. *p*-value
CSNK1D	2.3540254	12	0.0000000
ARSB	2.0489307	1	0.0079073
MKRN1	2.0415143	7	0.0000000
NDEL1	1.9840196	4	0.0000017
YWHAG	1.9814540	1	0.0006731
CLPTM1L	1.9484724	18	0.0000000
CENPO	1.9397006	1	0.0022207
ADCY3	1.9104136	8	0.0000000
TMEM127	1.8771575	1	0.0320309
SRP68	1.8716046	1	0.0466669
TCONS_00002153	1.4382761	1	0.0133434
TCONS_00000054	1.4382761	1	0.0133434
TCONS_l2_00027396	1.4146060	2	0.0005258
TCONS_l2_00027395	1.4146060	1	0.0069316
TCONS_l2_00026658	1.4146060	2	0.0005258
TCONS_00027260	0.7380067	1	0.0290904
TCONS_00027256	0.7380067	1	0.0290904
TCONS_00026762	0.7380067	1	0.0290904
TCONS_l2_00006855	0.7116053	2	0.0044826
TCONS_l2_00006854	0.7116053	1	0.0228919
UBE2G2	0.5238908	1	0.0228263
EPC2	0.5195640	3	0.0003885
PNPLA3	0.5180252	2	0.0000076
PPP4R3B	0.5083592	1	0.0124231
VPS52	0.5071418	9	0.0000000
ABCB7	0.5066111	1	0.0028273
KARS1	0.5035317	7	0.0000002
PHF6	0.4802143	3	0.0000048
EIF2B5	0.4669790	5	0.0000003
AP3M1	0.3582646	1	0.0004534

2.3.3 LIVER HEPATOCELLULAR CARCINOMA

Expressions of ten randomly selected samples of the liver hepatocellular carcinoma dataset TCGA-LIHC are shown in Figure 6A and B, before and after TMM normalization, respectively. As expected, medians of the exon expressions were more uniformly distributed after normalizing.

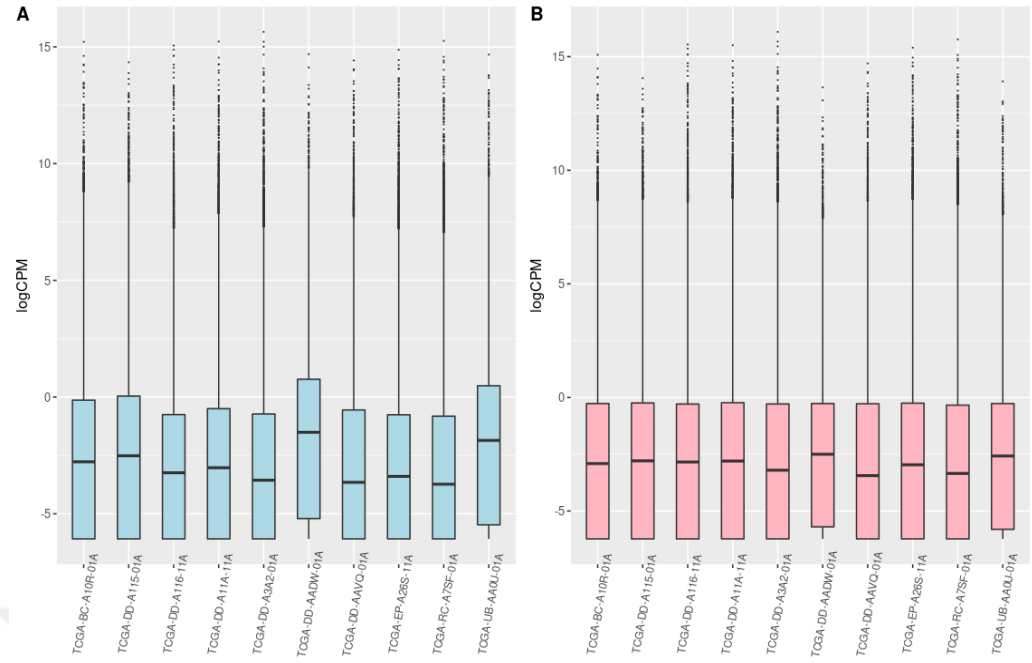

Figure 6: Effect of TMM normalization on LIHC samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Individual exons of transcriptome-wide significance and impact that have passed the log$CPM > 1$ count filter for the LIHC dataset are listed in Table 8 for those with the ten highest and ten lowest HRs. Altogether, 13,134 exons with transcriptome-wide significance ($p < 0.05$) have been identified for the LIHC dataset with log$CPM > 1$ counts > 255. The exon with the highest HR value in LIHC, 16316, belongs to the *SRSF10* gene, a member of the serine-arginine (SR) family involved in RNA splicing. Exon 197300 with the lowest HR value belongs to *AKAP13*.

Table 8: Exons with the highest and lowest HR values in the coding hg19 transcriptome in TCGA-LIHC. Exons with the ten highest and ten lowest HRs with transcriptome-wide significance that have passed the logCPM > 1 count filter were listed.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
16316	chr1	24298340	24298502	163	-	1.9040971	0.0053322
139948	chr10	27404989	27405062	74	-	1.8607774	0.0024414
139947	chr10	27403450	27403536	87	-	1.8601184	0.0013594
38486	chr2	27607540	27607955	416	-	1.8448756	0.0001117
38487	chr2	27608614	27608746	133	-	1.8448756	0.0001117
38488	chr2	27609061	27609146	86	-	1.8448756	0.0001117
38489	chr2	27609956	27610025	70	-	1.8448756	0.0001117
231062	chr17	80206751	80206890	140	-	1.7961281	0.0001026
139949	chr10	27405148	27405274	127	-	1.7922261	0.0036138
51782	chr3	101309059	101309114	56	+	1.7713003	0.0173665
49615	chr3	43388831	43392634	3804	+	0.6277882	0.0034649
193745	chr15	31233769	31235310	1542	+	0.6228399	0.0013793

200940	chr15	55651737	55653142	1406	-	0.6000988	0.0001076
200936	chr15	55647421	55648607	1187	-	0.6000988	0.0001076
200939	chr15	55651737	55651888	152	-	0.6000988	0.0001076
200941	chr15	55653038	55653142	105	-	0.6000988	0.0001076
200937	chr15	55647421	55653142	5722	-	0.6000988	0.0001076
200936	chr15	55647421	55648607	1187	-	0.6000988	0.0001076
200940	chr15	55651737	55653142	1406	-	0.6000988	0.0001076
197300	chr15	86287859	86292589	4731	+	0.5703750	0.0001293

Exons with transcriptome-wide significance and impact for TCGA-LIHC, that were identified in the hg19 noncoding lincRNA transcriptome are indicated in Table 9 for five exons with the highest and lowest HRs. In total, 152 exons with transcriptome-wide significance with sufficient $\log CPM > 1$ counts have been identified for LIHC. Interestingly, exons with protective HRs in TCGA-STAD for the noncoding transcriptome on the region on chromosome 12 mapping to *MAPKAPK5-AS*, were found to be associated with hazardous HRs in TCGA-LIHC. The most protective exons in LIHC mapped to a region on chromosome 15 containing *OIP5-AS1*, an important regulator of proliferation in embryonic stem cells and BC067230.

Table 9: Exons with the highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-LIHC.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
35594	chr12	112279488	112279779	292	-	1.3553973	0.0124276
35595	chr12	112279509	112279752	244	-	1.3553973	0.0124276
35590	chr12	112279131	112279771	641	-	1.3553973	0.0124276
35588	chr12	112279131	112279274	144	-	1.3553973	0.0124276
35593	chr12	112279488	112279749	262	-	1.3553973	0.0124276
24652	chr7	65844814	65844990	177	-	0.7972080	0.0211625
24651	chr7	65841031	65844990	3960	-	0.7972080	0.0211625
24650	chr7	65837573	65844990	7418	-	0.7972080	0.0211625
39569	chr15	41590239	41598727	8489	+	0.6719687	0.0141063
39570	chr15	41591874	41598734	6861	+	0.6455960	0.0010542

Genes and lincRNAs identified from aggregation of filtered exons, with the highest and lowest mean HR values are given in Table 10. The number of exons used in the adjusted p-value calculations is indicated in the "# of sig. exons" column. *PAFAH1B2* contributed one exon with the highest HR value ($p < 0.01$), while the *AKAP13* gene also contributed one exon with the lowest HR ($p < 0.001$).

Table 10: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-LIHC.

Gene symbol	Mean HR of sig. exons	# of sig. exons	Adj. *p*-value
PAFAH1B2	1.7652744	1	0.0086555
SRSF10	1.7125499	2	0.0009418
C1orf109	1.7072782	1	0.0009557
SUPT7L	1.7062648	1	0.0039786
FKBP1A	1.7054627	2	0.0000600
NRAS	1.6888862	1	0.0002454
BMI1	1.6875856	1	0.0015323
AKIRIN1	1.6812822	1	0.0004406
ARID1A	1.6772556	1	0.0055023
PCNP	1.6601721	3	0.0006238
TCONS_00020932	1.3553973	1	0.0124276
TCONS_00020931	1.3553973	1	0.0124276
TCONS_00020930	1.3553973	1	0.0124276
TCONS_00020928	1.3553973	1	0.0124276
TCONS_00020927	1.3553973	2	0.0015098
TCONS_l2_00027387	0.7972080	1	0.0211625
TCONS_l2_00026646	0.7972080	1	0.0211625
TCONS_00023384	0.6719687	1	0.0141063
TCONS_00023383	0.6719687	1	0.0141063
TCONS_00023382	0.6455960	1	0.0010542
ZNF181	0.6626963	1	0.0041562
STX17	0.6584223	1	0.0145307
UBR1	0.6542178	1	0.0141417
PCGF5	0.6505692	1	0.0027837
RTF1	0.6470719	1	0.0167282
SNRK	0.6277882	1	0.0034649
FAN1	0.6228399	1	0.0013793
DNAAF4-CCPG1	0.6000988	2	0.0000002
CCPG1	0.6000988	5	0.0000000
AKAP13	0.5703750	1	0.0001293

2.3.4 BREAST INVASIVE CARCINOMA

Expression distributions of ten randomly selected primary tumor samples of the breast invasive carcinoma dataset TCGA-BRCA are shown in Figure 7A and B, before and after *TMM* normalization, respectively. Medians of the exon expression distributions were more uniformly distributed after normalization, as expected.

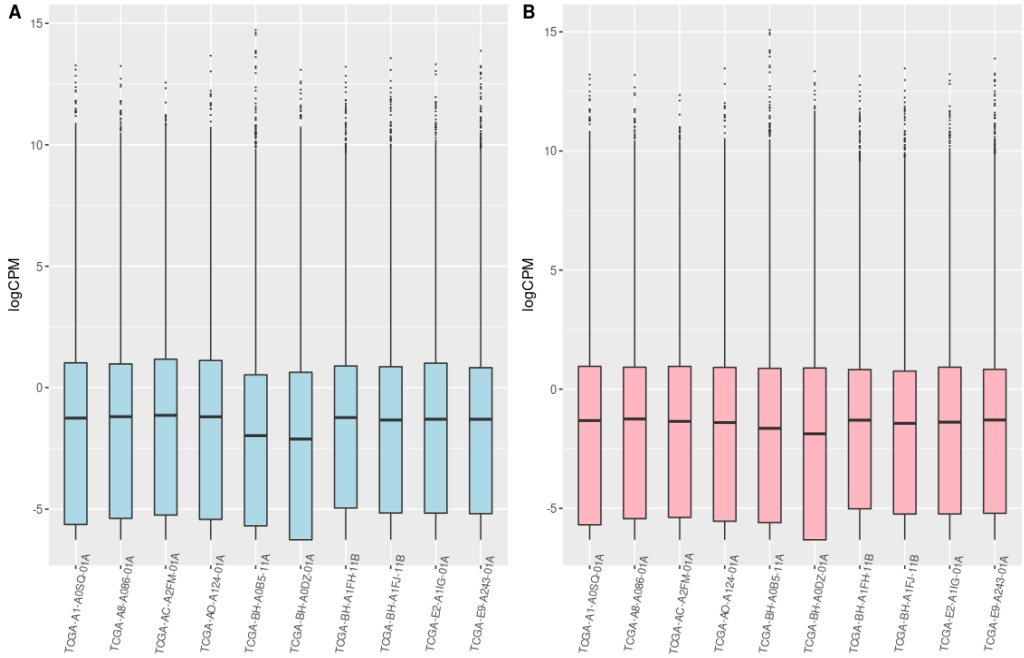

Figure 7: Effect of TMM normalization on BRCA samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Specific exons of transcriptome-wide significance and impact that have passed the log$CPM > 1$ count filter for the BRCA dataset are listed in Table 11 for those with the ten highest and ten lowest HRs. Overall, 16,226 exons with sufficient log$CPM > 1$ counts and transcriptome-wide significance ($p < 0.05$) were identified in BRCA. Exon 152872 with the highest HR belongs to *CARS1*, while Exon 19836 with the lowest HR belongs to *RBMXL1*.

Table 11: Exons with the highest and lowest HR values in the coding hg19 transcriptome in TCGA-BRCA. Ten highest and ten lowest HRs with transcriptome-wide significance were indicated.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
152872	chr11	3050533	3050673	141	-	1.8087083	0.0004268
83964	chr5	141014378	141014520	143	-	1.7925599	0.0062596
83965	chr5	141014378	141014729	352	-	1.7925599	0.0062596
170750	chr12	57058560	57058584	25	-	1.7844373	0.0002892
75620	chr5	68414326	68414455	130	+	1.7805727	0.0006538
69856	chr4	24543496	24543645	150	-	1.7798749	0.0007970
75095	chr5	52095719	52096954	1236	+	1.7744619	0.0005804
87647	chr6	32137070	32137278	209	+	1.7643648	0.0006190
152873	chr11	3059280	3059429	150	-	1.7592323	0.0005579
215619	chr17	2575949	2576051	103	+	1.7548734	0.0011136
184149	chr14	21859992	21860105	114	-	0.5704081	0.0019326
186735	chr14	78187054	78187171	118	-	0.5664713	0.0045958
227665	chr17	42285232	42285279	48	-	0.5585182	0.0006912
186736	chr14	78189524	78189620	97	-	0.5549966	0.0012836
182159	chr14	75295916	75296046	131	+	0.5486603	0.0004329

3246	chr1	36766487	36766685	199	+	0.5391220	0.0045708
182532	chr14	89069172	89069389	218	+	0.5303510	0.0005058
182533	chr14	89069172	89069404	233	+	0.5303510	0.0005058
186737	chr14	78197331	78197472	142	-	0.5253923	0.0011007
19836	chr1	89445139	89449749	4611	-	0.4539806	0.0002523

Individual exons with transcriptome-wide significance and impact for TCGA-BRCA, that were identified in the hg19 noncoding lincRNA transcriptome are indicated in Table 12. In total, 109 exons in the noncoding transcriptome were identified for BRCA that passed the $\log CPM > 1$ count threshold for lincRNA exons, with univariable survival p-values < 0.05. The top five exons mapped to BC067244 on chromosome 8, while the region on chromosome 9 containing exons with the lowest HRs mapped to *MGC21881*.

Table 12: Exons with the highest and lowest HR values in the noncoding hg19 transcriptome for TCGA-BRCA.

Exon ID	Chromosome	Start	End	Width	Strand	HR	*P*-value
26832	chr8	92082159	92082367	209	-	1.4208159	0.0017615
26831	chr8	92082159	92082359	201	-	1.4208159	0.0017615
26833	chr8	92082159	92082384	226	-	1.4208159	0.0017615
26829	chr8	92080146	92082050	1905	-	1.3831498	0.0084303
26830	chr8	92081807	92082050	244	-	1.3831498	0.0084303
35030	chr12	8383708	8385203	1496	-	0.7838735	0.0057415
28802	chr9	41953619	41953779	161	-	0.7810066	0.0245105
28803	chr9	41954504	41954648	145	-	0.7810066	0.0245105
28804	chr9	41954504	41954688	185	-	0.7810066	0.0245105
28801	chr9	41952399	41954651	2253	-	0.7810066	0.0245105

Genes and lincRNAs identified from aggregation of filtered exons, with the highest and lowest mean HR values are given in Table 13. The number of exons used in the adjusted p-value calculations is indicated in the "# of sig. exons" column. The *CARS1* gene was the most hazardous of genes found in BRCA with the contribution of three exons. *RBMXL1*, which has a single coding exon, was found to be the most protective for BRCA in the gene list resulting from the exon aggregation analysis.

Table 13: Genes and lincRNAs with the highest and lowest mean HR values identified with exon aggregation for TCGA-BRCA.

Gene symbol	Mean HR of sig. exons	# of sig. exons	Adj. *P*-value
CARS1	1.7433900	3	0.0000002
HDAC3	1.7359662	3	0.0000988
TOR1A	1.6876767	3	0.0000061
PPT2-EGFL8	1.6807746	3	0.0000008
DDX23	1.6630557	3	0.0000234
CHTF8	1.6445052	1	0.0133898

VPS41	1.6418278	2	0.0001107
DLG3	1.6361961	3	0.0000001
TMEM41A	1.6309503	4	0.0000001
DHX16	1.6262715	3	0.0000560
TCONS_l2_00028315	1.4019828	2	0.0001799
TCONS_l2_00028314	1.4019828	2	0.0001799
TCONS_l2_00028313	1.4019828	2	0.0001799
TCONS_l2_00019954	1.3756808	1	0.0308316
TCONS_l2_00022977	1.3697201	2	0.0139294
TCONS_l2_00006097	0.7838735	1	0.0057415
TCONS_l2_00006096	0.7838735	1	0.0057415
TCONS_00016318	0.7810066	2	0.0050568
TCONS_00015598	0.7810066	1	0.0245105
TCONS_00015562	0.7810066	1	0.0245105
SNW1	0.6301761	7	0.0000000
RBM15B	0.6247706	1	0.0065336
FBXO34	0.6239790	1	0.0151046
SUPT16H	0.6232955	2	0.0001214
CHD2	0.6209654	1	0.0064491
YLPM1	0.6120256	4	0.0000038
THRAP3	0.6115103	2	0.0003170
ZSCAN25	0.5932670	1	0.0023482
HNRNPK	0.5929845	2	0.0051627
RBMXL1	0.4539806	1	0.0002523

2.3.5 ANALYSIS OF GENE AND LINCRNA SETS FOR THE FOUR DATASETS

After the exclusion of genes with exons that were found to have opposing effects on survival from each of the STAD, LIHC and BRCA datasets; the remaining numbers of genes, divided into groups of hazardous and beneficial (or protective) based on their mean HR values for each of the four datasets are given in Table 14.

Table 14: Number of genes from the exon-based analysis after the filtering and aggregation steps for each cancer dataset.

Dataset	Hazardous	Protective	Total
STAD	848	791	1639
GBM	976	429	1405
LIHC	864	929	1793
BRCA	673	1179	1852

The number of lincRNAs, divided into groups of hazardous and beneficial based on their mean HR values for each of the four datasets are given in Table 15. The table represents the total number of lincRNAs identified in each dataset from the exon-based aggregation analysis in the noncoding hg19 transcriptome.

Table 15: Number of lincRNAs from the exon-based analysis of the noncoding transcriptome after the filtering and aggregation steps for each cancer dataset.

Dataset	Hazardous	Protective	Total
STAD	32	26	58
GBM	87	33	120
LIHC	61	47	108
BRCA	28	44	72

The numbers of genes with unidirectional exons in terms of significant HR values from the univariable expression-based survival analysis, that are overlapping or distinct between the four datasets are represented as percentages in the Venn diagrams of Figure 8. According to this, the highest percentage of overlap for hazardous genes were for GBM-STAD and LIHC-BRCA with 3%; while for the protective gene sets it was for LIHC-BRCA and BRCA-STAD with also 3%. For the lincRNA transcriptome, highest percentage of overlap was found for hazardous lincRNAs for GBM-STAD and GBM-BRCA with 6%; while the highest percentage of overlap was 2% for GBM-BRCA among protective lincRNAs.

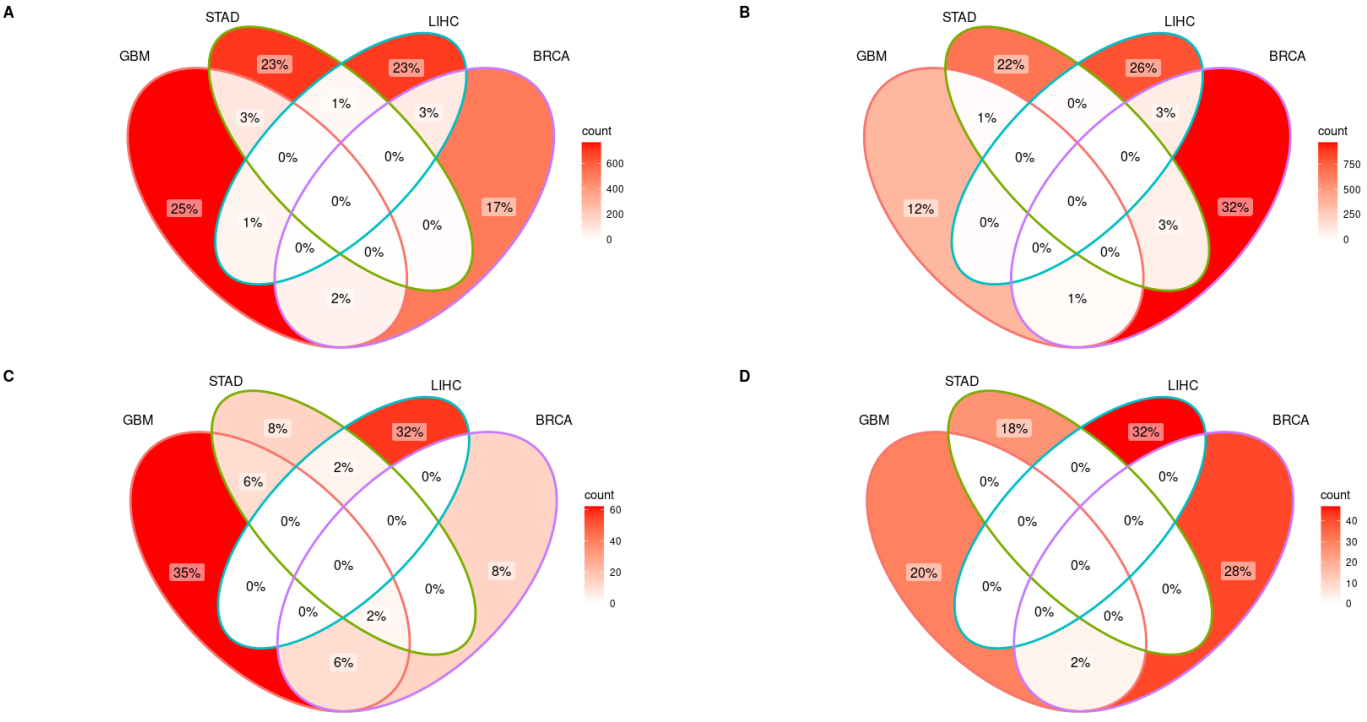

Figure 8: Venn diagrams showing the numbers of genes in percentage that were found to be overlapping or distinct between the four datasets after the filtering and aggregation steps. (A) Genes that were designated hazardous with mean HRs > 1. (B) Protective genes with mean HRs < 1. (C) Hazardous lincRNAs with mean HRs >1. (D) Protective lincRNAs with mean HRs < 1.

2.3.6 Enrichment analysis of gene sets

Gene set enrichment analysis (GSEA) plots of Figures 9-12 indicate that the four datasets are enriched for different GO terms, and the enrichment profiles for the top three GO terms for each dataset are different as well. The STAD dataset in Figure 9 displays a GSEA profile whereby genes with protective exons, i.e., genes with mean HR < 1 cluster at the end of the plot, indicating negative correlation with the listed GO terms. The GSEA profile in Figure 10 for LIHC indicates a positive correlation of the listed GO terms mitotic cell cycle and those related to RNA splicing with the identified hazardous genes of this dataset.

Genes of the BRCA set were overall enriched for vesicle-mediated transport (Figure 11), while genes with mean HR > 1 were specifically enriched for GO terms related to Golgi vesicle transport. For GBM, on the other hand, protective genes with HR < 1 were negatively enriched for GO terms related to mRNA 3'-end processing and RNA splicing reactions (Figure 12).

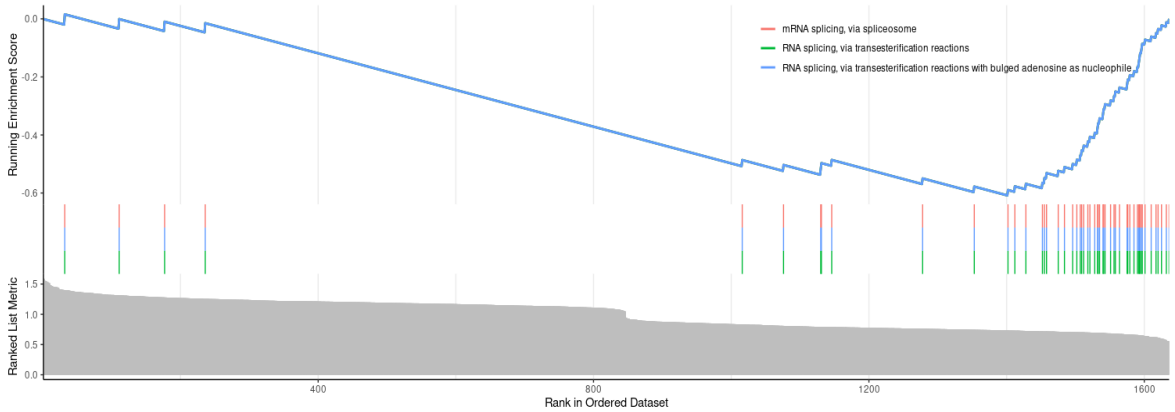

Figure 9: GSEA plot for STAD. The top three GO terms with the most significant enrichment scores are indicated with colors red, green, and blue from first to third. These terms were negatively enriched for the protective genes of this dataset.

Figure 10: GSEA plot for LIHC indicates that mainly the hazardous genes in the gene set have been enriched for terms related to the mitotic cell cycle and RNA splicing.

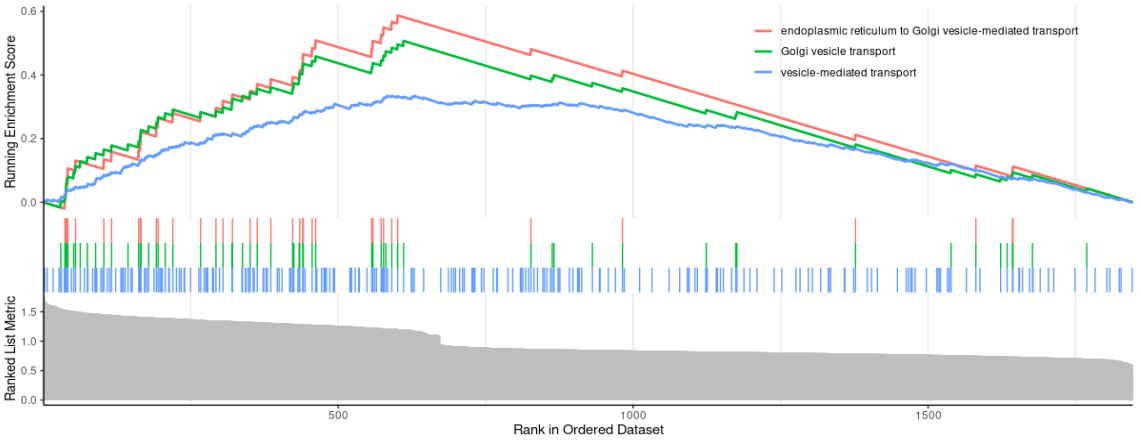

Figure 11: GSEA plot for BRCA indicates enrichment for GO terms related to vesicle-mediated transport.

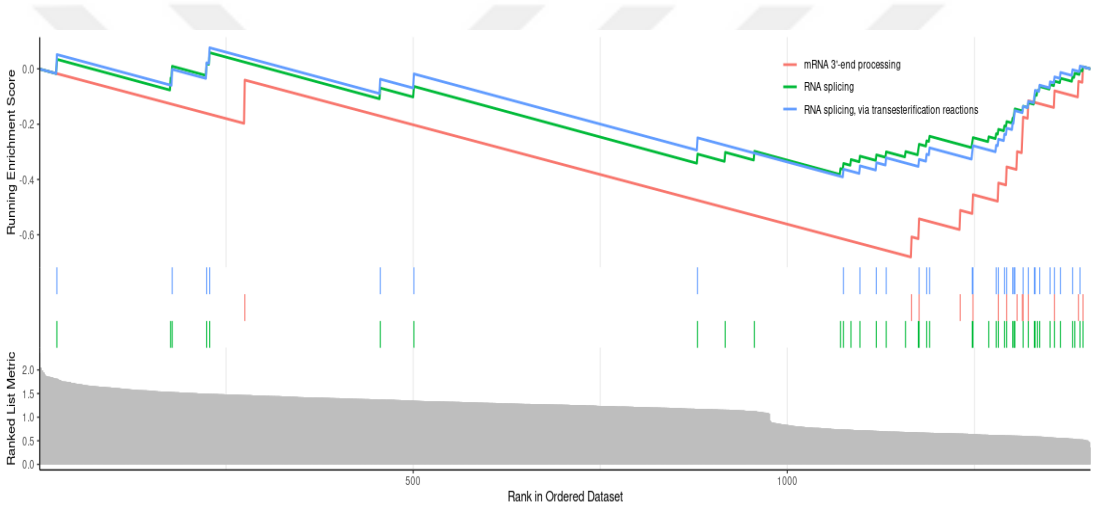

Figure 12: GSEA plot for GBM indicated negative enrichment of protective genes for the listed GO terms.

Sets of genes for each dataset, that were formed from genes with mean HR > 1 and after genes containing oppositely directed exons were excluded, are displayed in the emap plot in Figure 13. We can observe from the plot that cancer datasets seem to be enriched mainly distinctly for KEGG pathways, with overlaps between two datasets for several pathways. Genes from the hazardous gene sets of GBM and STAD were found to be jointly enriched for several KEGG pathways including those for ECM-receptor interaction, actin cytoskeleton regulation, and proteoglycans. Several KEGG pathways that are common between LIHC and BRCA datasets were also identified such as those for the proteasome, carbon metabolism, nucleocytoplasmic transport, and aminoacyl-tRNA biosynthesis. The KEGG pathway for the lysosome contained genes from both the GBM and BRCA datasets.

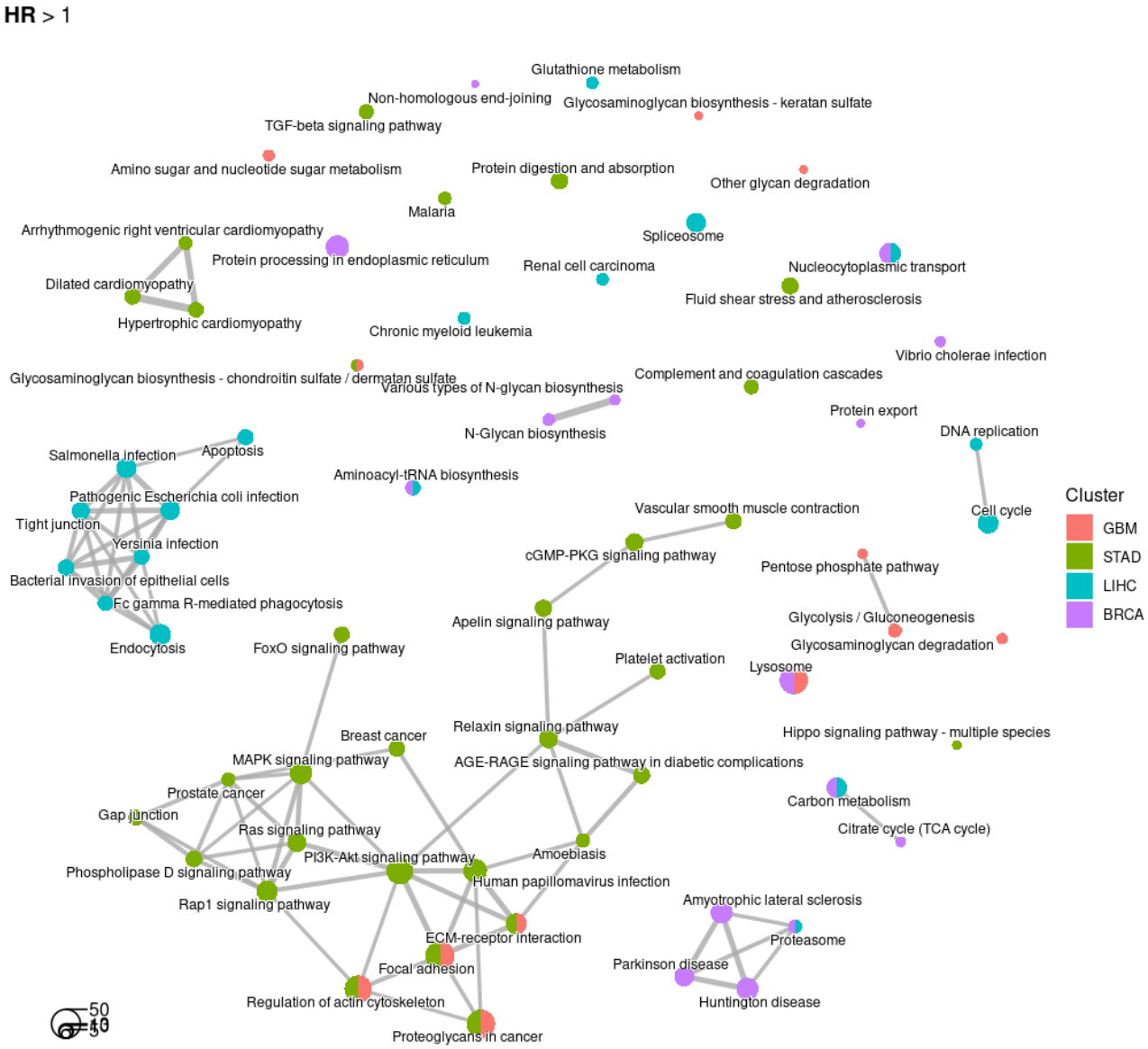

Figure 13: Emap plot for KEGG pathway enrichment of each dataset revealed several common pathways for the hazardous genes of these datasets; such as nucleocytoplasmic transport and aminoacyl-tRNA biosynthesis for LIHC-BRCA, i.e. genes with mean HR > 1 based on exon expressions.

Gene sets for each dataset, that were formed from genes with mean HR < 1 and after genes containing oppositely directed exons were excluded, are displayed in the emap plot in Figure 14. Protective genes the cancer datasets are enriched distinctly for most of the identified KEGG pathways, with overlaps in only three pathways. Protective genes from the LIHC and BRCA datasets were jointly enriched for the *Staphylococcus aureus* infection pathway, while genes from the GBM and BRCA datasets were jointly enriched for the Herpes simplex virus 1 infection pathway. The GBM, STAD, BRCA datasets contributed roughly equal numbers of genes to be enriched for the spliceosome pathway.

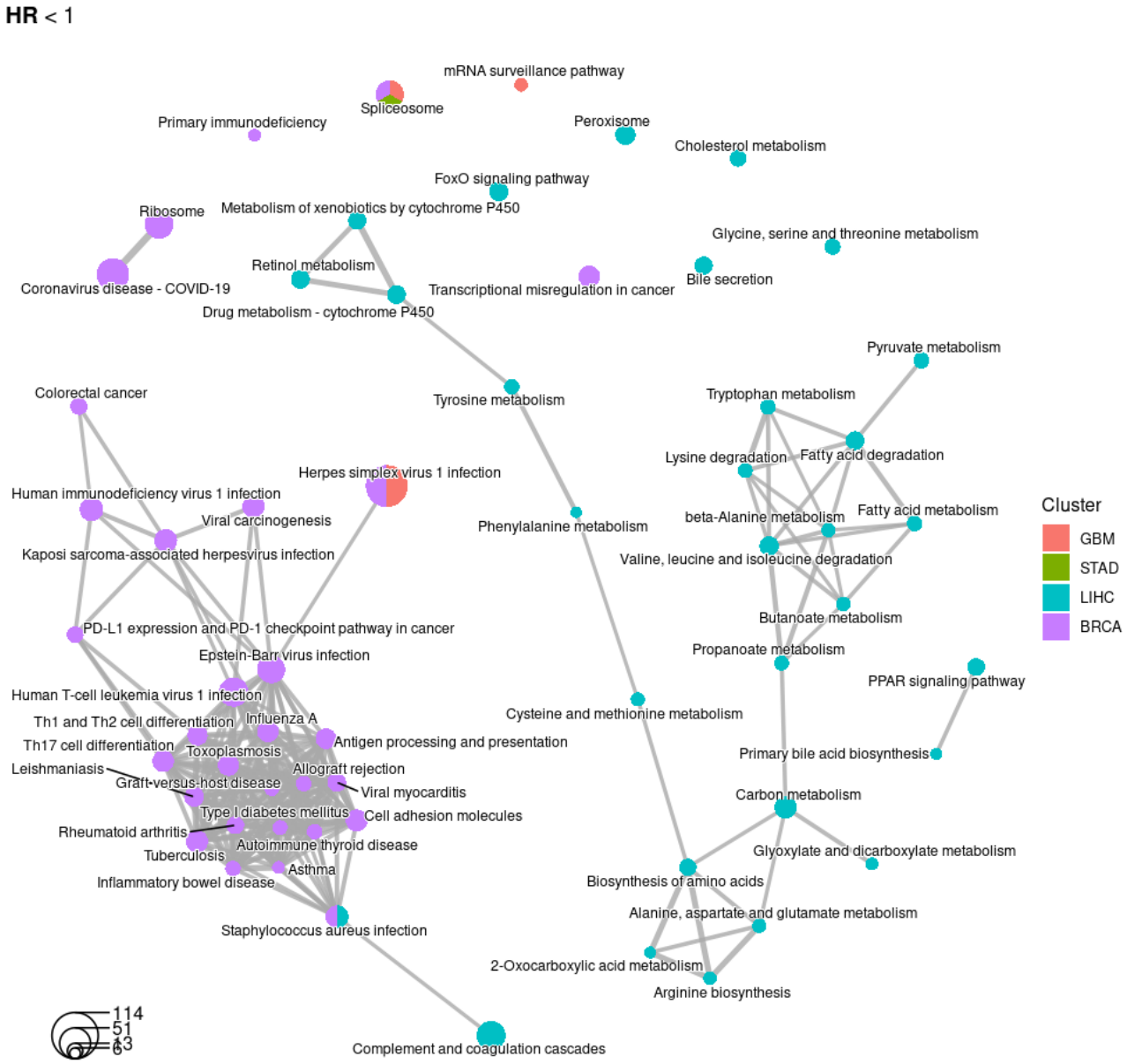

Figure 14: Emap plot for KEGG pathway enrichment of each dataset indicates that the spliceosome pathway as common for datasets GBM, STAD and BRCA, for their protective genes, i.e. genes with mean HR < 1 based on exon expressions.

2.4 Discussion

Transcriptome-wide exon-based analysis of four TCGA datasets revealed interesting results regarding each dataset separately, and when the results from the four datasets were compared. Starting with the STAD dataset, the gene with the highest HR value after exon aggregation was *GNA12* (HR $\simeq$ 1.60, adj. $p < 0.05$), which encodes the Gα_{12} subunit of guanine nucleotide-binding proteins (G proteins); known to activate ARHGEF proteins that function as guanine nucleotide exchange proteins involved in regulating the actin cytoskeleton. *GNA12* was previously shown by (Jian et al., 2013) to upregulate proinflammatory cytokines in oral squamous cell carcinoma. While the gene was not

found to be significantly DE in primary tumor vs. normal (solid tissue) cells in the STAD dataset (Welch's t-test $p \simeq 0.16$), K-M analysis with expression levels from UCSC Xena indicates that the downregulation of this gene has a positive effect on survival ($p < 0.05$), as shown in the plot in Figure 15.

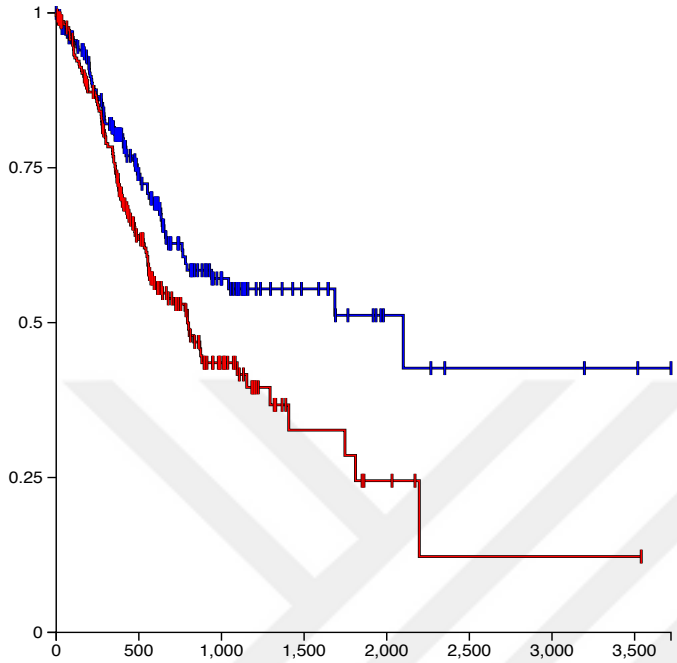

Figure 15: K-M plot with *GNA12* expressions for survival rates in STAD. The blue curve indicates the survival rate for samples with less than the median (= 10.76) expression for *GNA12* ($n = 219$), while the red curve indicates survival for samples with more than or equal to the median expression of the gene ($n = 224$).

The gene *YTHDC1* was found to be the most protective gene in STAD with HR $\simeq$ 0.56 (adj. $p < 0.01$), which is a regulator of alternative splicing (exon skipping/inclusion, see Roundtree et al. (2017)). However, when $\log_2$-transformed expressions of this gene were compared between STAD primary tumors and normal tissue, a significant difference was not observed (Welch's t-test $p \simeq 0.99$).

For GBM, the gene with the highest HR value after the filtering and aggregation steps was *CSNK1D* with a mean HR $\simeq$ 2.35 for 12 exons (adj. $p < 0.0001$). This gene is a member of the casein kinase I (CKI) family, whose members regulate important cytoplasmic and nuclear processes including DNA replication and repair. It has previous implications as an oncogene, where it was found to be highly expressed in choriocarcinomas (Stoter et al., 2005). However, a significantly differentiated prognosis was not obtained for high vs. low expression of *CSNK1D* ($p > 0.05$), as seen in the K-M plot in Figure 16.

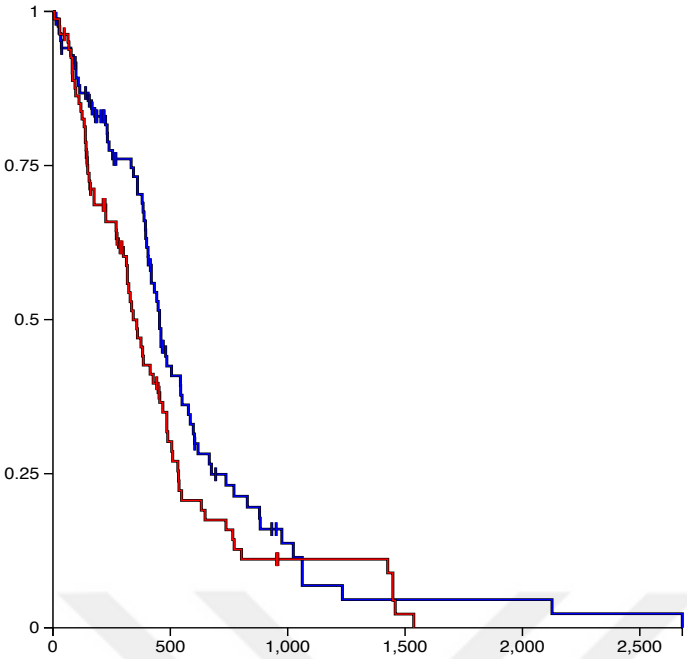

Figure 16: K-M plot with *CSNK1D* expressions for survival rates in GBM. The blue curve indicates the survival rate for samples with less than the median (= 11.91) expression for *CSNK1D* ($n = 85$), while the red curve indicates survival for samples with more than or equal to the median expression of the gene ($n = 81$).

The most protective gene in the GBM exon aggregation analysis was identified as *AP3M1* (HR $\simeq 0.36$, adj. $p < 0.001$). This gene encodes the medium subunit of the heterotetrameric AP-3 complex; which is involved in vesicle transport and protein sorting from the Golgi and has mainly been associated with neuropsychiatric and neurodegenerative disorders (Grupe et al., 2006; Hashimoto et al., 2009). Like *CSNK1D*'s result, a significant differentiation was not observed ($p > 0.1$) between samples with expressions higher or lower than the median expression of *AP3M1* (= 10.03).

The gene that was found to have the highest HR value after the processing steps in LIHC was *PAFAH1B2* (HR $\simeq 1.77$, adj. $p < 0.001$), for which a single exon had transcriptome-wide significance. This gene encodes for a subunit of an acetylhydrolase that inactivates platelet-activating factor and was previously implicated in pancreatic cancer metastasis (Ma et al., 2018). However, K-M analysis indicates LIHC samples do not significantly differentiate ($p > 0.1$) with respect to median (= 7.725) *PAFAH1B2* expression. *AKAP13* was identified as the most protective gene in TCGA-LIHC from the exon aggregation analysis in this chapter (HR $\simeq 0.57$, adj. $p < 0.001$). Indeed, comparison of normalized and $\log_2$-transformed *AKAP13* expressions between primary tumor and normal samples for LIHC indicates that this gene is downregulated in tumors, (Welch's t-test $p < 0.001$) with $\log_2$ fold-change $\simeq -0.15$ (Figure 17).

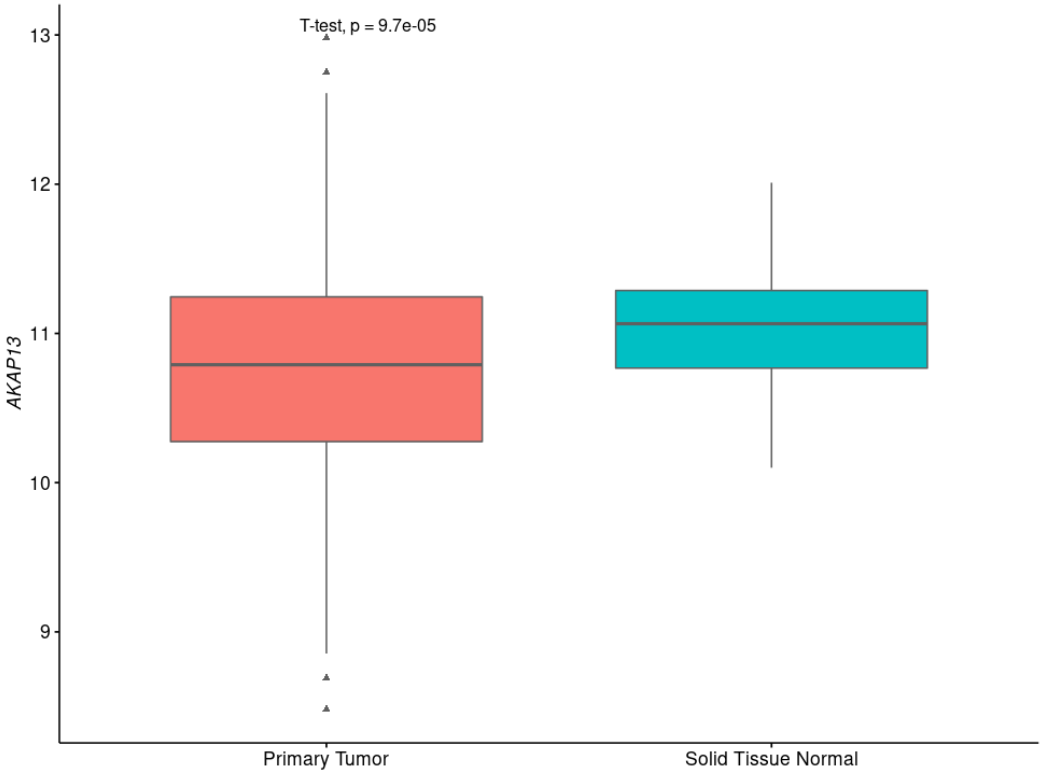

Figure 17: Comparison of normalized expressions of *AKAP13* in TCGA-LIHC for primary tumors vs. normal tissue.

The *CARS1* gene was identified to have as the most hazardous gene in TCGA-BRCA, with HR $\simeq$ 1.77 (adj. $p < 0.0001$), which is a cysteinyl-tRNA synthetase previously identified within a fusion protein involved in ALK-positive anaplastic large-cell lymphoma (Cools et al., 2002). Using Welch's t-test with $\log_2$-transformed expressions of this gene ($p < 0.0001$) obtained from UCSC Xena, *CARS1* was found to be upregulated in tumor cells with $\log_2$ fold-change ($\log FC$) $\simeq$ 0.33 as indicated in the boxplot of Figure 18.

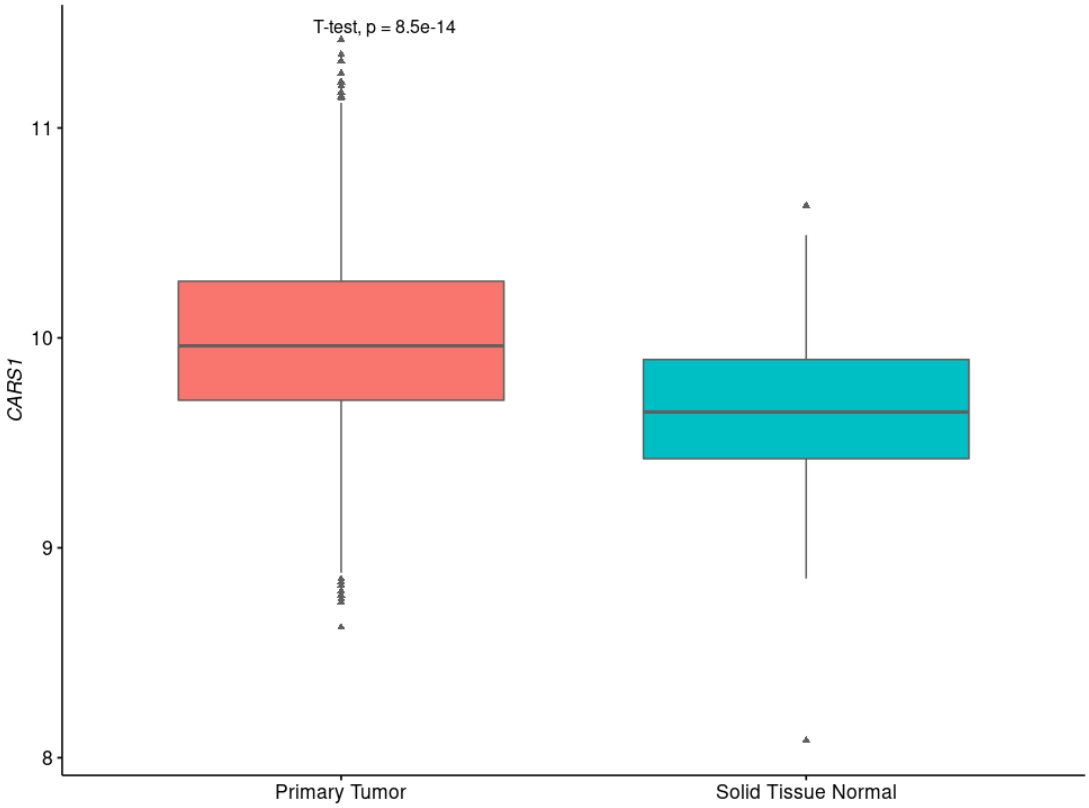

Figure 18: Comparison of normalized expressions of *CARS1* in TCGA-BRCA between primary tumor and normal samples.

The most protective gene identified in BRCA was *RBMXL1* (HR $\simeq 0.45$, adj. $p < 0.001$), a retrogene of an X-linked protein *RBMX*. This gene is significantly downregulated in primary tumors of BRCA when compared to normal breast tissue (Welch's t-test $p < 0.0001$, $\log FC \simeq -0.56$). *RBMX* and *RBMXL1* RNA-binding proteins were recently found to be involved chromatin maintenance in myeloid leukemia (Prieto et al., 2021).

An interesting outcome of the analyses in this chapter was the identification of individual exons with transcriptome-wide impact and significance; for which examples can be given from GBM with exons 38132 (HR $\simeq 2.36$, $p < 0.001$) and 54759 (HR $\simeq 0.38$, $p < 0.01$), belonging to genes *ADCY3* and *EIF2B5*, respectively. *ADCY3* encodes adenylyl cyclase 3, a membrane-associated enzyme involved in the catalysis of the signaling molecule cyclic adenosine monophosphate (cAMP) in response to G-protein signaling, and its loss-of-function has been associated with severe obesity (Saeed et al., 2018). *EIF2B5* is a subunit of the eukaryotic translation initiation factor 2B, previously found to have a hazardous role in diffuse large B-cell lymphoma (Unterluggauer et al., 2018), while a rare variant of this gene was found to be protective in ovarian cancer (Goode et al., 2010).

Designing a drug that aims to inactive the region corresponding to Exon 38132, the eighteenth exon on *ADCY3* which encodes for its nucleotide cyclase activity in GBM patients could be useful in a therapeutics study. Similarly, a drug that aims to enhance the

activity of the region encoded by Exon 54759, which corresponds to the third exon of *EIF2B5* and forms part of the nucleotide-diphospho-sugar transferase domain of its protein could be designed as a therapeutic aid for GBM.

Unlike the conventional method of ranking genes based on p-values for GSEA, since the gene lists of each cancer dataset were already filtered for significance and expression; mean HR values calculated from the genes' exons were used for ranking purposes. For each dataset, genes consisting of exons with HR values that were either all < 1 or all > 1 were ranked according to their mean HR values from highest to lowest before being used inside the gseGO() function.

Each dataset revealed a distinct GO term enrichment profile, for both their hazardous and protective gene sets, as observed in the GSEA and emap plots of Section 2.3.6. However, the emap plot for the protective genes indicate that these sets are more distinctly enriched than the hazardous gene sets. The KEGG pathways that were prominent for the hazardous gene sets included those related to infection, apoptosis, phagocytosis, endocytosis, tight junctions, DNA replication, and the cell cycle for LIHC and MAPK, Ras, PI3-Akt, cGMP-PKG, phospholipase D, Rap1, Relaxin, Apelin, FoxO, and AGE-RAGE signaling, as well as those for gap junctions and cardiomyopathy for STAD. Pathways enriched for BRCA included those for Parkinson ad Huntington disease, interestingly, and the N-glycan biosynthesis pathway. KEGG pathways enriched for GBM hazardous genes included those for glycolysis/gluconeogenesis and the pentose phosphate pathway.

In the emap plot for the protective genes of each dataset, KEGG pathways were mainly divided between the BRCA and GBM datasets and consisted of mostly interconnected pathways. Protective pathways identified in GBM were mainly metabolic, including those for carbon, butanoate, pyruvate, cholesterol, and fatty acid metabolism, as well as metabolic, biosynthetic, and degradation pathways for various amino acids such as arginine, cysteine, methionine, valine, leucine, phenylalanine, tryptophan, and tyrosine. GBM protective pathways also included several signaling pathways such as PPAR and FoxO. Protective pathways identified in BRCA included those for different types of viral infection such as EPV, human immunodeficiency virus 1, human T-cell leukemia, and influenza A. Other pathways identified from the protective gene set in BRCA included those for Th1, Th2, and Th17 cell differentiation, the PD-1 checkpoint pathway, tuberculosis, asthma, and inflammatory bowel disease.

While there were no genes or lincRNAs, hazardous or protective, that were common in the sets of all four cancers; several have been identified that were common in three datasets. For instance, the *UTP14C*, a member of the ribonucleoprotein complex small-subunit processome, came out as protective in all three of GBM, STAD and LIHC datasets. The lincRNAs *TCONS_l2_00003186*, *TCONS_l2_00003187* and *TCONS_l2_00003188*, for which functional annotation seems lacking, were found to be hazardous in datasets BRCA, GBM and STAD. These genes and lincRNAs, identified as commonly involved in multiple cancers, could be important as drug development targets for drugs that are aimed at treating more than one type of cancer.

Through the novel workflow developed with computational methods applied to four TCGA datasets in this chapter; exons, lincRNAs and genes that had not been previously

implicated in these cancers were identified. Additionally, GO terms and KEGG pathways that were distinctly and commonly enriched for each dataset were revealed. This univariable survival analysis followed by exon aggregation methodology could be applied to the remaining datasets in TCGA to identify similarities and differences for the genes and enrichment profiles identified from each dataset across TCGA-PANCAN. It is important to note that the identified exons should be confirmed for their functional significance *in vivo* with respect to the cancer dataset they were identified in, to understand their role in cancer. Similar validation experiments could be performed with the gene sets identified for each dataset to determine whether the prioritized genes have a role in the development and prognosis of each cancer.

3 UNIVARIABLE SURVIVAL ANALYSIS OF FOUR TCGA DATASETS WITH ISOFORM EXPRESSIONS

3.1 INTRODUCTION

3.1.1 ALTERNATIVE SPLICING

Alternative splicing is the process of rearrangement of the pattern of intron and exon elements of a gene, which are joined by splicing, to alter the mRNA coding sequence, which in turn enables the synthesis of different protein variants with distinct cellular or functional properties. There are five main types of alternative splicing: constitutive splicing (also defined as exon skipping or cassette exon), in which an exon may be alternatively spliced out or retained; alternative donor site, when an alternative 5' splice junction changes the upstream exon's boundary; alternative acceptor site, when an alternative 3' splice junction changes the downstream exon's boundary; mutually exclusive exons, when either one of two exons is retained after splicing; and intron retention, when an intron may be alternatively spliced out or retained (Wang et al., 2015). Different transcripts produced from the same gene with alternative splicing are indicated in Figure 19.

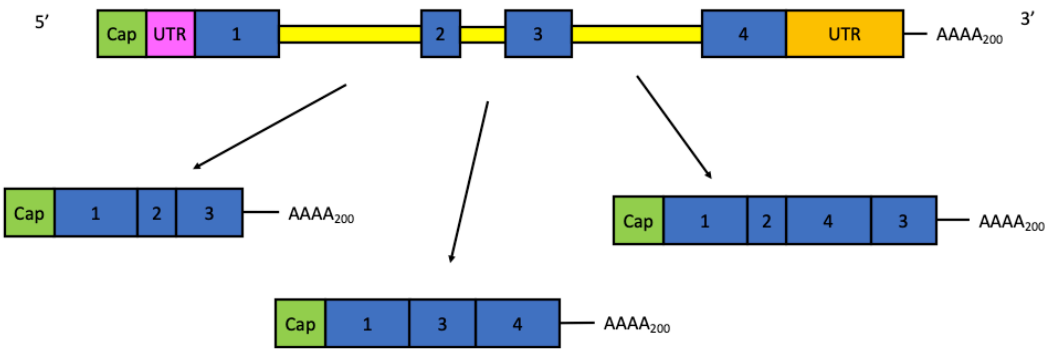

Figure 19: Different transcripts of the same gene, produced by alternative splicing. In the left transcript, the fourth exon is excluded in the mRNA; in the transcript in the middle, the second is excluded, while the third and fourth exons have switched orders in the transcript on the right.

Several previous studies have focused on associating the dysregulation of mRNA splicing with various types of cancer, and cancer cells are known to have cancer type and subtype-specific alterations in splicing patterns with prognostic value (Bonnal et al., 2020). Alternative splicing variants can have distinct effects on cancer prognosis. For example, the *BCL2L1* gene can be transcribed into either a small isoform Bcl-x_S, which is pro-apoptotic; or a large isoform Bcl-x_L retaining an intron, which promotes stemness and aggressiveness in melanomas and glioblastomas (Trisciuoglio et al., 2017). Recent cancer studies have focused on drugs that target the splicing machinery to prevent the production

of oncogenic isoforms of certain proteins, such as the angiogenic isoform of VEGF, thereby aiming to inhibit angiogenesis (Bonnal et al., 2020).

In the glioblastoma study by Sadeque et al. (2012), AEU by the *TTN* gene was found to be the most significantly associated with survival. In another study involving high-grade serous ovarian carcinoma, various isoforms of *TP53* were compared and the Δ133p53 isoform was associated with survival (Bischof et al., 2019) using CPH models involving additional clinical factors. In another study involving *TP63* isoforms in twenty-nine TCGA datasets, the gene's DNp63 isoform was positively associated with survival, while its TAp63 isoform was negatively associated with survival in bladder, breast, and non-small cell lung cancer cohorts (Bankhead et al., 2020). While individual studies have associated distinct isoforms of certain genes with survival in specific cancers, a study that univariably computes and ranks HR values of isoform expressions in the hg19 transcriptome in multiple cancer cohorts is not present in literature. Therefore, this approach formed the basis of the computations and analyses in this chapter.

Although there exist numerous studies regarding the role of AEU and alternative splicing in cancer, in which various isoforms have been identified and associated with cancer progression, transcriptome-wide analyses involving multiple cancer datasets in which isoforms have been ranked based on their univariable HR values.

3.1.2 TCGA DATASETS USED IN THE TRANSCRIPT-BASED ANALYSES

The same TCGA datasets, namely STAD, GBM, BRCA, and LIHC that were investigated in the exon-based analyses were investigated also in the transcript-based analyses; whereas this time samples of RNA-Seq counts of isoforms were used. Sample sizes for the isoform-based analyses of the four datasets are shown in Table 16.

Table 16: Sample sizes of TCGA datasets used in the transcript-based analyses, consisting of primary tumors for each dataset.

STAD	GBM	BRCA	LIHC
382	155	1061	363

3.2 METHODS

3.2.1 DATA ACQUISITION

For each of the four cancer datasets, the GDCdownload() function of the "TCGAbiolinks" Bioconductor package (Colaprico et al., 2016) was used to download the folders containing the TXT files of isoform expression quantification from the GDC legacy archive, as in Chapter 2. The GDCquery() function of the same package was then used to access cancerous tissue sample names and information for each of the downloaded TXT files. This information was then used to construct a raw counts matrix for each of the four cancer datasets from raw, unnormalized RNA-Seq counts of primary

tumor samples sequenced with the same Illumina HiSeq 2000 v2 platform that were submitted from UNC. The sample names in the raw counts matrix were matched and confirmed with those in the survival TXT file downloaded earlier with "UCSCXenaTools".

3.2.2 PRE-PROCESSING

Pre-processing of the isoform raw counts matrices for all four TCGA datasets was carried out using default parameters of the calcNormFactors() function of the package "limma" (Ritchie et al., 2015) and the cpm() function of the package "edgeR" (Robinson et al., 2010) to obtain their TMM-normalized $\log CPM$ expression values.

3.2.3 DATA ANALYSIS AND INTERPRETATION

Entrez GENEIDs were accessed through UCSC's knownGene library, while the same library's transcriptsBy() function was used to obtain transcript names of each GENEID. The HR and p-value with using coxph(), as well as the $\log CPM > 1$ counts, mean $\log CPM$ and additional calculations were performed on pre-processed counts of each isoform in the matrix that had matching transcript names in the hg19 library. Isoform names that were not found in the hg19 coding transcriptome were designated "not in panel" and calculations were not performed on them. The $\log CPM > 1$ count for each isoform was used as a filter to remove lowly-expressed isoforms; with thresholds set at 270, 108, 255, and 750 for the datasets STAD, GBM, LIHC, and BRCA, respectively. Only isoforms with HRs that had transcriptome-wide significance, i.e., those with $p <$ 0.05 were used in the downstream analyses. The filtered isoforms table was then further annotated with the addition of "GENENAME" and "SYMBOL" columns using the "org.Hs.eg.db" annotation package (version 3.11.4), for each of the four investigated datasets.

3.2.4 AIM

- Identify transcripts in the hg19 transcriptome with an impact on survival rates in the four investigated TCGA datasets: STAD, GBM, LIHC, and BRCA, with univariable survival analyses.

3.3 RESULTS

3.3.1 STOMACH ADENOCARCINOMA

The effect TMM normalization on ten TCGA-STAD isoform samples is represented as before-after boxplots in Figure 20. Median expressions are more uniformly distributed after normalization as expected, although they would be more uniform if the filtering step was not postponed as in the pipeline of this study.

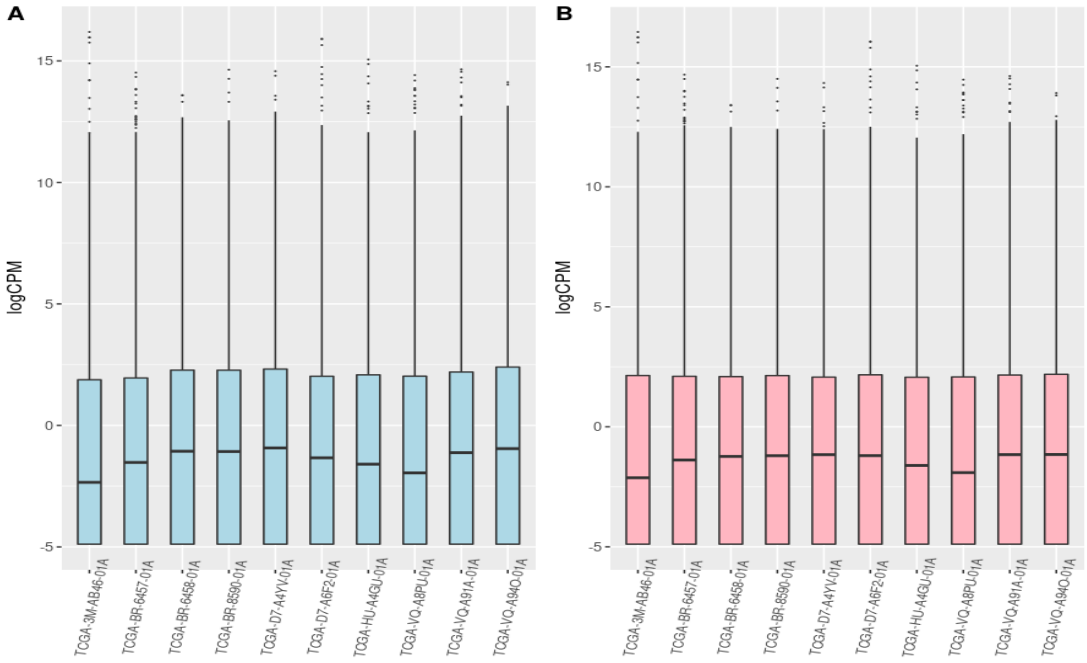

Figure 20: Effect of TMM normalization on STAD isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Transcripts identified in the hg19 transcriptome with transcriptome-wide impact and significance for the STAD dataset, with the highest and lowest HRs were listed in Table 17. Genes that each transcript in the list belongs to were indicated. The transcript at the top of the list, uc003tdb.2, belongs to *KBTBD2* and the transcript uc002lqr.1, which was found to have the lowest HR value in STAD, belongs to *TMEM259*.

Table 17: Transcripts with the ten highest and lowest HR values whose expressions have transcriptome-wide significance for survival in TCGA-STAD.

Transcript name	HR	*P*-value	Gene symbol
uc003tdb.2	1.49030938323074	0.0322415162126896	*KBTBD2*
uc003iah.2	1.47177202507974	0.00041398217319473	*FAM241A*
uc003krt.1	1.46370071888738	0.00814968537788309	*COMMD10*
uc001mel.1	1.46305699231201	0.0196449927142444	*TPP1*
uc002elf.1	1.46281553349279	0.0306146538576733	*ARL2BP*
uc002ndu.2	1.44825305449565	0.01446718388509	*AP1M1*
uc002uwe.2	1.42139272685279	0.000444732778624805	*CLK1*
uc001zxc.1	1.38646638920435	0.00162774386006401	*EID1*
uc009zwk.1	1.37989919175844	0.0374243224992327	*MAP1LC3B2*
uc002kbw.1	1.3629532653746	0.0163378073727372	*ANAPC11*
uc001cce.1	0.641201094157814	0.00700236237274391	*MTF1*
uc003bck.1	0.641131354395721	0.000711102631523756	*TCF20*

uc004dly.1	0.635092440515765	0.017368104807835	*GRIPAP1*
uc002qfw.2	0.626160681193593	0.00437936592055777	*LENG8*
uc004clr.1	0.616393409765814	0.00164070524034723	*ANAPC2*
uc002mtz.2	0.613746892764784	0.00181969118954915	*ZNF490*
uc002mrg.1	0.61241929741883	0.00110150917702067	*SWSAP1*
uc001ogy.1	0.612141770373844	0.0215124236004199	*SF3B2*
uc001noc.1	0.604071486039911	0.01209757238344	*OSBP*
uc002lqr.1	0.604041099827669	0.00386651200556031	*TMEM259*

Genes with oppositely directed transcripts in the hg19 transcriptome identified in TCGA-STAD samples are given in Figure 21. In Figure 21A, two transcripts of *CLK1* are shown with transcriptome-wide significance and oppositely directed HR values, in which uc002uwe.2 expression was positively associated with survival (HR $\simeq$ 0.87, $p < 0.001$) while uc002uwf.2 expression was negatively associated (HR $\simeq$ 1.42, $p < 0.05$). Figure 21B represents *BICD2* gene's transcripts: uc004aso.1, which was negatively associated with survival (HR $\simeq$ 1.28, $p < 0.01$) in STAD and uc004asp.1, which was positively associated (HR $\simeq$ 0.83, $p < 0.05$). In Figure 21C, transcripts of *ESYT2* are represented, whereby uc003wob.1 is slightly negatively associated (HR $\simeq$ 1.05, $p < 0.05$) and uc003wod.1 slightly positively associated (HR $\simeq$ 0.88, $p < 0.01$) with survival.

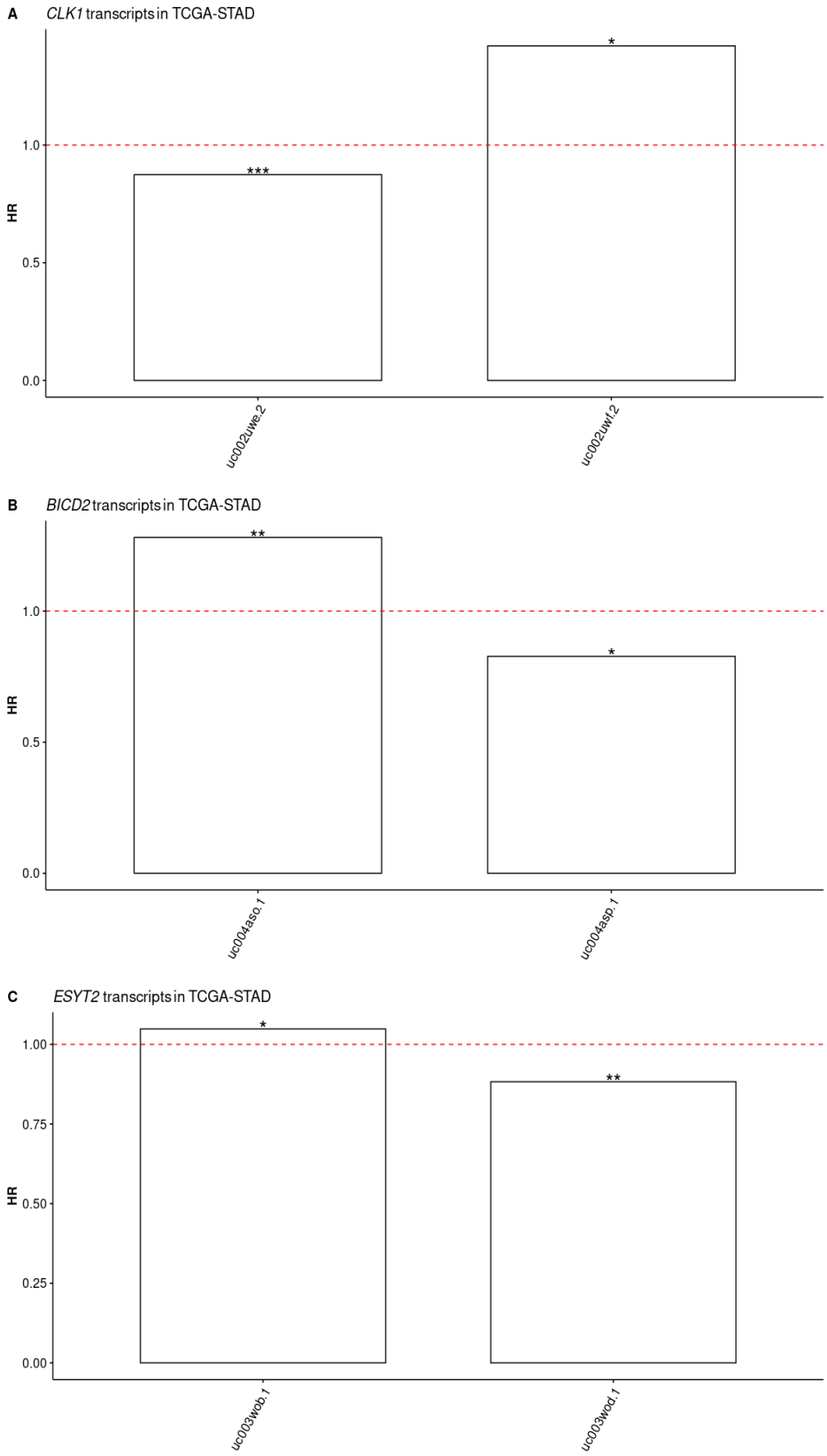

Figure 21: Genes with transcripts that have opposite effects on survival in STAD. (A) *CLK1* with two transcripts of opposite impact on survival rates. (B) One hazardous and one protective transcript for *BICD2* were identified. (C) *ESYT2* has a hazardous and protective transcript with transcriptome-wide significance. Univariable significance levels for the isoforms were indicated on each bar with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$.

3.3.2 GLIOBLASTOMA MULTIFORME

The effect *TMM* normalization on ten TCGA-GBM isoform samples is represented as before-after boxplots in Figure 22. Median isoform expressions of the samples are more uniformly distributed, as expected, after the normalization step.

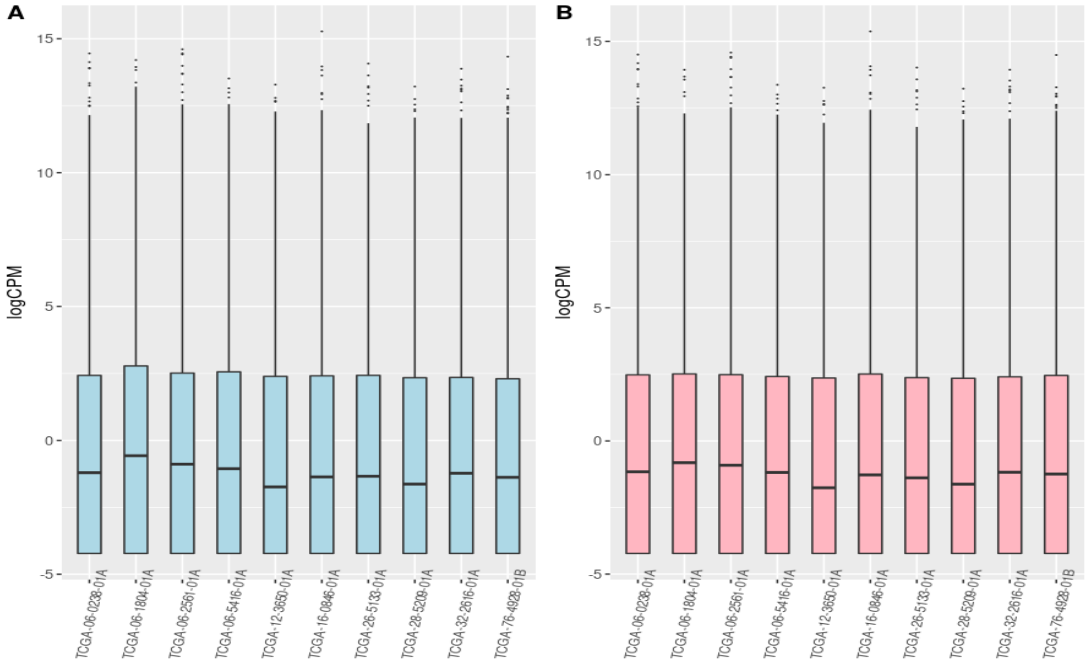

Figure 22: Effect of *TMM* normalization on GBM isoform samples. (A) Distribution of log*CPM* expressions of raw (unnormalized) samples. (B) Distribution of log*CPM* expressions of *TMM*-normalized samples.

Transcripts with the highest and lowest HR values associated with survival in GBM are listed in Table 18. The transcript with the highest univariable transcriptome-wide HR was found to be uc002dbg.1, which belongs to *TXNDC11*. The uc002vpc.2 transcript of *AGFG1* was found to have the lowest HR with transcriptome-wide significance.

Table 18: Transcripts identified in GBM with the highest and lowest HRs whose expressions associated either positively or negatively with survival.

Transcript name	HR	*P*-value	Gene symbol
uc002dbg.1	2.36191814323174	0.00434068294365401	*TXNDC11*
uc001loe.2	2.09831043708603	0.000954719488294137	*BET1L*
uc003jam.1	2.06506720275083	0.00855094937922387	*CCDC127*
uc002huc.2	2.02598611351402	0.00406201287284883	*MSL1*
uc003vwi.2	1.98184483249975	0.00949833844872197	*AGK*
uc003knj.1	1.94969067136353	0.0021058185579853	*FAM174A*
uc003uzh.2	1.89019181557051	0.00462526618422144	*PRKRIP1*
uc011kgj.1	1.78359238385237	0.00299510722313977	*YWHAG*
uc001ulk.1	1.74338070408723	0.0013336207700299	*ZNF605*

uc003udw.1	1.6812965255194	0.00197991686200317	*RHBDD2*
uc010oxe.1	0.603322044928111	0.0269081900614912	*WDR3*
uc002ysa.2	0.584929568616365	0.00962693778468905	*GART*
uc001oxh.1	0.56915284836011	0.00896490879898311	*THAP12*
uc001hpd.1	0.558625149367795	0.000339792932693755	*ENAH*
uc003fjr.1	0.548159407043107	0.0327828255399704	*ZNF639*
uc001knd.1	0.546957405673378	0.00231290017834643	*FRAT2*
uc001kjq.1	0.513140010251272	0.00658234430769432	*NOC3L*
uc002sqw.2	0.508731232119671	0.0242134368775433	*PTCD3*
uc001khx.1	0.497130075970697	0.0061657245645428	*MARCHF5*
uc002vpc.2	0.494162046733363	0.00969383175144585	*AGFG1*

3.3.3 LIVER HEPATOCELLULAR CARCINOMA

TMM normalization on ten isoform expression samples from TCGA-LIHC are represented in before-after boxplots in Figure 23. Due to filtering of lowly-expressed isoforms after the analysis stage, the difference in median expressions before and after normalization cannot be clearly observed.

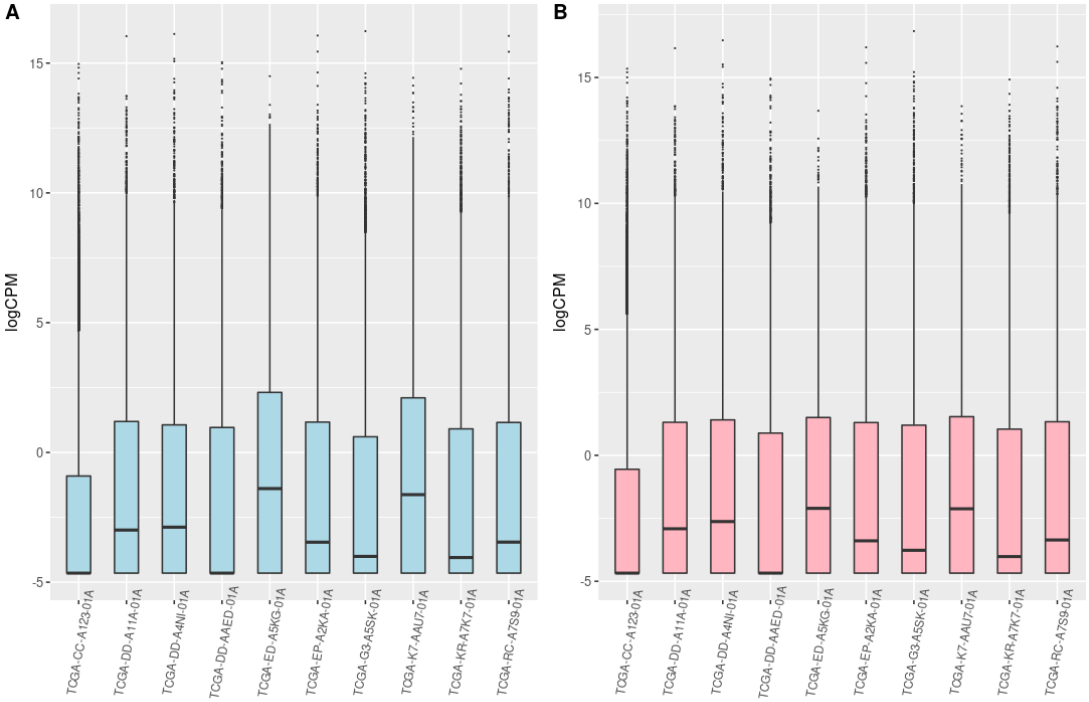

Figure 23: Effect of *TMM* normalization on LIHC isoform samples. (A) Distribution of log*CPM* expressions of raw (unnormalized) samples. (B) Distribution of log*CPM* expressions of *TMM*-normalized samples.

Transcripts identified from the univariable transcriptome-wide analysis with TCGA-LIHC with transcriptome-wide significance are listed in Table 19 for transcripts with the ten highest and lowest HRs. The transcript with the highest HR, uc001tqo.2, belongs to *ANAPC7* and uc002ano.2, with the lowest HR, belongs to *OAZ2*.

Table 19: Transcripts with the highest and lowest HRs with transcriptome-wide significance in TCGA-LIHC.

Transcript name	HR	*P*-value	Gene symbol
uc001tqo.2	1.98774917999966	0.000128874309860887	*ANAPC7*
uc001yks.2	1.97182816888499	0.000144948587809726	*DYNC1H1*
uc009vsm.1	1.94907064315932	0.000585008067307853	*ARID1A*
uc003jtn.1	1.88282281208279	0.0014781230030603	*CWC27*
uc010qfe.1	1.85354133876577	0.000262680185275297	*HNRNPA3P1*
uc001nqv.1	1.82444719075355	0.00136008191348999	*VPS37C*
uc001cor.1	1.81697597308357	0.000102859072801516	*TMEM69*
uc001ctj.1	1.79507762455527	0.00260789146214868	*KTI12*
uc001dot.2	1.74540984568353	0.00123930877108164	*RPAP2*
uc003fyc.2	1.73063391490209	0.00405118211924759	*RUBCN*
uc001ktu.2	0.703903719799845	0.0242653143081643	*OGA*
uc002dbg.1	0.701481108447651	0.0199737987175209	*TXNDC11*
uc001zon.2	0.69443985438689	0.00283275386867636	*JMJD7*
uc010cnt.1	0.683219888951774	0.0047909330683391	*VAMP2*
uc002oin.1	0.67665719946426	0.0134726483214993	*FAM98C*
uc001uzy.1	0.671457575545272	0.00184840682207335	*TPT1*
uc001zkq.2	0.667057710411985	0.0394547266636981	*SRP14*
uc002nen.1	0.58929945492222	0.0211407820838448	*MED26*
uc002sgb.1	0.581697134690302	0.000418137114490637	*PCBP1-AS1*
uc002ano.2	0.54042085286879	0.000932827918852521	*OAZ2*

The gene identified in LIHC, *AKAP8L*, was found to have two transcripts with transcriptome-wide significance; uc002naw.1, which was negatively associated (HR ≃ 1.48, $p < 0.01$) and uc002nax.1, which was positively associated (HR ≃ 0.86, $p < 0.01$) with survival as shown in Figure 24.

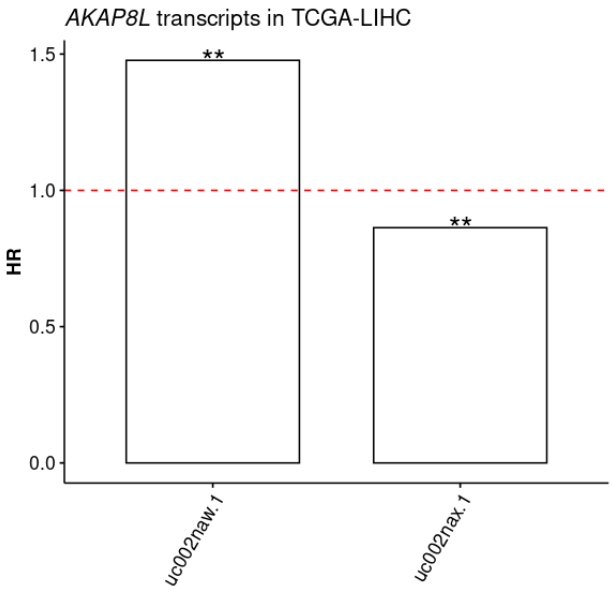

Figure 24: Two different transcripts of *AKAP8L* with opposing effects on survival in TCGA-LIHC. Univariable significance levels for the isoforms were indicated on each bar with ** for $p < 0.01$.

3.3.4 BREAST INVASIVE CARCINOMA

Ten randomly selected samples from TCGA-BRCA are shown in Figure 25, before and after TMM normalization. The reason for median expression values of TCGA samples not being completely aligned after TMM normalization is that lowly-expressed isoforms were removed only after the normalization and transcriptome-wide analysis steps.

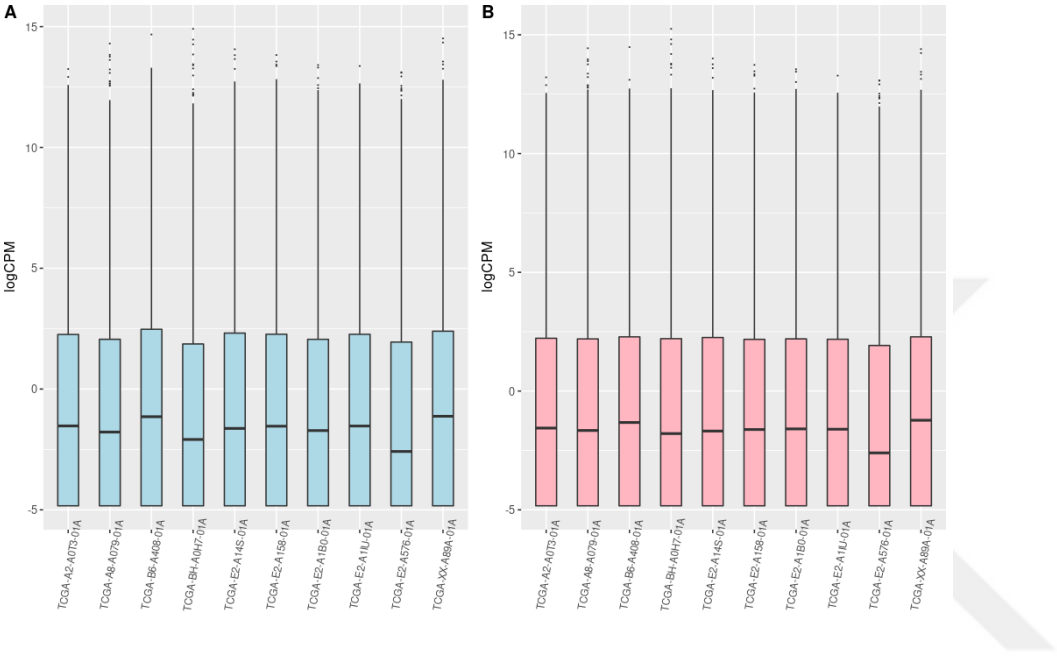

Figure 25: Effect of TMM normalization on BRCA isoform samples. (A) Distribution of logCPM expressions of raw (unnormalized) samples. (B) Distribution of logCPM expressions of TMM-normalized samples.

Transcripts with the highest and lowest HR values identified from the univariable transcriptome-wide survival analysis with isoform expressions in TCGA-BRCA are listed in Table 20. The transcript with the highest HR, uc003qqv.1, belongs to *SNX9* and the transcript with the lowest HR, uc011bax.1, belongs to *PTPN23*.

Table 20: Transcripts associated with survival univariably in TCGCA-BRCA with the ten highest and lowest HR values.

Transcript name	HR	*P*-value	Gene symbol
uc003qqv.1	1.70913425790335	0.00124941435116211	*SNX9*
uc003qdg.2	1.68789422514354	0.00154155843420925	*STX7*
uc003lgg.1	1.66395791301949	0.0215650013479996	*SLC35A4*
uc010jaj.1	1.64067689105933	0.00404870960596577	*CRSP8P*
uc002sfp.2	1.63722943369199	0.00762333157199924	*AAK1*
uc004byk.1	1.62304512623039	0.00279801815189127	*TOR1B*
uc003llf.2	1.62168492430993	0.0196737513790482	*HDAC3*
uc002elt.1	1.60277003382119	0.00242615714425921	*POLR2C*
uc003qez.2	1.5923632828916	0.000360483286205488	*HBS1L*

uc003ymw.1	1.58668561442004	0.000259218817186538	*EMC2*
uc001enp.1	0.698581754036258	0.00942437446221363	*POLR3GL*
uc002vbf.1	0.694671915714624	0.00221610726938053	*EEF1B2*
uc001noh.2	0.684621550166467	0.0255198525526045	*MRPL16*
uc001bxc.1	0.67978154325122	0.0305253308429461	*ZNF362*
uc003hmm.1	0.678693300615922	0.0299894866006606	*HNRNPD*
uc010xtb.1	0.67598806976324	0.0076442709313439	*THAP8*
uc003zzu.1	0.658625945631463	0.00824992488957586	*GRHPR*
uc001vgz.1	0.65185161570644	0.00922537268127928	*HNRNPA1L2*
uc002szs.1	0.622704887403631	0.00955045244181836	*MITD1*
uc011bax.1	0.557252876131617	0.000706650843598064	*PTPN23*

Two genes, *DDX6* and *EEF1B2*, were identified in TCGA-BRCA with transcripts that were found to have opposing effects on survival with transcriptome-wide significance are shown in Figure 26. Two transcripts of *DDX6* are indicated in Figure 26A: uc001pua.2, which was slightly positively associated with survival (HR $\simeq$ 0.92, $p < 0.05$) and uc001puc.2, which was slightly negatively associated (HR $\simeq$ 1.06, $p < 0.05$). In Figure 26B, two transcripts of *EE1B2* are shown with opposite effects on survival: uc002vbf.1's expression which was found to be univariably protective (HR $\simeq$ 0.70, $p < 0.01$) and uc002vbh.1's expression which was univariably hazardous (HR $\simeq$ 1.41, $p < 0.05$).

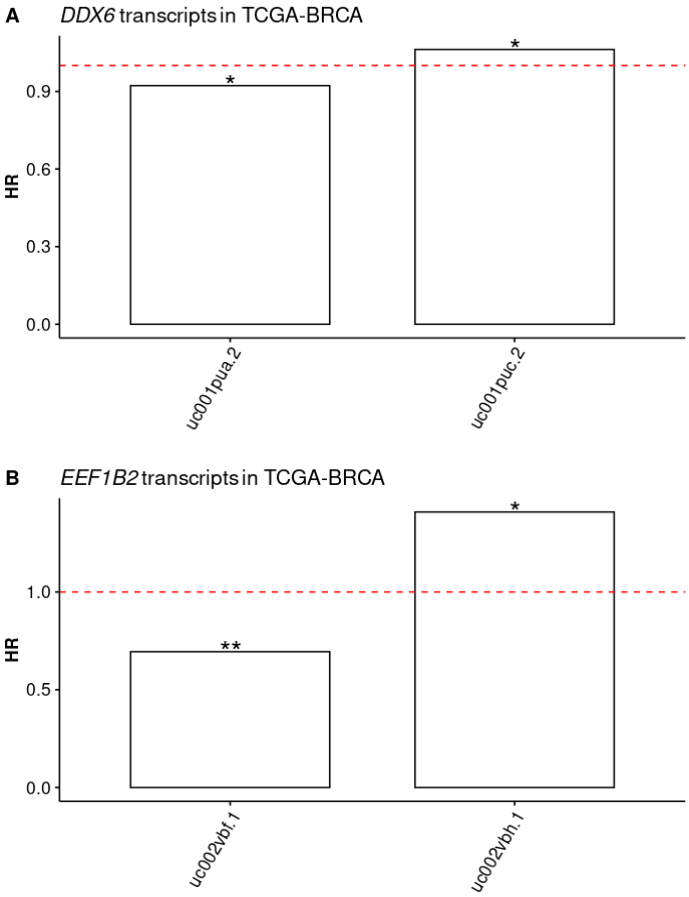

Figure 26: Genes with transcripts that were found to have opposite effects on survival. (A) *DDX6* had two significantly associated transcripts with survival in BRCA. (B) *EEF1B2* also had two significantly associated transcripts with survival in BRCA. Univariable significance levels for the isoforms were indicated on each bar with * for $p < 0.05$ and ** for $p < 0.01$.

3.4 Discussion

For the TCGA-STAD dataset, the gene *KBTBD2*'s transcript uc003tdb.2 was found to be the most hazardous transcriptome-wide (HR $\simeq$ 1.49, $p < 0.05$). The gene this transcript belongs to was found to cause insulin resistance in mice by Z. Zhang et al. (2016). The most protective transcript for STAD was identified as uc002lqr.1 (HR $\simeq$ 0.60, $p < 0.01$) belonging to gene *TMEM259*. This gene encodes for the Membralin protein, which protects motor neurons against endoplasmic reticulum (ER) stress (Yang et al., 2015).

The *TXNDC11* gene's uc002dbg.1 isoform was found to be the most hazardous (HR $\simeq$ 2.36, $p < 0.01$) in the GBM dataset univariably. This gene encodes for the thioredoxin domain containing protein 11, a redox regulator in DUOX protein folding and has not yet been associated with a disorder (Stelzer et al., 2016). The uc002vpc.2 transcript of the *AGFG1* gene was determined as the most protective (HR $\simeq$ 0.49, $p < 0.01$) in GBM. The gene encodes the Arf-GAP domain and FG repeat-containing protein 1, which is related to nucleoporins and is involved in nuclear export of Rev-directed RNAs (Bogerd et al., 1995).

When we look at the isoforms coming from the LIHC dataset, the *ANAPC7* gene's uc001tqo.2 transcript stands out with HR $\simeq$ 1.99 ($p < 0.0001$). The *ANAPC7* gene encodes for a component of the anaphase promoting complex which controls the cell cycle (Yu et al., 1998). On the other hand, the uc002ano.2 transcript of the *OAZ2* gene was determined as the most protective (HR $\simeq$ 0.54, $p < 0.001$) with respect to its expression levels. This gene encodes the ornithine decarboxylase antizyme 2, which has not been yet associated with a disorder according to the GeneCards database (https://www.genecards.org/, Stelzer et al. (2016)).

The uc003qqv.1 transcript of the *SNX9* gene was found to be the the most hazardous in TCGA-BRCA, with a transcriptome-wide HR $\simeq$ 1.71 ($p < 0.01$). The *SNX9* gene, or Sorting Nexin-9, is a member of the sorting nexin family that contain a phox (PX) domain. Members of this family of proteins are involved in intracellular trafficking, and have been emphasized as potential drug targets in cancer (Yang et al., 2020), and is also associated with Wiskott-Aldrich syndrome (Badour et al., 2007). The uc011bax.1 transcript of *PTPN23*, a non-receptor type protein-tyrosine phosphatase, had the lowest HR value ($\simeq$ 0.56, $p < 0.001$) in the univariable analysis with isoform expressions, thereby having the most beneficial effect on survival in breast cancer. The product of this gene is known to be involved in vesicular trafficking via sorting of ubiquitinated cargos and its variant was associated with neurodevelopmental delay and brain abnormalities (Bend et al., 2020).

Genes found with transcripts that had opposing effects on survival have been interesting results of this chapter. These were *CLK1*, *BICD2*, and *ESYT2* for the STAD dataset, *AKAP8L* for the LIHC dataset, and *DDX6* and *EEF1B2* for BRCA. For each of these genes, two transcripts with transcriptome-wide significance have been identified: one with HR > 1 and one with HR < 1. Experiments to understand the role of each alternative transcript inside the cell and the conditions under which each is transcribed from its gene could be carried out in future studies. This could make possible the promotion of cancer cells of each dataset to transcribe the protective transcript from its

respective gene under the specified conditions the gene prefers to transcribe the protective alternative.

As with the findings with exon expressions in Chapter 2, the isoforms identified with univariable transcriptome-wide analysis to be protective or hazardous in their respective datasets in this chapter, should be confirmed *in vivo* for understanding their exact role in cancer prognosis.

4 MULTIVARIABLE ANALYSES AND APPLICATION DEVELOPMENT WITH SHINY

4.1 INTRODUCTION

4.1.1 UNIVERSITY OF CALIFORNIA SANTA CRUZ XENA

The UCSC Xena platform (http://xena.ucsc.edu/) is a high performance online visualization tool that allows researchers the option to analyze their own data, as well as data from public cancer genomics resources such as TCGA and GDC (Goldman et al., 2020). Once Xena is launched, users initially select their choice of study and then the data type they want to analyze out of choices: “Genomic”, “Phenotypic” and for some studies “Analytic”. Users are asked to enter the input gene names once they have selected the “Genomic” radio button. The example screenshot in Figure 27 shows the Xena interface of the TCGA-PANCAN dataset when the Slit-Robo family members are selected. In the third panel on the right, additional “Phenotypic” “cancer type abbreviation” and “sample_type” choices have been selected.

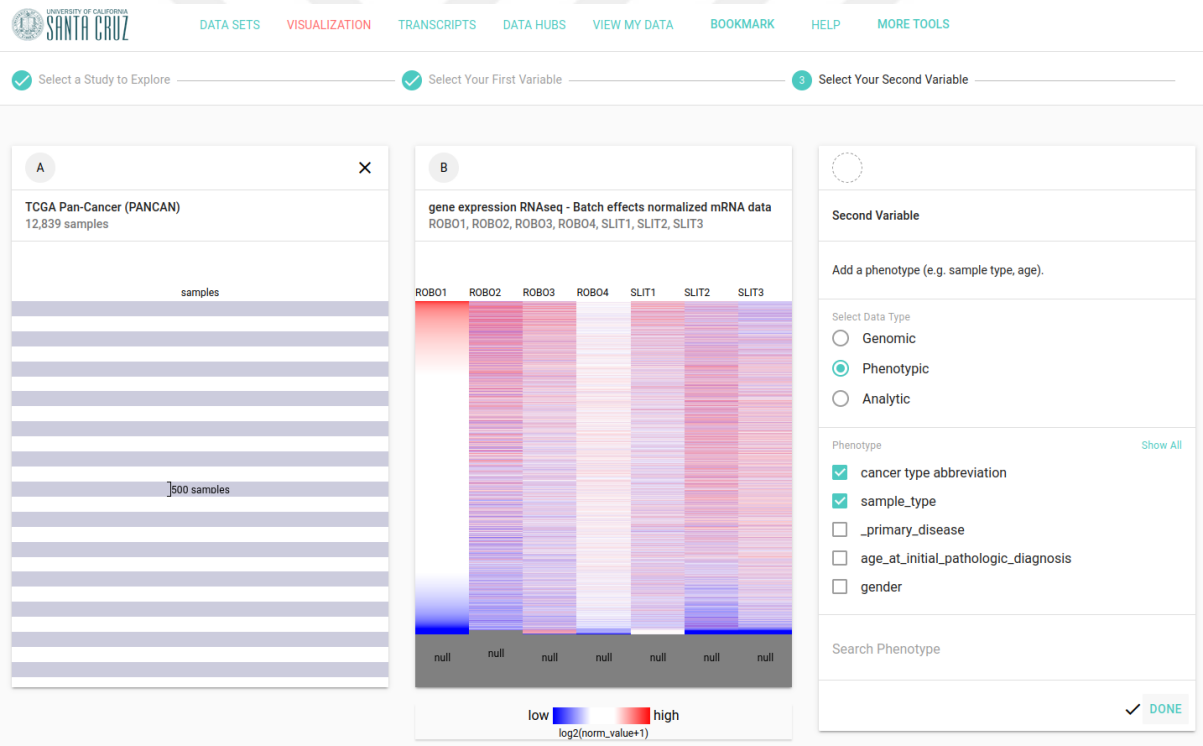

Figure 27: UCSC Xena user interface after the selection of genes. In Panel A on the left, the TCGA-PANCAN study was selected, indicating the total number of samples in the study, which is 12,839. In Panel B in the middle, the $\log_2(norm\_value + 1)$ expression values of the selected Slit-Robo genes are listed.

Users can then select and add phenotypic information columns from the third panel and click on “DONE”, after which they would be able to download the TSV file containing $\log_2(norm\_value + 1)$ batch effect normalized expression values of their selected genes and phenotypic information regarding the TCGA samples. Another feature

of Xena allows users to visualize the expressional distribution of their selected genes amongst TCGA-PANCAN datasets, as in Figure 28.

Figure 28: Screenshot from Xena depicting $\log_2(norm\_value + 1)$ distributions of each Slit-Robo gene across TCGA-PANCAN.

Xena has additional visualization and analysis tools for survival analysis, scatter plots, bar charts, gene set analysis, and spreadsheets; and it is possible to perform a wide variety of statistical tests on the platform. Xena makes it possible to download gene expression values as shown in Table 21, which can then later be analyzed in a user-dependent manner. However, Xena does not allow survival analysis of more than one gene at a time online. This is a significant limitation for gene sets, that act in coordination.

Table 21: Ten rows of a downloaded example Xena TSV file containing TCGA samples and their expression values of netrins, as well as TCGA study names and tumor type information. This column structure is required as input for the apps in this study.

sample	samples	NTN1	NTN3	NTN4	NTN5	NTNG1	NTNG2	cancer type abbreviation	sample_type
TCGA-DX-A7ER-01	TCGA-DX-A7ER-01	13.46	3.80	5.02	4.71	2.95	11.57	SARC	Primary Tumor
TCGA-09-0369-01	TCGA-09-0369-01	13.40	4.75	7.29	3.91	4.32	1.84	OV	Primary Tumor
TCGA-IE-A4EH-01	TCGA-IE-A4EH-01	13.34	1.01	6.29	1.60	6.26	3.27	SARC	Primary Tumor
TCGA-S9-A89V-01	TCGA-S9-A89V-01	13.33	4.99	13.19	4.44	9.21	7.94	LGG	Primary Tumor
TCGA-BR-8371-01	TCGA-BR-8371-01	13.32	0.82	10.39	2.88	4.83	5.26	STAD	Primary Tumor
TCGA-L5-A4OQ-11	TCGA-L5-A4OQ-11	13.31	1.16	10.53	2.27	7.54	5.68	ESCA	Solid Tissue Normal
TCGA-B5-A0K4-01	TCGA-B5-A0K4-01	13.30	2.71	8.06	3.12	0.93	5.03	UCEC	Primary Tumor
TCGA-KO-8414-01	TCGA-KO-8414-01	13.28	1.39	10.88	4.84	0.86	2.83	KICH	Primary Tumor
TCGA-A3-3328-01	TCGA-A3-3328-01	13.27	2.97	11.20	2.14	1.16	1.52	KIRC	Primary Tumor
TCGA-DU-6404-01	TCGA-DU-6404-01	13.26	5.80	10.20	2.70	9.10	11.63	LGG	Primary Tumor

4.1.2 BEST SUBSET SELECTION

Best subset selection methods using expressions of a set of genes are used regularly in identifying the most important genes or biomarkers in CPH models (Gentles et al., 2015; Yan et al., 2020). Best subsets of gene sets can be selected directly from the CPH model from genes with significant ($p < 0.05$) coefficients, or by using one of various algorithms for best subset selection that can be applied to Cox models. Such methods include the least absolute shrinkage and selection operator (lasso), ridge regression, and elastic net, which linearly combines lasso and ridge (Tibshirani, 1996; Zou, 2005). Both lasso and ridge aim to solve the function $\min_{\beta_0,\beta}\left\{\frac{1}{n}\|y-\beta_0-\mathrm{X}\beta\|_2^2\right\}$, but with different constraints: $\|\beta\|_1 \leq c$ for lasso and $\|\beta\|_2^2 \leq c$ for ridge, where c is a positive number.

While lasso can set coefficients to zero due to the square shape of the constraint region defined by the l^1-norm, ridge cannot because of the circle formed by the l^2-norm's constraint region. The elastic net method combines the lasso and ridge methods in $\min_{\beta\in\mathbb{R}^p}\{\|y-\mathrm{X}\beta\|_2^2+\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2\}$ which is equivalent to $\min_{\beta_0,\beta}\{\|y-\beta_0-\mathrm{X}\beta\|_2^2\}$ subject to constraint $(1-\alpha)\|\beta\|_1+\alpha\|\beta\|_2^2 \leq c$, where $\alpha=\frac{\lambda_2}{\lambda_1+\lambda_2}$. The elastic net can be adjusted through the parameter α, as the ridge constraint cancels when $\alpha = 1$ and the lasso constraint cancels when $\alpha = 0$. Each gene of the best subset, selected through lasso, ridge or elastic net regression is assigned a coefficient, which could then be used in prognostic index (PI) calculations.

4.1.3 PROGNOSTIC OUTCOME ANALYSIS

A PI can be determined for each patient of a disorder such as cancer using coefficients from a CPH model formed by combinations of factors that can be categorical such as sex, age, tumor grade therapy status (Haybittle et al., 1982) or continuous such as expression values of genes (Denkert et al., 2009). Once coefficients are determined for a best subset of genes selected with one of the methods explained in the previous section, Equation 4 can be used to calculate a PI value for each sample in a selected dataset with the expression values of the genes:

$$\sum_{i=1}^{n} gene_i \times coeff_{gene} \tag{4}$$

as in Xue et al. (2019), where $[gene]_i$ is the expression of an input gene for the i^{th} sample in the selected cancer dataset, $coeff_{gene}$ is the coefficient of the gene from the best subset selection method and n is the sample size of the dataset. Afterwards, usually the median PI value of the cancer dataset is used to label samples with a PI value greater than the median PI as high-risk, while remaining samples are labeled low-risk.

While PI analyses have been frequently used in different cancer datasets to identify factors or GSs important for that cancer, analyses of axon guidance ligand-receptor families in this regard have remained relatively unexplored, forming the main motivation of this chapter. The cancer dataset with samples labeled either high or low-risk can be used to compare cancer prognostic patterns in these two groups based on the genes

(covariates) whose coefficients are involved, using K-M plots explained in Section 1.1.1. The magnitude of the difference between the two prognosis groups can be assessed by a low:high risk ratio using a univariable survival model with the PI values determined for the dataset.

4.1.4 CLUSTERING, BICLUSTERING AND CLUSTER COMPARISON

Cluster analysis or clustering is an unsupervised learning process whereby a set of objects are grouped in a way that the objects in the same group form what is called a cluster (Madigan, 2013). Objects in a cluster are more like each other than to objects in other groups. The clustering exploratory data analysis task has applications in a variety of fields including graph theory, pattern recognition, image analysis and in particular bioinformatics; where it is mainly used for the functional annotation, tissue classification or motif identification of microarray or RNA-Seq gene expression data (Karatzas et al., 2021; Karim et al., 2021).

Clustering is also used in elucidating protein-protein interaction networks (Pizzuti & Rombo, 2014). Connectivity-based cluster models such as hierarchical clustering can be top-down (divisive) or bottom-up (agglomerative) and are based on distance connectivity (Karim et al., 2021). The results for hierarchical clustering are usually represented as a dendrogram. Commonly used distance metrics for hierarchical clustering include the Euclidean distance, represented by: $\|x-y\|_2 = \sqrt{\sum_{i=1}(x_i+y_i)^2}$, the Manhattan distance, represented by: $\|x-y\|_1 = \sum_{i=1}|x_i-y_i|$ and the maximum distance, represented by: $\|x-y\|_\infty = \max_{i=1}|x_i-y_i|$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are vectors in an n-dimensional real vector space. Other distance metrics include the Canberra distance: $\sum_{i=1}^{n}\frac{|x_i-y_i|}{|x_i|+|y_i|}$, the Minkowski distance: $(\sum_{i=1}^{n}|x_i-y_i|^p)^{\frac{1}{p}}$ and the binary (Hamming) distance which represents the number of different entries in $\mathbf{x}$ and $\mathbf{y}$ (Ehsani & Drablos, 2020).

Linkage criteria in hierarchical clustering use the pairwise distance function between observations to determine the distance between sets of observations. Between two sets of observations X and Y, commonly used linkage criteria include complete (maximum) linkage: $\max\{d(x,y) : x \in X, y \in Y\}$ and single (minimum) linkage: $\min\{d(x,y) : x \in X, y \in Y\}$ (Valentini, 2009). There is also the bottom-up unweighted average linkage clustering or unweighted pair group method with arithmetic mean (UPGMA, see Datta and Datta (2003) for the method's description) represented by $\frac{1}{|X|\cdot|Y|}\sum_{x\in X}\sum_{y\in Y} d(x,y)$. In the bottom-up weighted average linkage clustering or weighted pair group method with arithmetic mean (WPGMA) hierarchical clustering, the nearest two clusters i and j are combined into a higher-level cluster $i \cup j$ at each step; after which the distance between this union cluster and another cluster k becomes $\frac{d_{i,k}+d_{j,k}}{2}$, the arithmetic mean of the distances between members of k and i and k and j. Ward's method is an agglomerative

hierarchical clustering method with a similar principal to analysis of variance (ANOVA) that aims to minimize variance through the error sum of squares (Valentini, 2009).

In partitional or centroid-based cluster models, each cluster is represented by its arithmetic mean (centroid) such as in *k*-means clustering; in which a pre-determined *k* number of clusters are used to assign objects to the cluster with the minimized distance to the mean of the cluster (also center or centroid of the cluster) to minimize within-cluster variance (Valentini, 2009). Other clustering models include distribution-based models that for instance rely on the expectation-maximization (EM) algorithm's multivariable normal distributions (Xing et al., 2006). There are also density-based (Mehmood et al., 2017), graph-based (Xu et al., 2002) and neural clustering models (Herrero et al., 2001).

Biclustering or co-clustering is the name of a collection of data mining techniques used to simultaneously cluster the rows and columns of a matrix which allows for discovery of similarity based homogeneous subsets or submatrices known as biclusters of the matrix. Biclustering was first applied to gene expression data by Cheng and Church (2000) using their greedy CC algorithm in an effort to overcome limitations in earlier clustering algorithms where each gene or experimental condition is assigned to only one cluster and each must be assigned to a cluster, which may fall short of reflecting the actual biological process.

The Bimax algorithm applies inclusion-maximal biclustering; whereby change with respect to the control experiment for a gene is represented by 1 and no change is represented by 0. The algorithm then looks for submatrices in which all elements are equal to 1 (Prelic et al., 2006). The plaid algorithm employs a plaid model whereby a general background layer of all genes and samples is initially fitted and then bicluster-specific layers exhibiting patterns different from the general model are added (Turner et al., 2005). The Xmotifs algorithm (Murali & Kasif, 2003) searches sample rows for conserved gene expression motifs across columns for identifying biclusters.

Comparing clusters or cluster sets could be desired for certain research questions, where relevant metrics apply for each method. Please refer to Karatzas et al. (2021) for an overview of different cluster comparison methods. Clustering, biclustering and cluster comparison methods, which have traditionally been applied to gene expression levels, were applied to $\log_2$HR values in the ClusterHR application (app) developed in this study.

4.1.5 MicroRNAs

A microRNA, or shortly miRNA, is a small (~19-25 nucleotides in length), noncoding RNA molecule usually involved in post-transcriptional repression. According to the miRNA database miRBase (http://www.mirbase.org/), there are currently 38,589 hairpin precursors and 48,860 mature miRNAs cataloged in the human genome (Kozomara et al., 2019). After being exported into the cytoplasm from the nucleus, miRNAs are further processed by the Dicer enzyme to be incorporated into the RNA-induced silencing complex (RISC), which then mediates gene silencing through mRNA degradation or translational repression (Tijsterman & Plasterk, 2004). MicroRNAs are known to be

heavily dysregulated in cancer cells. For example, miR-143 and miR-145 are known to be deleted in lung cancers (Peng & Croce, 2016), while the miR-17/92 cluster is upregulated in hepatocellular carcinoma and osteosarcoma (Mogilyansky & Rigoutsos, 2013). Furthermore, miR-424 was found to promote angiogenesis in endothelial cells in vitro under hypoxic conditions (Peng & Croce, 2016).

While SurvMicro allows for multivariable survival analysis and best subset selection with miRNAs, SmulTCan is the only existing tool allowing interactive multivariable analyses with input miRNAs; that also performs downstream PI analyses with the selected best subset from the input miRNA set (Ozhan et al., 2021).

4.1.6 AXON GUIDANCE MOLECULES IN CANCER

Axon guidance involves the coordinated action of different types of molecular cues; both local and short-range, which can either be attractive or repellent. It is the process during neural development whereby soluble or membrane-associated cues guide extending axons. The canonical axon guidance cues of the nervous system include the ligand families netrins, semaphorins (Semas), Slits, and ephrins (Squire, 2013). Semas can be either secreted or one of the four types of GPI-linked transmembrane proteins. There are ~20 Semas in mammals; receptors of Semas include plexins of classes A-D and neuropilins (Alto & Terman, 2017). Plexins and neuropilins form complexes in semaphorin (Sema) signaling (Figures Figure **29**-Figure **30**). Ephrin receptors are members of the Eph family; where the GPI-anchored Ephrin-A binds EphA and the transmembrane Ephrin-B binds EphB (Squire, 2013).

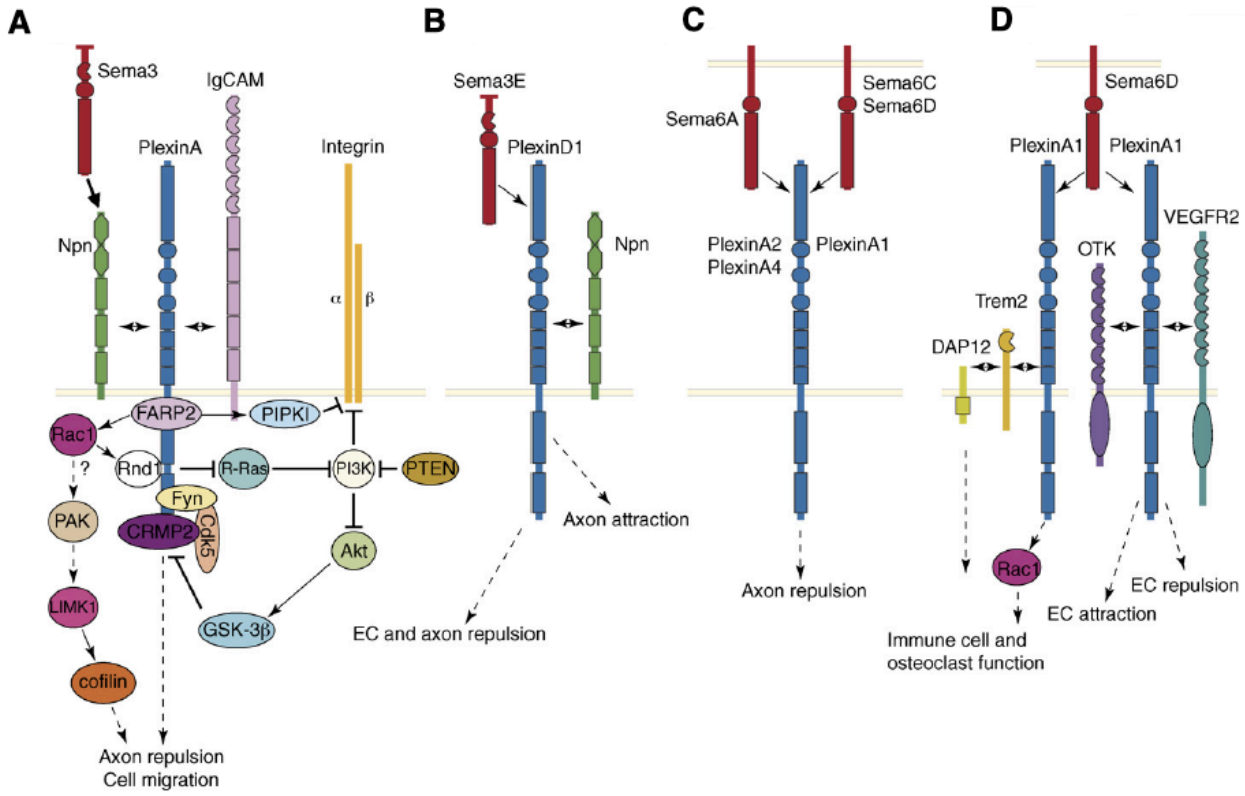

Figure 29: Diagram of Sema signaling, where plexinAs (blue) receive signals from Sema3s and Sema6s (red). Sema3s also require neuropilins (green) to bind plexins. (A) Sema3-neuropilin-plexin complex requires IgCAMs for axon guidance. This complex also regulates actin and microtubule dynamics, as well as endocytosis. (B) When Sema3E binds directly to PlexinD1, it results in endothelial cell and axon repulsion. Involvement of neuropilin in this interaction results in axon attraction. (C) Sema6s bind plexinAs during neural development, eliciting axon repulsion. (D) Sema6D can bind plexinA1 to elicit endothelial cell attraction through OTK interaction and repulsion through VEGFR2 interaction. Sema6D can also regulate immune cell and osteoclast function through Trem2. Figure taken from Zhou et al. (2008), see Permissions at the end of this document.

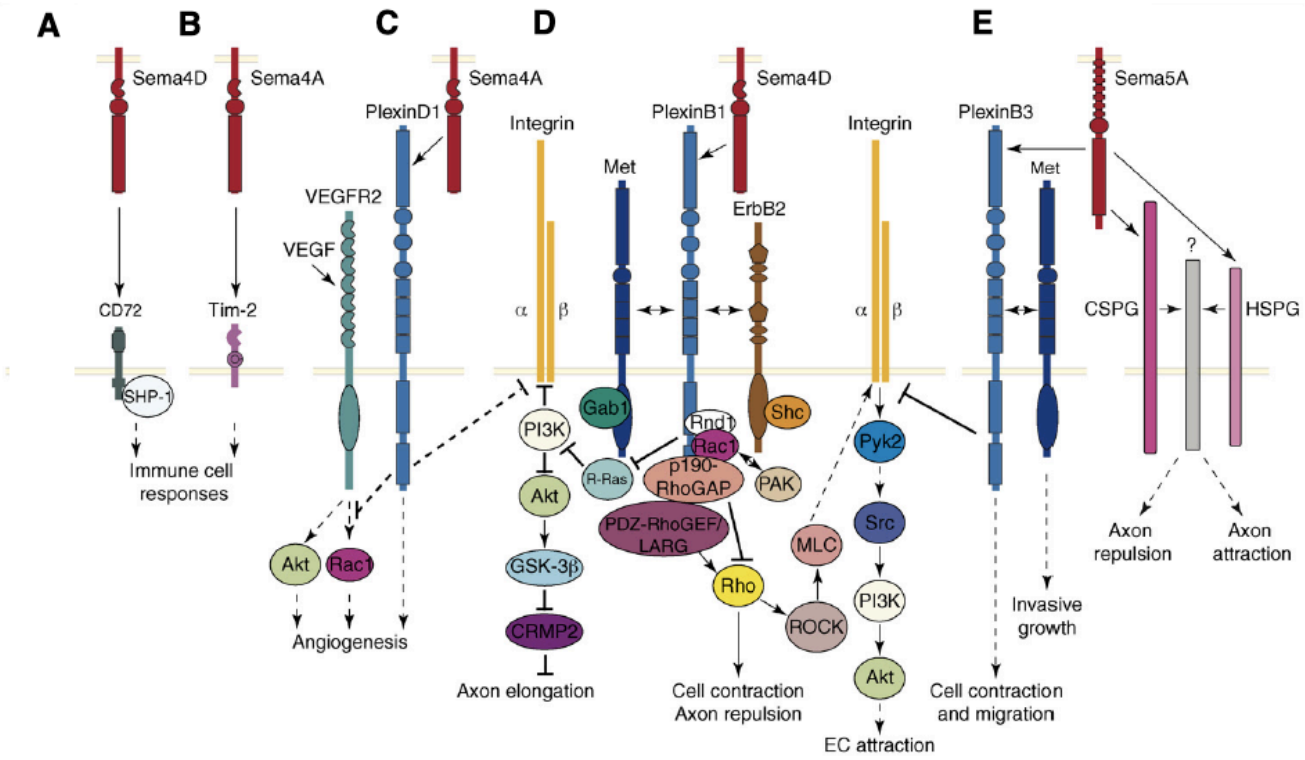

Figure 30: Diagram of signaling through Sema4s and Sema5s (red) through common and distinct receptors. (A, B) Sema4D binds to CD72 and Sema4A binds to Tim-2 receptors located on immune cells. (C) Sema4A binding to PlexinD1 inhibits angiogenesis. (D) The Sema4D-PlexinB1 complex mediates cell migration and invasive growth through interaction with Met and cell migration and growth cone collapse through interaction with ErbB2. (E) Sema5A binding to plexinB3 promotes cell migration through repulsion, while invasive growth is regulated through Met. Axon guidance through Sema5A requires proteoglycans HSPG and CSPG. Figure taken from Zhou et al. (2008), see Permissions at the end of this document.

Humans contain classes 3-7 of Semas, that include the Sema3s *SEMA3A*, *SEMA3B*, *SEMA3C*, *SEMA3D*, *SEMA3E*, *SEMA3F*, and *SEMA3G*; Sema4s *SEMA4A*, *SEMA4B*, *SEMA4C*, *SEMA4D*, *SEMA4F*, and *SEMA4G*; Sema5s *SEMA5A*, *SEMA5B*; Sema6s *SEMA6A*, *SEMA6B*, *SEMA6C*, and *SEMA6D*, and *SEMA7*. Class A plexins are *PLXNA1*, *PLXNA2*, *PLXNA3*, *PLXNA4* and Class B plexins are *PLXNB1*, *PLXNB2* and *PLXNB3* (Alto & Terman, 2017). Additional plexin members include the Class C plexin *PLXNC1* and Class D plexin *PLXND1*. There are two forms of the transmembrane glycoproteins neuropilins: *NRP1* and *NRP2*. In addition to axon guidance, Sema signaling influences a variety of biological processes including organ morphogenesis and homeostasis, as well as cell migration and cytokine release (Alto & Terman, 2017). Caused by to the pleiotropic nature of the Sema signaling pathway's members, variants of genes in this pathway were found to have roles in several disorders including schizophrenia, neurodegenerative disease, and cancer (Zhou et al., 2008).

The Slit-Robo signaling pathway mediated axon repulsion in the nervous system, though it is now known to be involved in a variety of cellular functions from cellular proliferation, cell migration, vascularization, and organ development (Tong et al., 2019).

There are four Roundabout or Robo receptors: *ROBO1*, *ROBO2*, *ROBO3*, and *ROBO4*; and three Slit ligands: *SLIT1*, *SLIT2* and *SLIT3*. Robos are members of the immunoglobulin (Ig) superfamily; the Ig domains are indicated with blue semi-circles in Figure 31 (Tong et al., 2019). Slits are large, secreted proteins found to be involved in the repulsion and branching of axons of the mammalian forebrain and the midline of the spinal cord, respectively. The epidermal growth factor (EGF)-like repeats on Slits are indicated with blue circles in Figure 31.

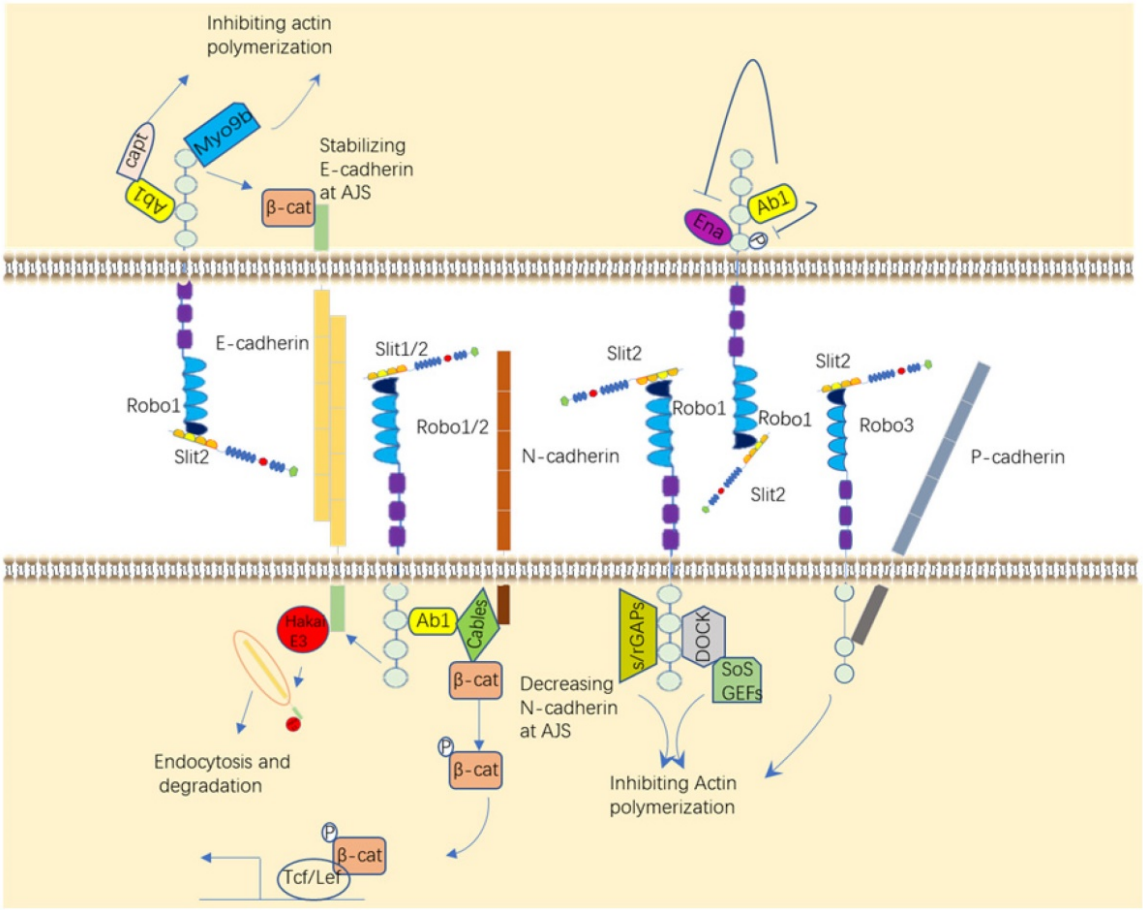

Figure 31: Slit-Robo signaling pathway and downstream targets. Slits also contain a Cystein-rich C-terminal domain shown as a green circle. Fibronectin type-III (FN3) domains on Robos are indicated with purple rectangles. Figure taken from Tong et al. (2019) , see Permissions at the end of this document.

Netrins are neuroguiding factors with receptors from the DCC and UNC-5 families, initially identified in studies with commissural axons in the spinal cord in vertebrates. They were shown to be secreted from floor plate cells and show a dorsal-to-ventral increasing gradient in the spinal cord (Squire, 2013). Netrin attraction is dependent on deleted in colorectal cancer (*DCC*), and repulsion is mediated by uncoordinated 5 (UNC-5) receptors, while combinations of these receptors and downstream signaling molecules also affect bifunctionality (Finci et al., 2014). Netrins include the secreted *NTN1*, *NTN2*,

NTN3, *NTN4*, and *NTN5*, as well as the membrane bound *NTNG1* and *NTNG2*. Like Slits, netrins contain EGF-like repeats, but they also contain N-terminal laminin domains (except for *NTN5*) and C-terminal netrin-like (NTR) domains (Meijers et al., 2020). Netrin receptors include the UNC-5 family: *UNC5A*, *UNC5B*, *UNC5C*, and *UNC5D*; as well as receptors *DCC* and Down syndrome cell adhesion molecule (*DSCAM*), which like Robos belong to the Ig superfamily (Rajasekharan & Kennedy, 2009).

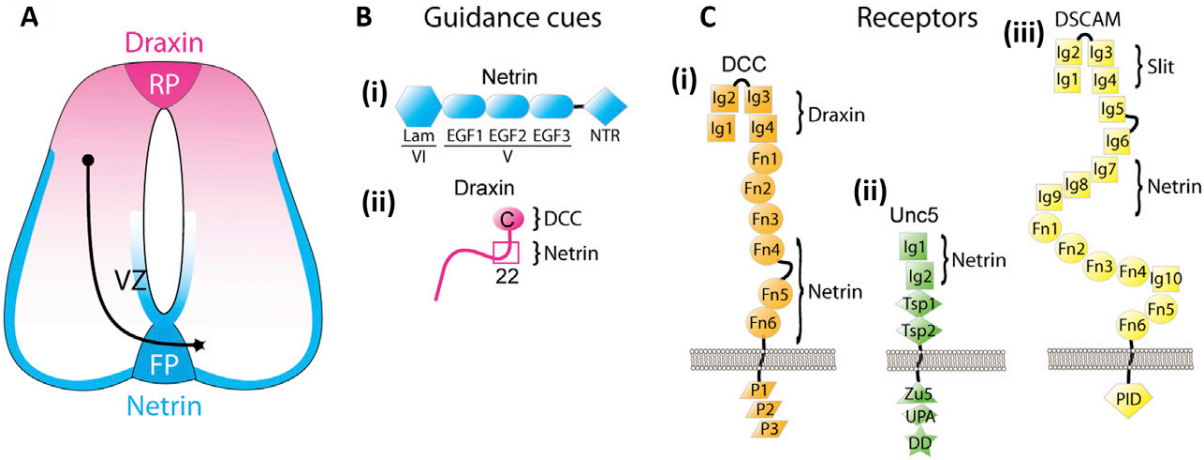

Figure 32: (A) Cross-sectional view of the early embryonic neural tube, from which the spinal cord and brain are developed. The ventricular zone (VZ) and floor plate (FP) express netrin (blue) and the roof plate (RP) expresses draxin (red). (B) Structural domains of the guidance cues netrin and draxin. (C) Binding sites for netrin, draxin and slit are indicated on each receptor. Figure taken from Meijers et al. (2020) , see Permissions at the end of this document.

Axon guidance molecules' roles in cancer progression have become the focus of research, as several types of neural guidance cues are implicated in various cancers. In particular, the Sema signaling pathway is known to be involved in several aspects of tumor development; with different members found to be involved in anti-angiogenesis, such as Semas 3A and 3B and Semas 3E, 3F and 3G; and pro-angiogenesis; such as *SEMA4D* (Sakurai et al., 2012).

A recent TCGA-PANCAN study with Sema3 expressions by Zhang et al. (2020) associated Semas 3A and 3E with poor survival and Sema 3G with survival advantage; further validated Sema pathway members as promising therapeutic targets in cancer. Neuropilins are known to enhance ligand-VEGF receptor binding and blocking *NRP2* function was found the inhibit lung metastasis (Caunt et al., 2008). However, Semas and their receptors have not been studied as whole family with respect to their role in survival using TCGA datasets.

Different Slit-Robo members have been associated with several types of cancer including ovarian and breast cancer (Rezniczek et al., 2019). In previous studies with mice, *ROBO4* was found to have an inhibitory role in angiogenesis (F. Zhang et al., 2016). Additionally, *SLIT2-ROBO1* were found to contribute to the Warburg effect, a

fermentative metabolic environment found in cancer cells and to activate the SRC/ERK/c-MYC/PFKFB2 pathway in osteosarcoma (Zhao et al., 2018).

The netrin family of secreted proteins, with contradictory roles in both tumor proliferation and inhibition of tumor development, was investigated in a recent study for their expression, mutation, and methylation profiles across TCGA-PANCAN. Several netrins including *NTNG1* and *NTNG2* were identified as putative diagnostic biomarkers for endocrine tumors (Hao et al., 2020).

Gene sets that were used to demonstrate the apps in this chapter include the seven Slit-Robo family members that make up the Slit-Robo gene set and the thirty-one Sema, plexin and neuropilin genes that make up the Semas-receptors gene set. Additionally, a gene set made up of twelve genes for the netrins and their receptors (netrins-receptors), has been used in the analyses as input. The reason for their selection includes them being ligand-receptor complexes and complex molecules with divergent interaction patterns, and the lack of studies in the literature that examine them as gene sets with respect to cancer.

4.1.7 TCGA DATASETS, SAMPLE SIZES AND SURVIVAL TYPES

As mentioned earlier in Section 1.2, TCGA contains survival and clinical data across 33 tumor types. TCGA's survival data are categorized into four different analysis types: namely, overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progress-free interval (PFI), each representing different forms of survival. The structure of a survival TXT file from GDC is shown in Table 22 for ten TCGA samples from the LIHC dataset, which includes information for the four different survival types.

OS is defined by the length of time that the patient lives after diagnosis or the start of treatment of cancer. DSS is the survival rate beginning at the time of diagnosis or the start of treatment, during which the patient has not died due to any other cause than the cancer itself. DFI, or relapse-free interval/survival, refers to the length of time after which a patient's primary treatment for a cancer has ended and the patient lives without any signs or symptoms of that cancer. PFI refers to the length of time after the patient has received treatment for cancer and lives with the tumor, without the worsening of their condition (Liu et al., 2018).

Table 22: The column format of a survival TXT file from GDC. The table shows ten TCGA-LIHC samples and their survival information, including event and time-to-event, for four different types of survival.

sample	_PATIENT	OS	OS.time	DSS	DSS.time	DFI	DFI.time	PFI	PFI.time	Redaction
TCGA-2Y-A9H5-01	TCGA-2Y-A9H5	1	555	1	555	1	258	1	258	NA
TCGA-2Y-A9H6-01	TCGA-2Y-A9H6	0	357	0	357	0	357	0	357	NA
TCGA-2Y-A9H7-01	TCGA-2Y-A9H7	0	1168	0	1168	1	1117	1	1117	NA
TCGA-2Y-A9H8-01	TCGA-2Y-A9H8	1	633	1	633	1	398	1	398	NA
TCGA-2Y-A9H9-01	TCGA-2Y-A9H9	0	697	0	697	1	59	1	59	NA
TCGA-2Y-A9HA-01	TCGA-2Y-A9HA	1	36	1	36	NA	NA	1	36	NA
TCGA-2Y-A9HB-01	TCGA-2Y-A9HB	0	260	0	260	0	260	0	260	NA
TCGA-3K-AAZ8-01	TCGA-3K-AAZ8	0	396	0	396	NA	NA	1	151	NA
TCGA-4R-AA8I-01	TCGA-4R-AA8I	1	262	1	262	1	158	1	158	NA
TCGA-5C-A9VG-01	TCGA-5C-A9VG	0	328	0	328	0	328	0	328	NA

It is important to study survival across different cancer types that contain adequate number of samples. TCGA provides this great opportunity to do so. To demonstrate why TCGA has been chosen as the main dataset resource, I have included the number of samples for each of the TCGA-PANCAN datasets incorporated into the applications (apps) in this study are given in Table 23. Each sample can have information for more than one type of survival file; while it should be noted that not all samples in the below table belong to primary tumors. Computations in the apps of this study are carried out with only primary tumor (and primary blood derived for TCGA-LAML) samples for the selected TCGA datasets, therefore the actual number of samples in generating the results in the apps is usually less than those shown in Table 23. The sample size depends also on the presence of expression values of the input gene(s) in the primary tumor samples in the respective TCGA datasets.

Table 23: TCGA-PANCAN total sample sizes for each survival type incorporated in the Shiny applications.

Dataset	Cancer	OS	DSS	DFI	PFI
LIHC	Liver hepatocellular carcinoma	437	424	371	437
STAD	Stomach adenocarcinoma	504	463	282	506
LUAD	Lung adenocarcinoma	632	593	381	632
GBM	Glioblastoma multiforme	602	561	3	602
BRCA	Breast invasive carcinoma	1235	1205	1054	1235
COAD	Colon adenocarcinoma	543	527	221	543
ESCA	Esophageal carcinoma	204	200	99	204
LUSC	Lung squamous cell carcinoma	498	446	303	499
ACC	Adrenocortical carcinoma	92	90	53	92
BLCA	Bladder urothelial carcinoma	435	417	195	436
LGG	Brain lower grade glioma	528	520	136	528
CESC	Cervical squamous cell carcinoma and endocervical carcinoma	312	308	178	312
CHOL	Cholangiocarcinoma	45	43	32	45
HNSC	Head and neck squamous cell carcinoma	603	573	142	603
KICH	Kidney chromophobe	89	89	42	89
KIRC	Kidney renal clear cell carcinoma	944	923	165	940
KIRP	Kidney renal papillary cell carcinoma	351	347	197	349
LAML	Acute myeloid leukemia	186	0	0	0
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	48	48	28	48
MESO	Mesothelioma	86	66	15	84
OV	Ovarian serous cystadenocarcinoma	599	561	299	599
PAAD	Pancreatic adenocarcinoma	196	188	72	196
PCPG	Pheochromocytoma and paraganglioma	187	187	164	187
PRAD	Prostate adenocarcinoma	566	564	396	566
READ	Rectum adenocarcinoma	183	177	51	183
SARC	Sarcoma	271	265	160	271
SKCM	Skin cutaneous melanoma	464	458	0	465
TGCT	Testicular germ cell tumors	139	139	110	139
THYM	Thymoma	125	125	0	125
THCA	Thyroid carcinoma	580	574	408	580

UCS	Uterine carcinosarcoma	57	55	27	57
UCEC	Uterine corpus endometrial carcinoma	582	580	452	582
UVM	Uveal melanoma	80	80	0	79

4.1.8 SHINY

Shiny (Chang et al., 2021) is an R package that helps in creating and publishing interactive web apps from R. App development with Shiny is made easier for newly introduced developers through sample apps in https://shiny.rstudio.com/gallery/ and the web forum at https://community.rstudio.com/ can be used for support from more experienced users. Shiny apps can be designed to include additional CSS and HTML functionality.

In recent years, Shiny apps have become prevalent as environments of database sharing and analysis due to the need in the research community for open-source and freely available web-based tools. Their relative simplicity of development and user-friendly interface make Shiny apps favorable in bioinformatics research. Recent examples of Shiny apps can be given from Quickomics for multidimensional statistical analysis of omics data (Gao et al., 2021) and the IPDmada tool for meta-analyses from individual patient data (Wang et al., 2021). Metabolite-Investigator is another web-based, open-source Shiny app which incorporates metabolomics data with tools for data pre-processing user-input disease datasets for metabolite identification and association (Beuchel et al., 2021).

Deployment of a Shiny app to the web can be done either through Shinyapps.io to the cloud, on-premises through the Shiny Server or on-premises commercially through RStudio Server. The applications developed in this chapter, SmulTCan for the multivariable survival analysis of TCGA data with gene sets and ClusterHR for the clustering-based joint analysis of HRs resulting from multivariable survival analysis of TCGA-PANCAN are based on R Shiny.

4.1.9 AIMS

- Develop and deploy the Shiny apps SmulTCan and ClusterHR, which focus on gene set HR analysis for a given cancer dataset, and clustering analysis of HR values (univariable and multivariable) for all or selected cancer datasets, respectively.
- Analyze TCGA-PANCAN for HR characteristics and signatures of three axon guidance gene sets: Slit-Robo, Semas-receptors and netrins-receptors; with their expression values obtained from Xena, using both Shiny apps created in this study.

4.2 METHODS

4.2.1 DATA ACQUISITION

The Slit-Robo TCGA-PANCAN TSV file was obtained from Xena for the four Robos and the three Slits with phenotypic information columns "cancer type abbreviation" and "sample_type" included. This file contains RNA-Seq $\log_2(norm\_value + 1)$ expression values of tissue samples across the thirty-three TCGA datasets for the selected genes. Similarly, the TCGA-PANCAN expression TSV file for the twelve netrins and receptor genes was downloaded from Xena and this file contains $\log_2(norm\_value + 1)$ expression columns for the six netrins and their six receptors. In both ClusterHR and SmulTCan apps, survival TXT files downloaded using the "UCSCXenaTools" package (Wang, 2019) from GDC are embedded in the app inside individual folders for each of the thirty three TCGA datasets.

4.2.2 DATA ANALYSIS AND INTERPRETATION

As in the univariable survival analyses of Chapters 2 and 3, the coxph() function of "survival" is used with its default parameters in the apps for calculating multivariable HRs of input gene sets. In the SmulTCan app, the "survminer" package (Kassambara et al., 2020) is used for the K-M, forest and Schoenfeld plots using functions ggsurvplot(), ggforest() and ggcoxzph(), respectively. The ggcoxzph() function requires the coxph() output to be used inside the cox.zph() function. For the percent area under the receiver operating characteristic (ROC) curve (AUC%) calculations, the Score.list() function of "riskRegression" (Gerds & Ozenne, 2020) is used with the "plots" parameter set to "ROC" and metrics parameter set to "AUC", while plots are drawn with the plotROC() function of the same package.

The output of the cph() function of "rms" (Harrell Jr, 2021) with the parameter "surv" set to TRUE, which extends the CPH model formula, was used inside anova() for plotting and ranking input genes based on their p and $\chi^2 - df$ values in the "ANOVA" tab of SmulTCan. CPH models in SmulTCan are validated with validate() of "rms" in the "Validation" tab. Best subset selection using the CPH model in SmulTCan is implemented by selecting genes whose coefficients have $p < 0.05$. The "Data", "Boxplot", "Forest plot", "Schoenfeld plots", "ROC", "ANOVA", "Validation" and "CPH" tabs are reactively connected to the main CPH model built from embedded TCGA data for the selected survival type with the default parameters of coxph().

The "glmnet" package (Simon et al., 2011) is used in SmulTCan to visualize the best subset of the input gene set selected with the elastic net method; whereby users can adjust the slider for the "alpha" parameter set to 1 by default, which corresponds to lasso; down to 0, which corresponds to ridge regression. The cv.glmnet() function's "family" parameter is globally set to "cox" for the computations. This function of "glmnet" uses 10-fold cross validation by default, but the user can also set the "Folds" parameter to "sample_size" from this main tab. The "caret" package's (Kuhn, 2021) createFolds() function is implemented for the "glmnet" tab's default "Folds" input. To prevent

variability in the results for the sub-tabs of the “glmnet” tab, the seed is set to 123 in the built-in function inside global.R for the “foldid” parameter of cv.glmnet().

The “BeSS” package (Wen et al., 2020) is included in another main tab, which uses the bess() function also with the “family” parameter set to “cox”. Reactivity with the embedded data is stopped in the “glmnet” and “BeSS” tabs using isolate() of “shiny”, and these main tabs build their own CPH models. The Breslow approximation method is used for handling ties in “glmnet”, while “BeSS” uses the default Efron approximation of coxph() for its ridge regression algorithm. The coxph() function is also used in computing Low:High risk ratios of the best subsets.

The ClusterHR app uses the Bioconductor packages “gtools” and “heatmaply” (Galili et al., 2018) for the heatmaps. When displaying the heatmap outputs, users may choose from different distance and hierarchical clustering methods and have the option of clustering based only on genes, only on cancers or both. Distance methods incorporated into the app through the heatmaply() function’s “dist_method” parameter are: “euclidean”, “maximum”, “manhattan”, “canberra”, “binary”, and “minkowski”. Through the “hclust_method” parameter of heatmaply(), users can choose from clustering linkage methods “complete”, “ward.D” (Ward’s method), “ward.D2” (squared Ward’s method), “single”, “average”, “mcquitty”, “centroid”, and “median”. Average linkage refers to UPGMA, McQuitty method refers to WPGMA, median refers to WPGMC, and centroid refers to UPGMC.

For biclustering, the currently included methods in ClusterHR are “BCBimax” (refers to Bimax biclustering), “BCCC” (refers to Cheng and Church’s algorithm), “BCPlaid” (refers to Plaid model biclustering), “BCQuest” (refers to Questmotif biclustering), and “BCXmotifs” (refers to Xmotifs biclustering). These biclustering methods have been incorporated into ClusterHR through the “biclust” package (Kaiser et al., 2021).

Biclusters with an equal number of cancer datasets can be compared based on their row names (datasets) using the “mclustcomp” package (You, 2021). Currently, the available comparison methods in ClusterHR include “adjrand” (refers to the adjusted Rand index), “chisq” (refers to the Pearson coefficient), “fmi” (refers to the Fowlkes-Mallows index), “jaccard” (refers to the Jaccard index), “mirkin” (refers to the Mirkin metric), “overlap” (refers to the overlap coefficient), “pd” (refers to the partition difference), “rand” (refers to the Rand index), “sdc” (refers to the Sørensen–Dice coefficient), “smc” (refers to the simple matching coefficient), “tanimoto” (refers to the Tanimoto index), “tversky” (refers to the Tversky index), “wallace1” (refers to the Type I Wallace criterion), and “wallace2” (refers to the Type II Wallace criterion).

These listed methods are from the counting pairs category of comparison methods of the “mclustcomp” package (You, 2021), please refer to the package’s manual for details of the other categories of available methods. All analyses in ClusterHR and SmulTCan can be carried out both univariably for each gene in the gene set against the transcriptome in each cancer dataset or in a multivariably; which can be used to analyze the comparative effects of the genes with respect to each other within the gene set in each cancer dataset. When the input gene set is formed from the members of a molecular pathway, these apps can provide insight into the functional mechanism of this pathway in the disorder.

4.2.3 WARNINGS

Both the SmulTCan and ClusterHR apps check that there is survival information for the selected dataset relevant to the selected survival tab, that the Cox model for the selected dataset with the input genes converges for the selected survival type, that all coefficients of the Cox model are finite, and that there are expression values of the input gene(s) for the selected dataset. A warning is displayed in the SmulTCan app if one of the conditions above is unmet for the selected dataset, survival type and input gene combinations. In those cases, ClusterHR app incorporates *NA* values for that dataset's rows in the multivariable HR matrix, thus the dataset does not appear in the output plots for the multivariable analyses. Individual *NA* entries are added to the results matrices for the univariate analyses if a condition declared above is unmet in the ClusterHR app. These warnings incorporated into the apps are intended to suppress potential coxph() warnings that might in turn affect Cox regression coefficient calculations and may result in wrongful interpretations and predictions.

4.2.4 DOWNLOADS

The SmulTCan app allows download formats TSV for data outputs and PDF and high-resolution PNG for plot outputs, with the name of the selected TCGA dataset and survival analysis type incorporated into the filename. The heatmaps of ClusterHR are downloadable owing to the "heatmaply" package's functionality.

4.3 RESULTS

4.3.1 SMULTCAN SHINY APPLICATION

In this section, SmulTCan Shiny app's architecture, interface, and plots generated from the different tabs of the app, which is available from http://konulabapps.bilkent.edu.tr:3838/SmulTCan/, are presented for selected datasets with results from selected axon guidance ligand-receptor family TSV file inputs.

4.3.1.1 ARCHITECTURE

The app architecture of SmulTCan is shown in the diagram of Figure 33, composed of the two main parts: model analysis and best subset selection. The model analysis part is made up of smaller modules that can be used for visualizing the data inside the CPH model built with the coxph() function, diagnostics of this CPH model, and preparing for prediction using this CPH model. Main tabs in the app are indicated inside olive green boxes in Figure 33, within the module circles they belong to. The app's functionality is centered around the TCGA data embedded inside it, to which the model analysis part's modules and the "CPH" main tab of the best subset selection part are reactively connected. Reactivity is stopped in the "BeSS" and "glmnet" main tabs of the best subset selection parts, which build their internal CPH models based on the embedded TCGA data and the user-input file from UCSC Xena.

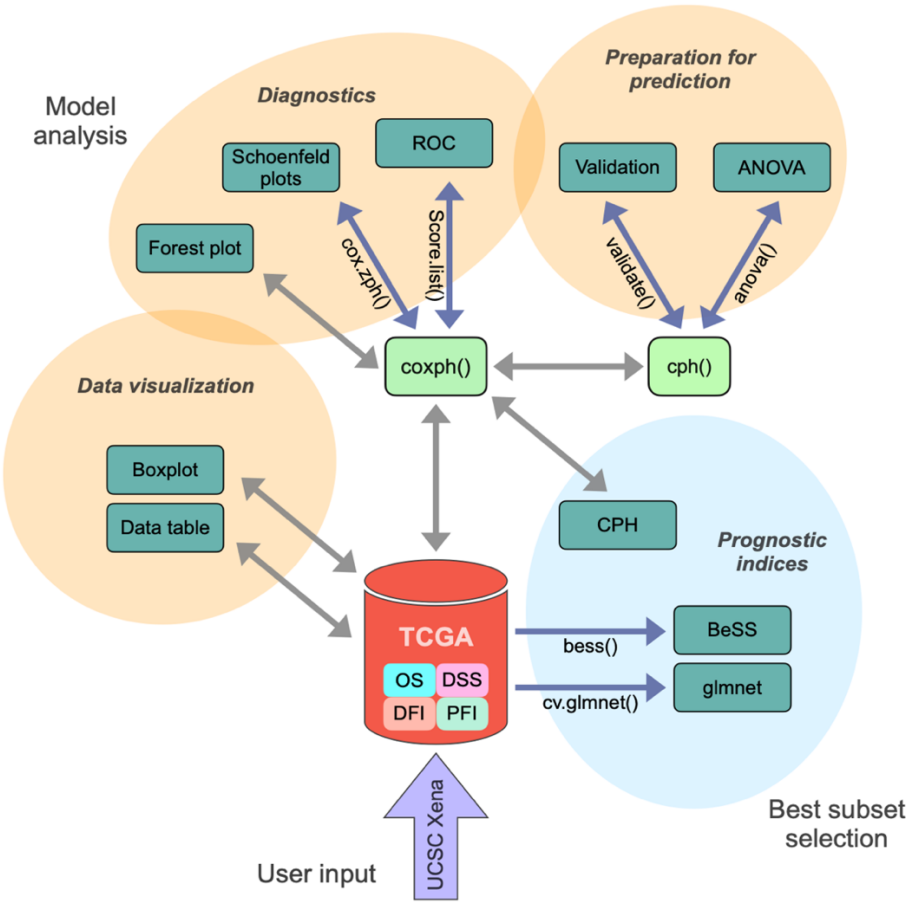

Figure 33: App architecture for SmulTCan indicates the two main parts of the app, model analysis and best subset selection. Reactivity in the app is shown with double-headed arrows. Functions required for the computations in each main tab, imported from existing packages for CPH model analysis are indicated on the arrows. Figure taken from Ozhan et al. (2021), see Permissions at the end of this document.

4.3.1.2 APP INTERFACE

The interface of the app is shown in Figure 34, where the input TSV file consists of expressions of the Slit-Robo genes. Users can select any one of the thirty-three TCGA datasets for which survival data is embedded within the app from the “Dataset” menu on the side-bar panel. The app also allows for determination of the “# of genes” in the input gene set from a slider on the side-bar panel. Input genes can then be added/removed from the “Genes” menu, while the updates in the main panel are displayed interactively with the change in the input genes. In in Figure 34, all Slit-Robo members have been selected for analysis with the app in the “Genes” menu, and their expression distributions were visualized with the “Boxplot” tab for the selected survival type, in this case DSS. The HTML functionality included in the app allows the selected main tabs and their sub-tabs to be visualized as medium slate blue.

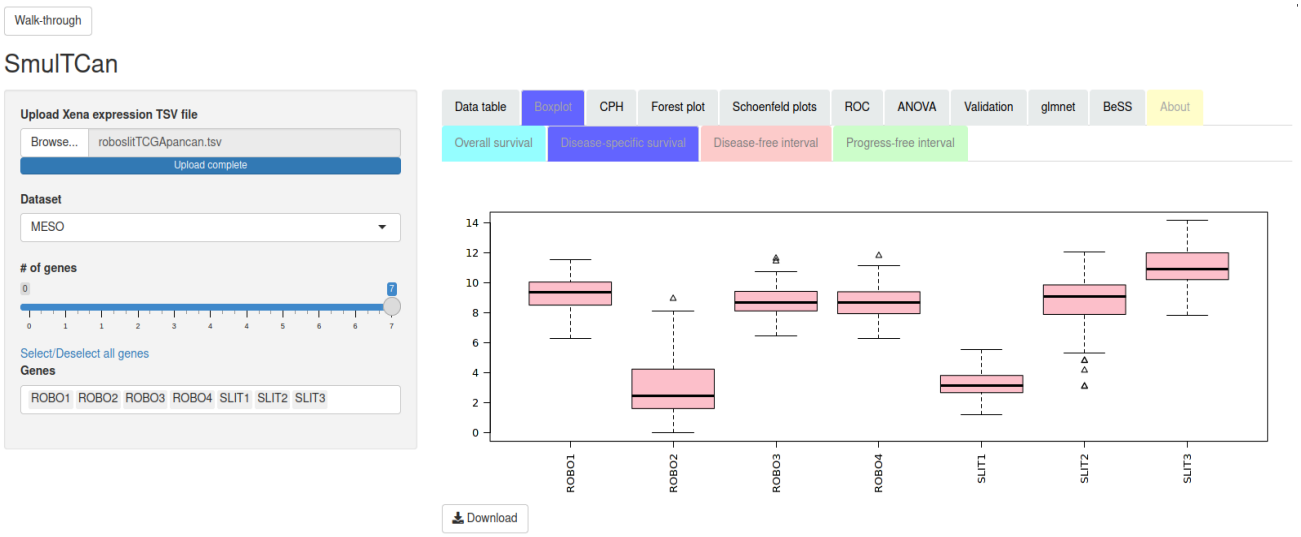

Figure 34: The SmulTCan app interface includes a side-bar panel where users can upload the file containing expressions of genes they would like to analyze, which they obtained from UCSC Xena earlier. Distribution of Slit-Robo expressions in TCGA-MESO.

Slit-Robo members are shown in the forest plot of the screenshot in Figure 35, in which the user has selected the ESCA from the "Dataset" menu. Each main tab ends with sub-tabs for the four different survival types in TCGA. In this forest plot, the sub-tab for DSS has been selected. The plot indicates *ROBO4* and *SLIT1* have slightly hazardous roles in ESCA with p-values < 0.05.

Figure 35: Screenshot from SmulTCan showing the forest plot of ESCA for DSS. The "Genes" menu in the side-bar panel indicates that all seven members of Slit-Robo have been selected.

Another screenshot from SmulTCan (Figure 36), this time indicating results from the "ANOVA" tab's "Plot" sub-tab, in which Slit-Robo members were ranked based on their p and $\chi^2 - df$ values. According to the ANOVA results, *SLIT2* is the best predictor among Slit-Robo for OS in the THYM dataset. The "Stats" sub-tab of the "ANOVA" main tab contains the table version of the ANOVA results.

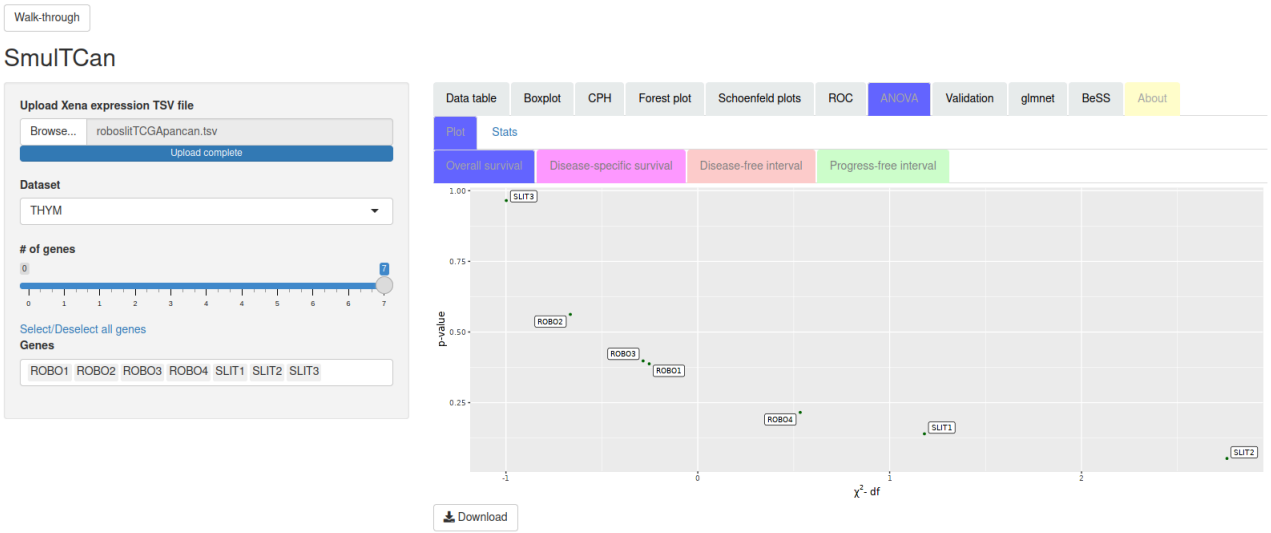

Figure 36: Results of ANOVA analysis of multivariable CPH models built with input genes sets can be visualized from the "ANOVA" tab's "Plot" sub-tab.

4.3.1.3 OUTPUTS FROM SMULTCAN WITH AXON GUIDANCE LIGAND-RECEPTOR GENE SETS

The prognostic HR signature for the complete netrins-receptors gene set in the colon adenocarcinoma dataset, TCGA-COAD, is displayed in the forest plot in Figure 37 for OS. The signature reveals the strongly protective role of the *UNC5C* receptor and the hazardous roles of the *NTN3* and *NTNG2* ligands in COAD. The global AIC score for this HR GS is 653.25.

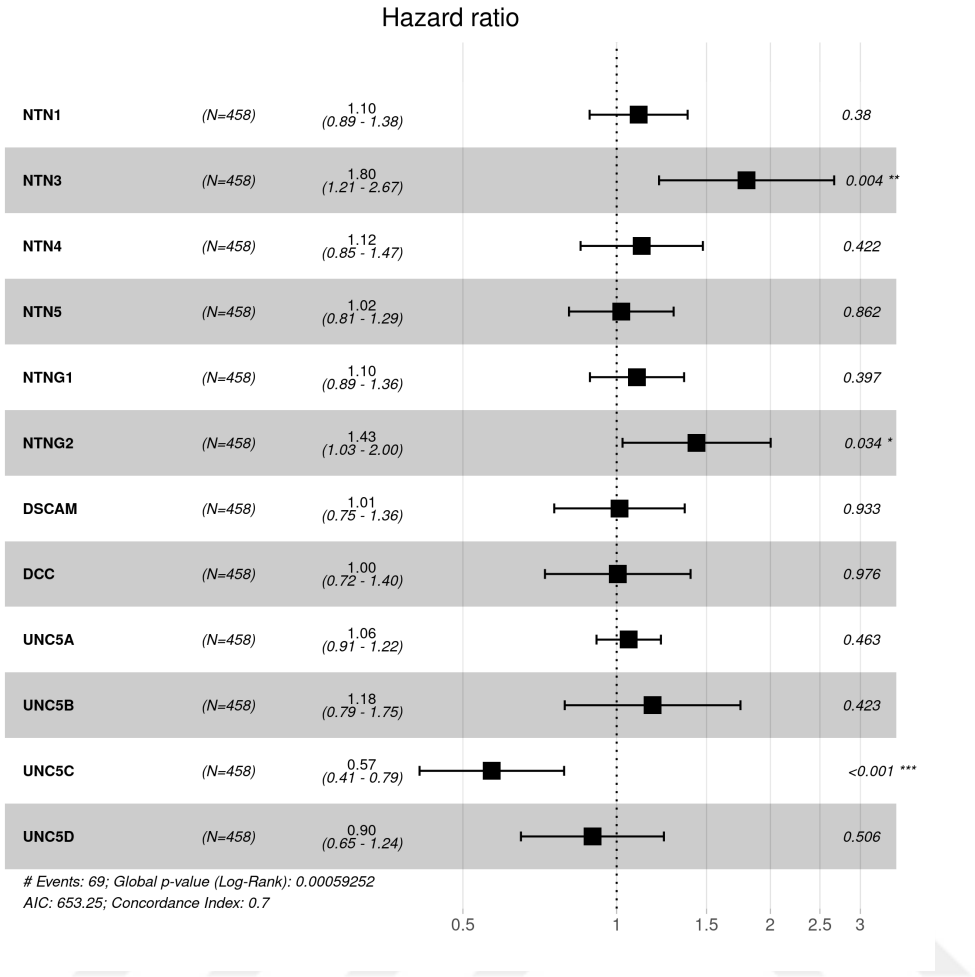

Figure 37: Forest plot of netrins-receptors in COAD indicates positive association of *UNC5C* and negative association of *NTN3* with OS.

The ROC plot of the same COAD dataset for OS indicates that the Cox model is a reasonable fit with a ≃70% AUC score with all twelve netrins-receptors Figure 38. This multivariable CPH model could be used as a training set for prediction.

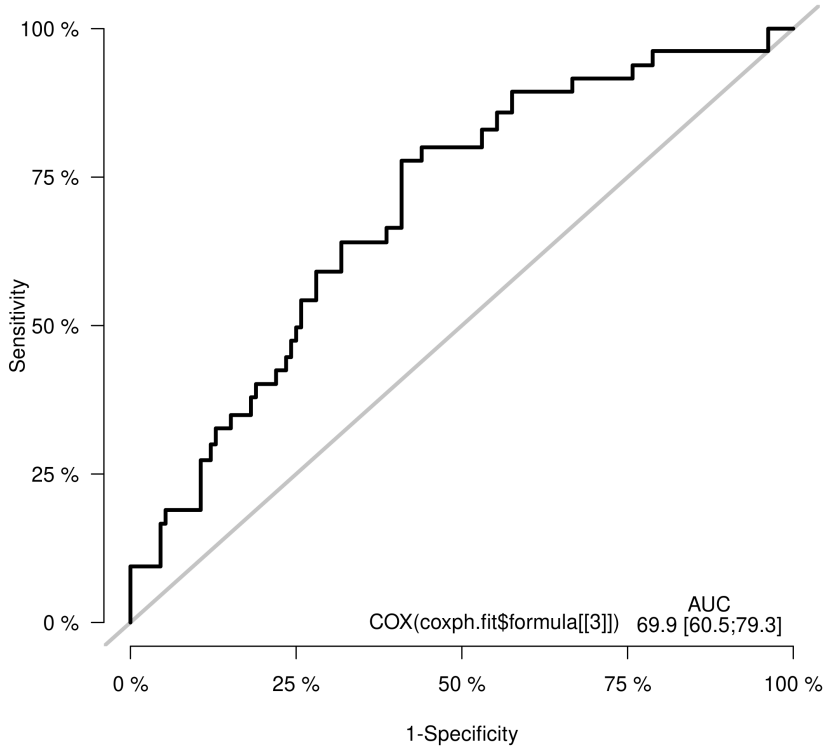

Figure 38: ROC plot for OS in COAD, whereby the CPH model includes all twelve netrins-receptors. The plot indicates the CPH model is a reasonably good predictor of prognosis.

ANOVA results plot Figure 39 of the CPH model of the cholangioma dataset TCGA-CHOL indicate that the best predictor of the netrins-receptors gene set is the *DCC* receptor, since it is the gene with the highest $\chi^2 - df$ value and lowest p-value.

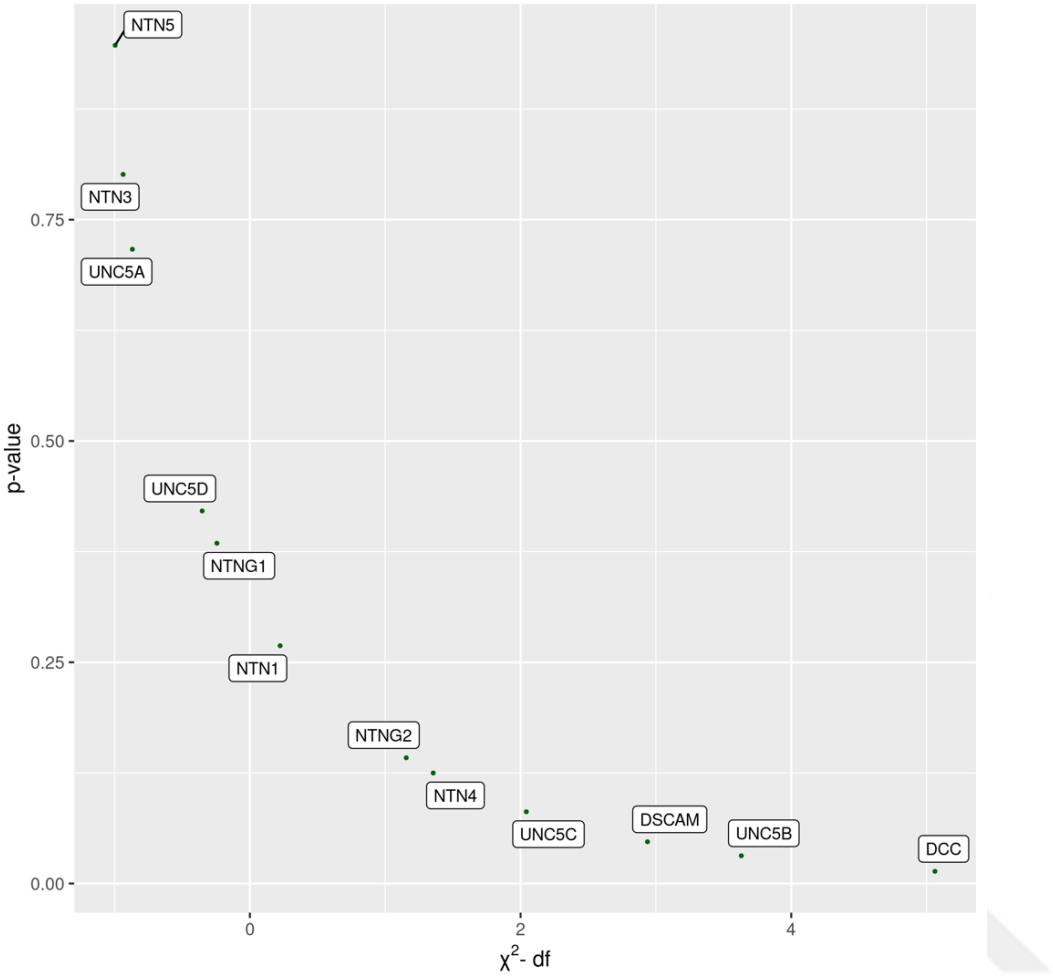

Figure 39: ANOVA results of CHOL for DSS indicate *DCC* as the best predictor among netrins-receptors.

Table 24 below lists the χ^2, degrees of freedom and p-values of netrins-receptors of the ANOVA plot in Figure 39, as well as the totals of these columns for the CPH model. The table was generated with the TSV file downloaded for DSS in CHOL from SmulTCan's "ANOVA" main tabs "Stats" sub-tab.

Table 24: ANOVA results table of CHOL for DSS, generated from the downloaded output TSV file from SmulTCan.

	χ^2	Degrees of freedom	*P*-value
NTN1	1.2235183	1	0.2686713
NTN3	0.0634162	1	0.8011758
NTN4	2.3546556	1	0.1249095
NTN5	0.0044159	1	0.9470180
NTNG1	0.7562438	1	0.3845066
NTNG2	2.1552416	1	0.1420841
DSCAM	3.9378143	1	0.0472121
DCC	6.0619321	1	0.0138126
UNC5A	0.1317439	1	0.7166308
UNC5B	4.6310904	1	0.0313976

UNC5C	3.0426877	1	0.0811015
UNC5D	0.6475528	1	0.4209890
TOTAL	11.8796684	12	0.4553915

The BeSS algorithm which employs ridge regression excluded only the *UNC5C* gene from the netrins-receptors set in TCGA-THYM for OS. Coefficients of the eleven genes for the best model are plotted in Figure 40, which was downloaded from the app. The coefficients indicate that according to the default BeSS model, *DCC*, *NTN1*, *NTN3*, and *UNC5B* positively contribute to prognosis, while *DSCAM*, *NTN4*, *NTN5*, *NTNG1*, *NTNG2*, *UNC5A*, and *UNC5D* negatively contribute to prognosis.

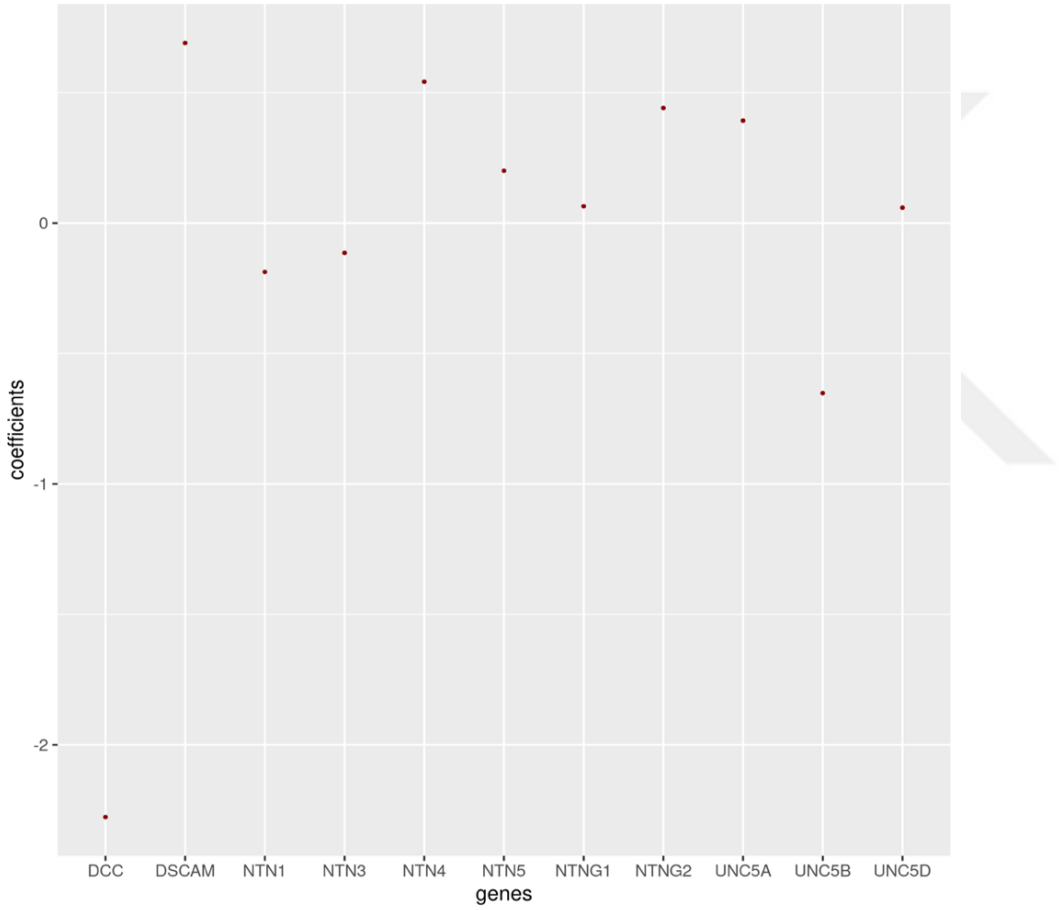

Figure 40: Plot of the coefficients of the best subset of netrins-receptors determined with the default parameters of the BeSS algorithm contain eleven genes in THYM for OS.

ANOVA results of the CPH model of netrins-receptors for OS in THYM corroborated the BeSS tab of SmulTCan, where *DSCAM* was found to have the lowest $\chi^2 - df$ value of the twelve genes Figure 41. The plot was downloaded from SmulTCan's "ANOVA"

tab's "Plot" sub-tab, used to visualize the ranking of input genes in the CPH models based on ANOVA results, which can be found in the same main tab's "Stats" sub-tab.

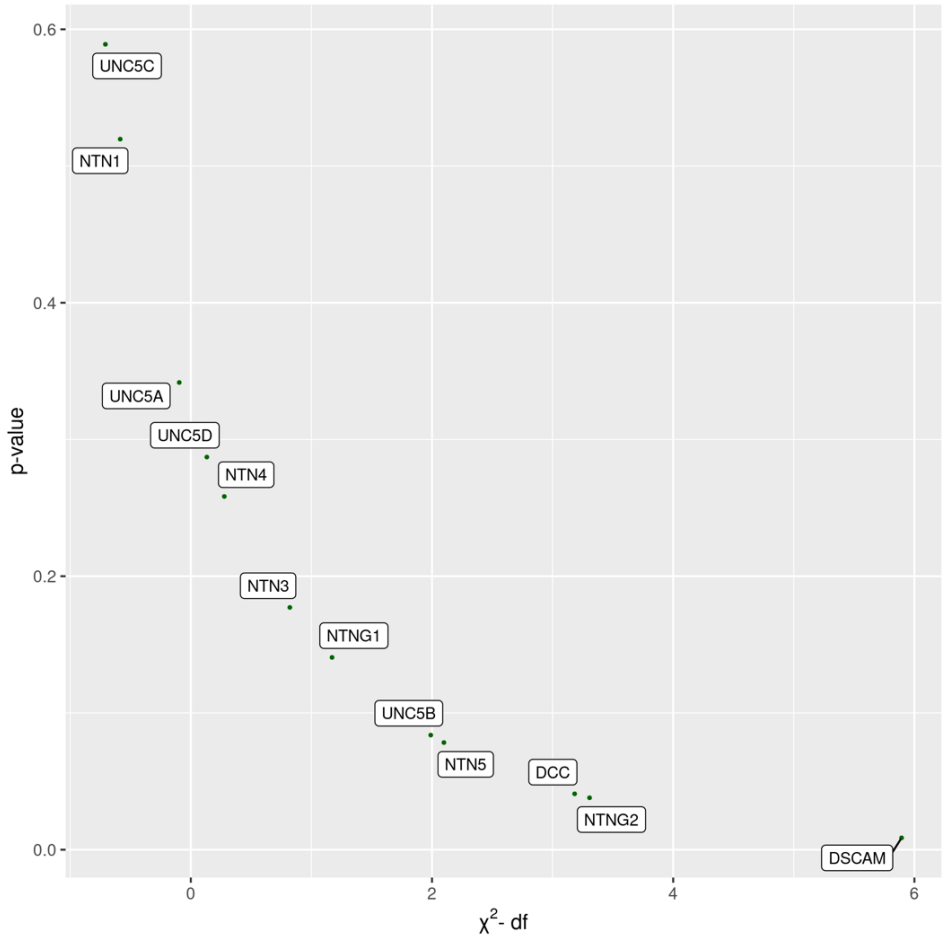

Figure 41: Plot of ANOVA results in THYM for OS with netrins-receptors. The best predictor of the gene set in this case is *DSCAM*.

Forest plot of OS for the Semas-receptors gene set in ACC indicates strongly protective roles for *NRP2*, *SEMA4D*, *SEMA5A*, and *PLXND1* Figure 42. The Akaike information criterion (AIC) score for the overall multivariable model with all thirty-one genes is 182.08, with a global p-value $< 10^{-7}$. Additionally, hazardous roles have been identified for the *SEMA3E*, *SEMA4B*, *SEMA4C*, *SEMA6B*, *SEMA6B*, *SEMA6C*, *SEMA7A*, and *PLXNA1*. Neuropilin members were found to have opposing effects on OS, with *NRP2* positively ($p < 0.001$) and *NRP1* negatively ($p < 0.1$) affecting OS rates.

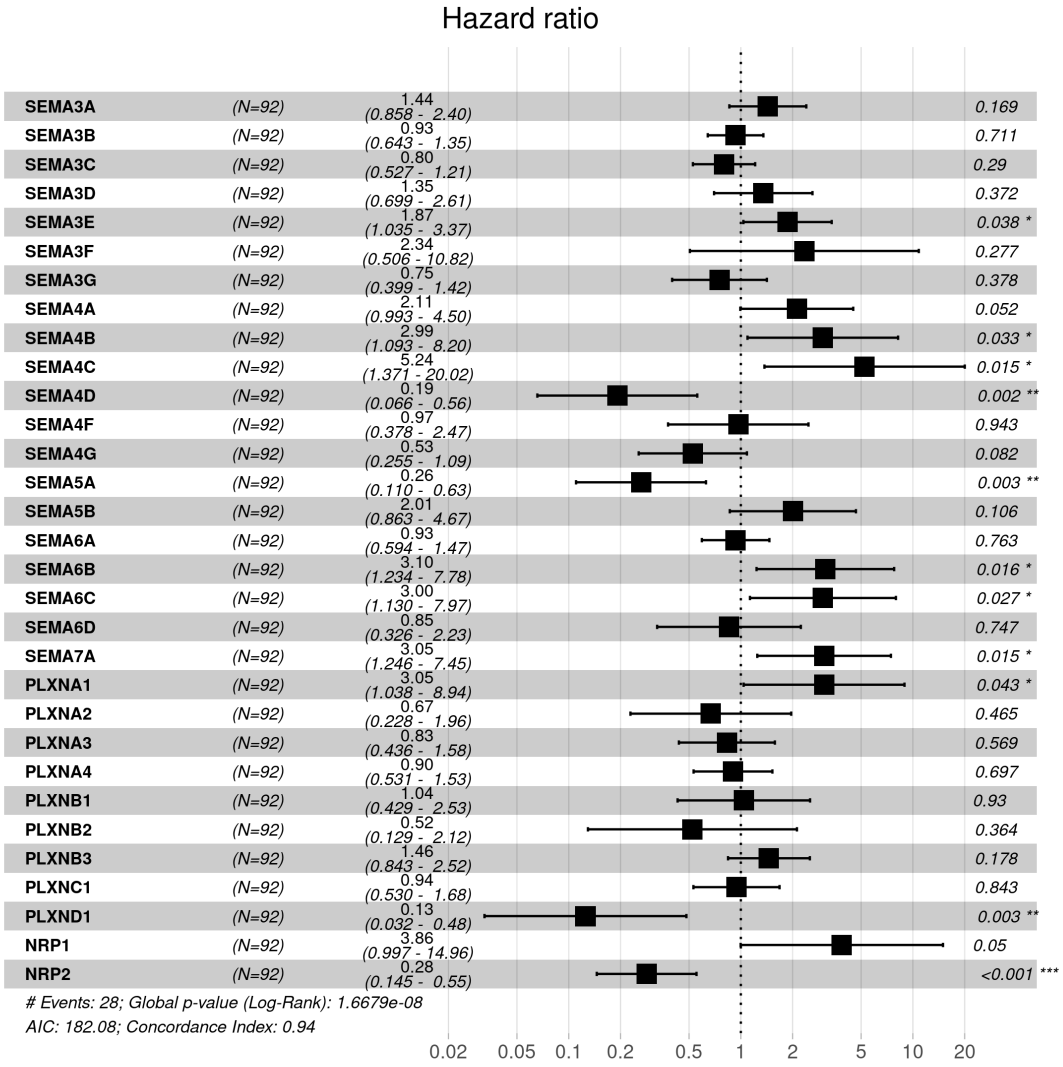

Figure 42: Forest plot of ACC for OS indicates a strongly protective role for *NRP2* and *SEMA4D* among Semas-receptors.

The “glmnet” main tab of SmulTCan was used with KIRC, with the number of folds in cross-validation set to the sample size and the elastic net adjusted to 1 (which corresponds to lasso), to obtain the best subset of Slit-Robo which consisted of *ROBO3* and *ROBO4* as observed from the “Coeffs” sub-tab for survival types OS, DSS, and the PFI (Figure 43). A best subset could not be found for DFS. The two members *ROBO3* and *ROBO4* from the seven members of the Slit-Robo family were able to significantly differentiate between prognostic outcomes for all three survival types in which they were identified.

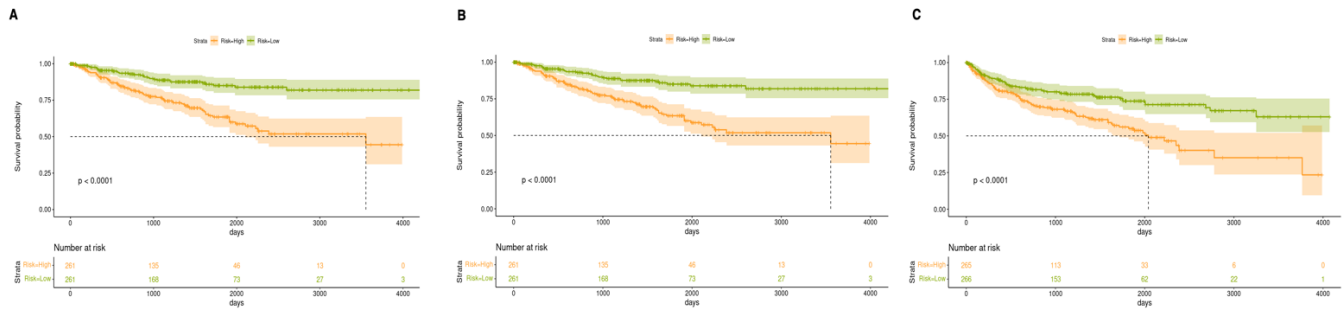

Figure 43: K-M plots of KIRC for (A) OS, (B) DSS, and (C) PFI obtained from SmulTCan for the Slit-Robo genes. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type with respect to the log-rank test.

On the other hand, the best subset of Slit-Robo for the KIRP dataset using lasso consisted of *ROBO2*, which contributed to low-risk prognosis and *SLIT3*, which contributed to high-risk prognosis for survival types OS, DSS, and the PFI, as shown in the K-M plots of Figure 44. This best subset consisting of two genes out of the seven was able to differentiate between prognostic outcomes in each indicated survival type. Again, the number of folds in the cross-validation was set to the sample size for each K-M analysis.

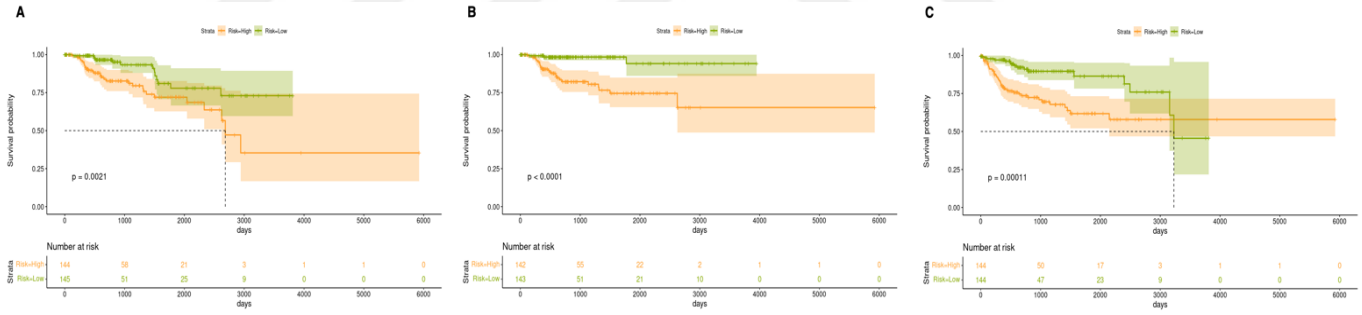

Figure 44: K-M plots of KIRP obtained from SmulTCan for (A) OS, (B) DSS, and (C) PFI with Slit-Robo. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type with respect to the log-rank test's results.

For the LGG dataset, a single gene, *SLIT1*, was able to differentiate between prognostic outcomes for OS, DSS, and the PFI, shown in the K-M plots of Figure 45A, B, and D, respectively. However, for the DFI, *ROBO3* and *SLIT3* in addition to *SLIT1* were involved; this best subset of three genes also differentiated between prognostic outcomes (Figure 45C). *SLIT1* contributed positively to prognosis in all four survival types, while *SLIT3* also contributed positively to prognosis, *ROBO3* contributed negatively in the DFI. Information of the prognostic contributions of each gene in the best subsets were obtained

from the “Coeffs” sub-tab of the “glmnet” main tab of SmulTCan. The number of folds in cross-validation was set to the sample size and the strictest elastic net parameter (= 1, corresponding to lasso) was used.

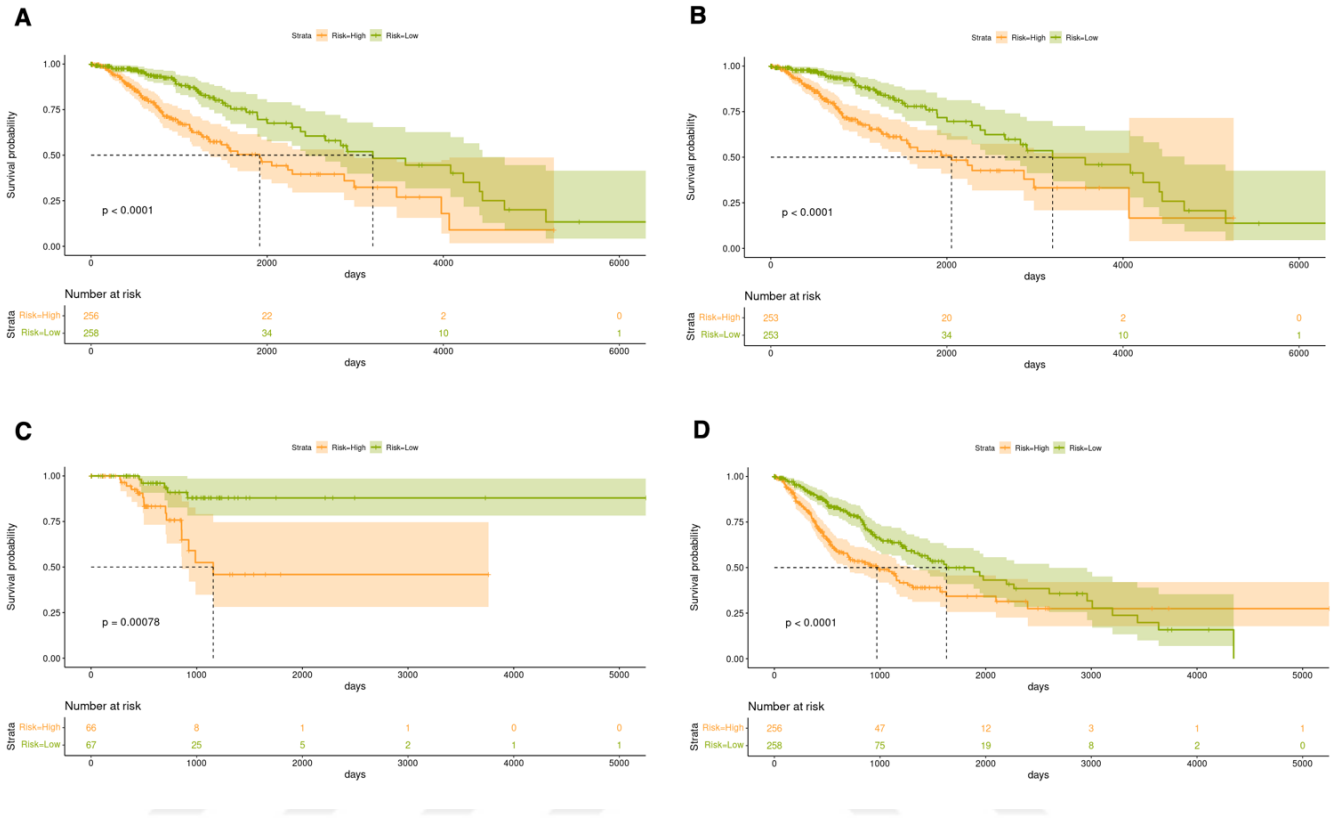

Figure 45: K-M plots of LGG for A) OS, (B) DSS, (C) DFI, and (D) PFI. Low and high-risk prognostic outcomes significantly differentiated based on the median PI for each survival type.

For DSS in LGG, SmulTCan selected only two genes *SEMA4G* and *SEMA6B*, out of the thirty-one genes in the Semas-receptors set as the best subset with lasso and the default ten-fold cross validation. According to this, both genes contribute positively to DSS, and they alone are sufficient for differentiating between High vs. Low-risk prognostic outcomes ($p < 0.0001$) by $\simeq$ 2.8-fold ($p < 0.0001$), which can be observed from the “K-M data” sub-tab of the “glmnet” main tab. The K-M plot of the best subset

consisting of *SEMA4G* and *SEMA6B* is shown in Figure 46, while the cumulative hazards plot of the same subset is shown in Figure 47.

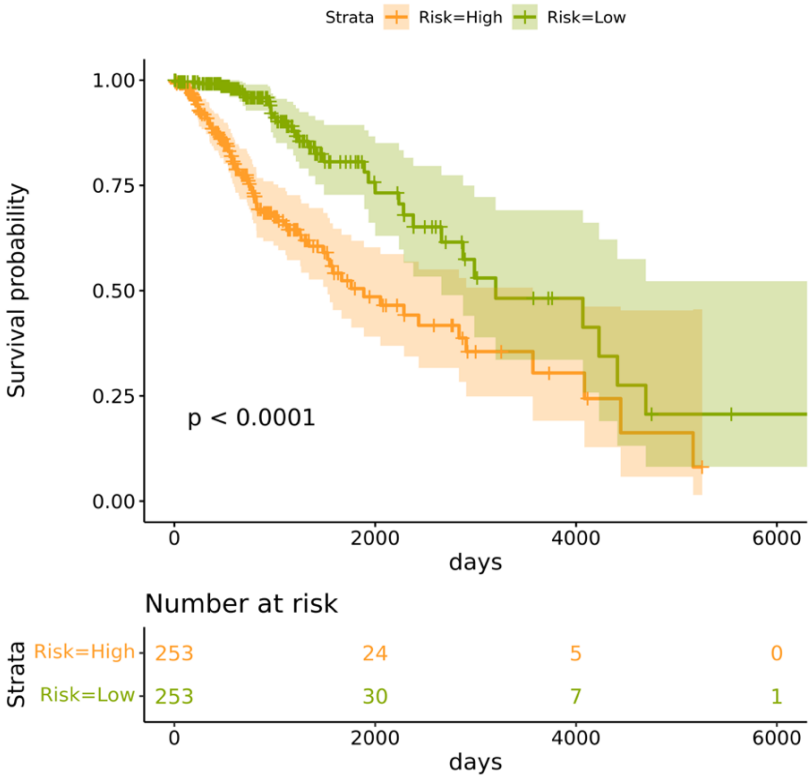

Figure 46: K-M plot of DSS in LGG for the best subset of Semas-receptors gene set identified with lasso.

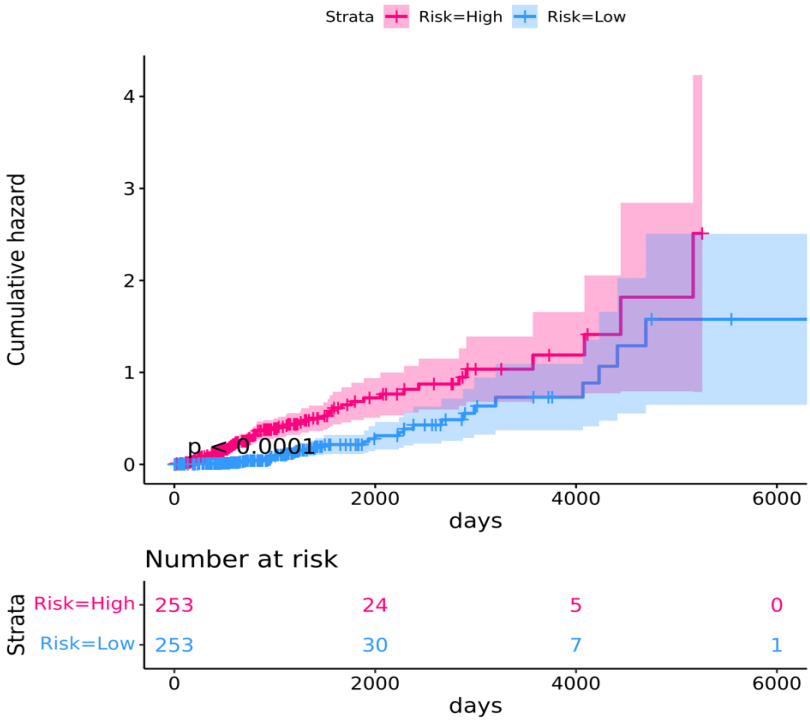

Figure 47: Cumulative hazards plot of DSS in LGG for the best subset of Semas-receptors identified with lasso has differentiated between prognostic outcomes.

The ROC plot for the lasso result of DSS in LGG shown in Figure 48 indicates that the PI computed from the *SEMA4G* and *SEMA6B* subset of Semas-receptors is a strong classifier of prognostic outcomes with an AUC% score of ≃ 83%.

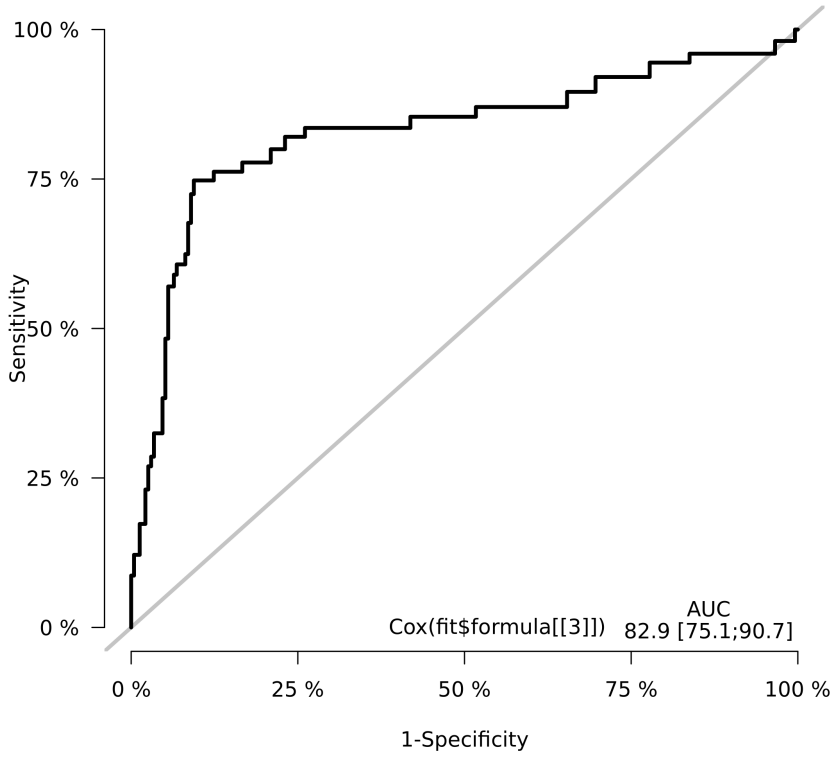

Figure 48: ROC plot of the PI covariate of the best subset consisting of *SEMA4G* and *SEMA6B* in LGG.

4.3.2 ClusterHR Shiny application

In this section, the TCGA-PANCAN HR clustering profiles of Slit-Robo, netrins-receptors and Semas-receptors were investigated using the ClusterHR Shiny app, available from http://konulabapps.bilkent.edu.tr:3838/ClusterHR/. The heatmap outputs of ClusterHR contain labels indicating significance of p-values with the following rule: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and + for $p < 0.1$.

4.3.2.1 App interface

A screenshot of the ClusterHR app is given in Figure 49 for Slit-Robo family's input file from UCSC Xena. The top thirteen datasets have been selected from the "Datasets" menu in the side-bar panel, with which the heatmap in the main panel has been drawn. The main panel for the "Heatmaps" main tab contains an additional internal side-bar panel which allows users to determine whether they would like to cluster the heatmaps by only rows (cancers), only columns (genes), or as in this case, both columns and rows. Like SmulTCan, ClusterHR has embedded HTML functionality to allow for visualization of selected tabs with medium slate blue.

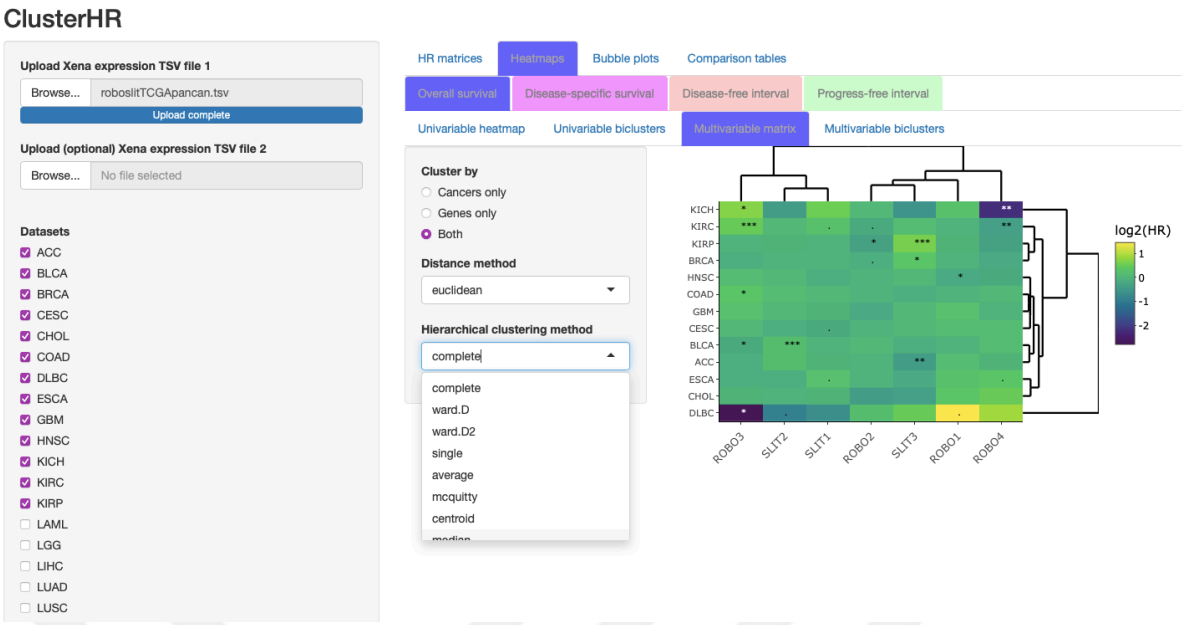

Figure 49: Screenshot of the ClusterHR app indicating the selected thirteen TCGA datasets on the side-bar panel. The "Multivariable matrix" sub-tab of the OS sub-tab of the "Heatmaps" main tab was selected. The user can select between clustering options, the distance methods in the clustering and the hierarchical clustering methods using the internal side-bar panel.

4.3.2.2 OUTPUTS FROM CLUSTERHR WITH AXON GUIDANCE LIGAND-RECEPTOR GENE SETS

The clustering heatmap in Figure 50 of Slit-Robo $\log_2$HR values from multivariable Cox regression results for DSS, which was drawn with Euclidean distance and complete linkage hierarchical clustering, indicates that the gene set is divided into two groups: *ROBO3-SLIT2-SLIT1-ROBO2* and *ROBO1-SLIT3-ROBO4*. Looking more closely, we can observe the Slit-Robo pairs *SLIT2-ROBO3*, *SLIT1-ROBO2* and *SLIT3-ROBO4*.

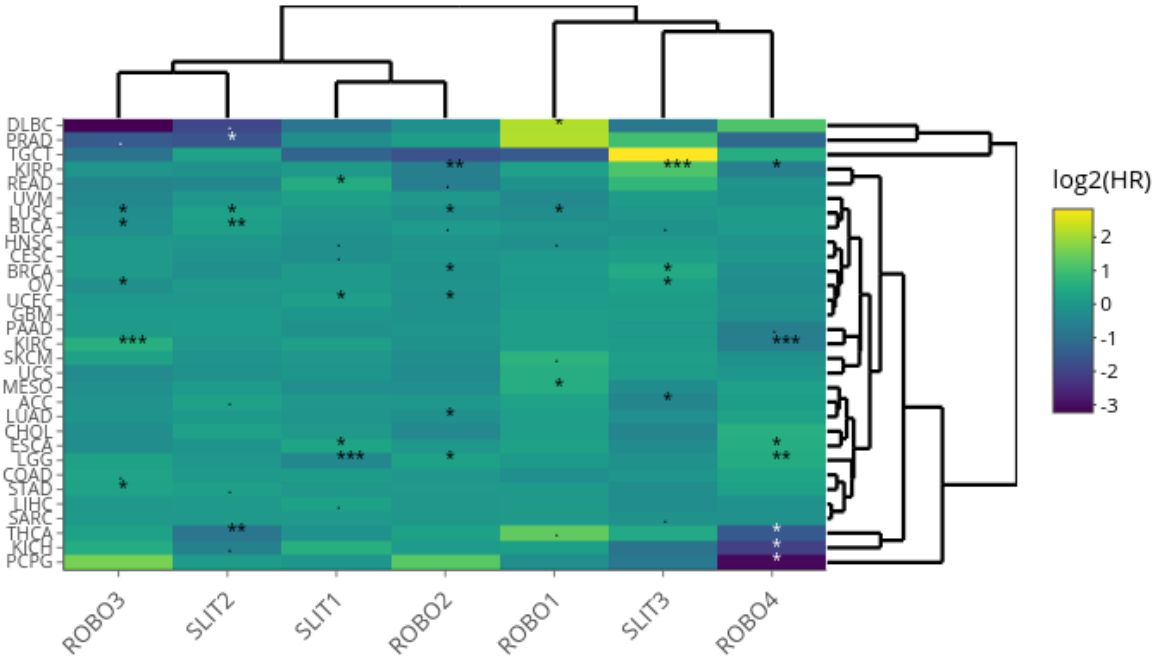

Figure 50: Slit-Robo multivariable log_2HR values across thirty-one TCGA datasets displayed in a clustering heatmap.

The heatmap in Figure 51 displays the clustering of log_2HR values in the netrins-receptors gene set for DSS. Initial observation indicates that *NTN3-UNC5B-DCC* form a separate cluster from the rest of the netrins-receptors. In the other cluster, the *NTN4* gene forms a separate branch from the other eight genes. In this heatmap, netrin-receptor pairs formed by log_2HR values can be observed as in the Slit-Robo DSS heatmap above. The ligand-receptor pairs in this case are *NTNG2-UNC5A*, *NTN1-DSCAM* and *NTN3-UNC5B*. The heatmap was drawn with Euclidean distance and complete linkage hierarchical clustering.

The TGCT dataset was intentionally excluded from the clustering analysis since none of the netrins-receptors genes' log_2HR results were significant for this dataset. This caused the unnecessary increase of the range of the log_2HR legend, making other datasets' log_2HR results difficult to observe. An antagonistic netrin-receptor pair can be seen in heatmap in Figure 51 in LGG for *NTN4-UNC5C*, where *NTN4* was positively associated while *UNC5C* was negatively associated with survival (p-values < 0.05).

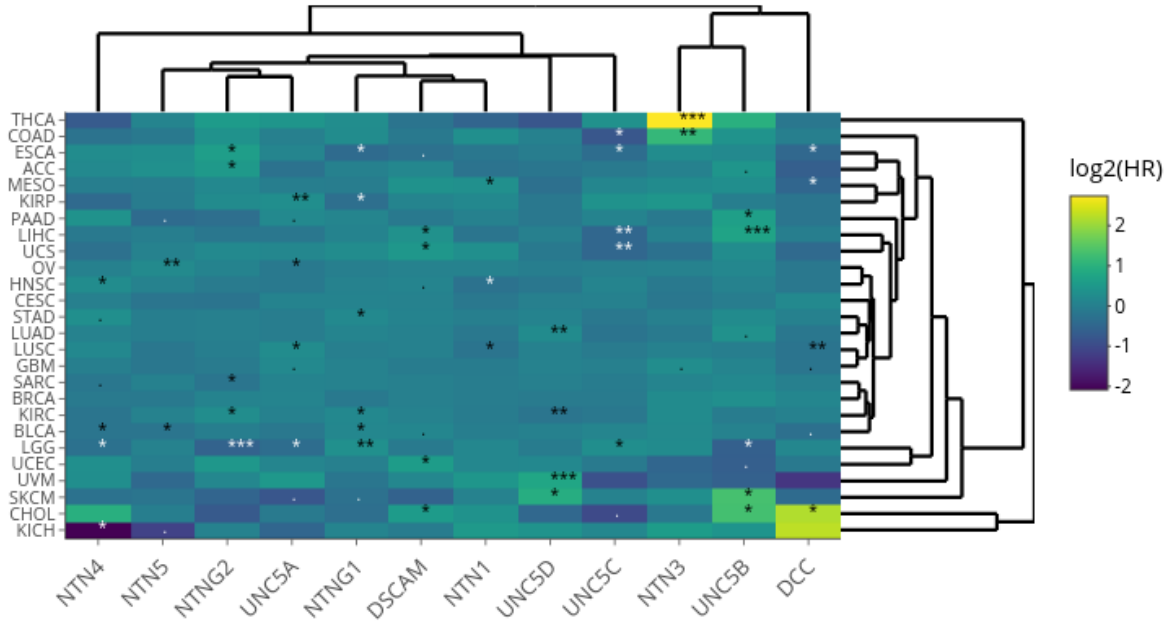

Figure 51: Multivariable analysis $\log_2$HR values computed from netrins-receptors across twenty-six TCGA datasets are displayed in the clustering heatmap.

Clustering of the multivariable $\log_2$HR values across TCGA datasets in the Semas-receptors gene set for the DFI is given in the heatmap in Figure 52. This clustering heatmap shows that *PLXND1* has a separate DFI $\log_2$HR profile across the indicated TCGA datasets. KIRC separates from the other eighteen TCGA datasets with respect to $\log_2$HRs of Semas-receptors, a spectrum of values from the most protective *SEMA4D* ($\log_2\text{HR} \simeq -20$, $p < 0.05$) to the most hazardous *PLXND1* ($\log_2\text{HR} \simeq 25$, $p < 0.01$). The very large and very small HR values are caused by the normalization method of Xena expressions, though they may nevertheless be relevant for understanding the comparative hazards within a gene set.

The Sema-receptor $\log_2$HR pairs identified across nineteen TCGA datasets in this gene set from this heatmap are: *SEMA7A-PLXNC1*, *SEMA4C-NRP2*, *SEMA3G-PLXNA4*, *SEMA3E-PLXNB3*, and *SEMA4B-NRP1*. Again, the Euclidean distance and complete linkage were used for hierarchical clustering in this heatmap. The PRAD and READ datasets had to be excluded from this clustering analysis due to a singularity error thrown by coxph() for the OS and DSS of these datasets, causing ClusterHR to crash.

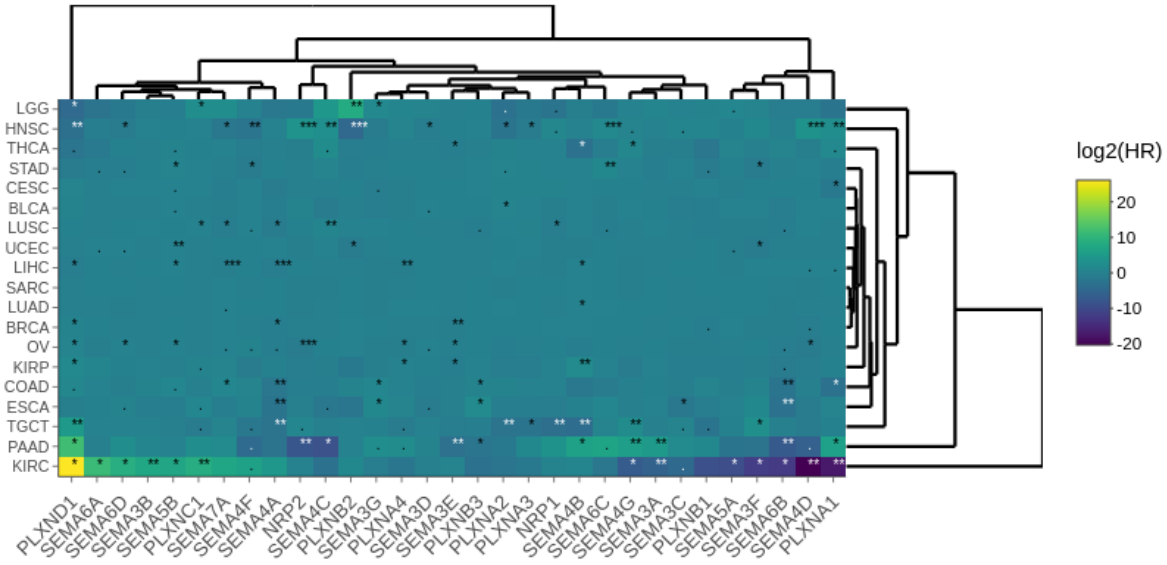

Figure 52: Multivariable analysis log_2HR values calculated from Semas-receptors across nineteen TCGA datasets are displayed in the clustering heatmap.

The data matrix of the main heatmap for the DFI in Figure 52 when used with the "BCCC" method generated two biclusters. These biclusters were then used in hierarchical clustering themselves with Euclidean distance and complete linkage, resulting in the heatmaps in Figures Figure **53**-Figure **54**. In the first bicluster in Figure 53 involving all thirty-one genes of Semas-receptors, *PLXNB2* shows a distinct log_2HR profile across seventeen rows of TCGA datasets. Then, *PLXND1* separates from the rest of the twenty-nine genes and those genes divide into two groups. Within this whole bicluster, LGG forms a separate branch with a range of log_2HR values, at odds with HNSC, which is also distinct from the remaining fifteen datasets.

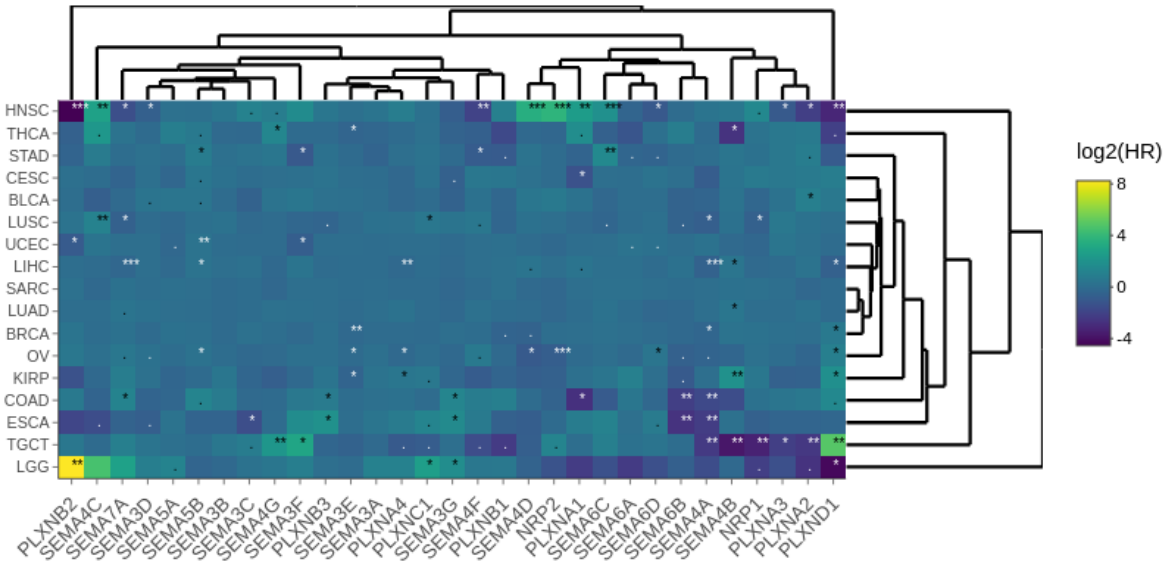

Figure 53: Bicluster generated from the Semas-receptors gene set with Cheng and Church's algorithm, where all members of the Semas and their receptors were included. There are seventeen TCGA datasets in this bicluster.

The second bicluster shown in Figure 54 contains only the two TCGA datasets PAAD and KIRC. Only ten of Semas-receptors are present in this bicluster. The *SEMA4C-NRP2* pair reappears here, clustering separately from the rest of the genes with the protective effect of *NRP2* in PAAD.

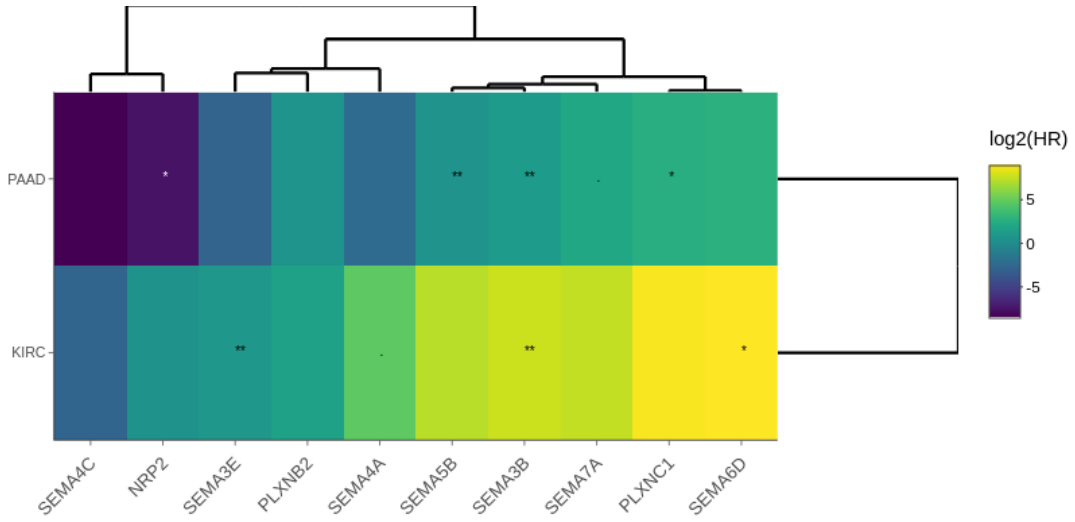

Figure 54: Another bicluster from Semas-receptors; this one contains nine genes from the gene set. We can observe that this subset of genes behaves differently in terms of HRs in PAAD and KIRC. For these two TCGA datasets, the nine members of Semas-receptors have similar $\log_2$HR profiles.

4.4 Discussion

Survival analysis and its applications are widespread in biomedical research, particularly in cancer where K-M analysis is frequently used to determine a covariate's effect on survival. This factor could either be continuous or categorical. In this study, two web-based Shiny applications that are both freely available and require no registration or the installation of additional software, have been created for the purpose of analyzing the impact of gene expression on survival rates of TCGA datasets. Both the SmulTCan and ClusterHR apps are based on the principals of Cox regression and have been used in this chapter to analyze gene sets made up of different types of axon guidance ligands and their receptors. Prognostic GSs are the expression levels of a group of genes associated with the tumor prognosis of a cancer type and are thought to have the potential to revolutionize cancer research (Y. Wang et al., 2017). In this study, the concept of HR GSs was developed, representing the HR profiles for sets of genes in cancer datasets.

Both the SmulTCan and ClusterHR apps have user friendly interfaces with modular structures, owing to Shiny's design principles. Both apps were designed to accommodate miRNA expressions, as well as gene-level copy number variations (CNVs) and methylation β-values. For gene-level methylation β-value sets, it is best if the user multiplies the values in the input TSV file downloaded from UCSC Xena by 100 before uploading the file onto SmulTCan or ClusterHR for analysis.

4.4.1 SmulTCan

SmulTCan analyzes the multivariable survival signatures of sets of genes in detail, using methods from existing packages that allow only for command-line analyses. A useful feature of the app includes its "Data table" tab, from which users can download tables of expressions of their input genes together with TCGA survival information, which they could then use as training sets for their own cancer survival data in prediction algorithms after analyzing the CPH model with other model analysis tabs in the app.

Another advantage of SmulTCan is its incorporation of several different best subset selection methods. While users can determine best subsets directly from the CPH model, they can also use the elastic net slider from the "glmnet" tab of the app to determine best subsets. SmulTCan also offers the less strict BeSS algorithm based on the primal dual active set (PDAS) algorithm to determine the best predictors from a gene set, which is faster than glmnet's coordinate descent algorithm (Wen et al., 2020). SmulTCan is unique among online multivariable gene/miRNA expression-based survival analysis tools to allow for prognostic analyses with the best subset selected from an input gene set.

4.4.2 ClusterHR

The ClusterHR Shiny application allows users to analyze multiple TCGA datasets at a time, while it is also possible to analyze two different gene sets from two different input files within a single session. ClusterHR helped visualize the HR clustering profiles across TCGA datasets of the axon guidance ligands-receptors sets chosen in this study and aided in the identification of synergistic and antagonistic ligand-receptor pairs in different cancers from each of the three gene sets in the study.

4.4.3 Slit-Robo

Results with Slit-Robo using best subset selection with the glmnet algorithm's lasso option (Simon et al., 2011; Zou, 2005) through SmulTCan indicated this pathway an important determinant for prognostic outcomes in kidney cancers and cancers of neuronal origin including LGG. Different best subsets were identified to be involved for each cancer, in which genes contributed differentially to prognosis. The *SLIT3-ROBO2* pair was found to be antagonistically involved in KIRP's best subset for prognosis. Antagonistic Slit-Robo pairs *SLIT3-ROBO3* and *SLIT3-ROBO1* were found to be involved in LGG's prognosis for the DFI.

Additionally, the *SLIT1-ROBO2* pair had an aggravating role in KICH. This pair, which is known for its role in trigeminal gangliogenesis (Shiau et al., 2008), was also found to have roles in both E-cadherin degradation through the recruitment of the Hakai ubiquitin ligase (Zhou et al., 2011) and the formation of the E-cadherin/β-catenin complex through the AKT/GSK3β/Snail pathway (Wu et al., 2017), as well as inhibition of N-cadherin-mediated cell adhesion (Rhee et al., 2002). The *SLIT1-ROBO2* pair has also been implicated in colorectal cancer by Grone et al. (2006).

ClusterHR results of Slit-Robo have shown that several members of this family were associated with survival rates in different types of kidney cancers. In KIRP, *ROBO2* and *ROBO4* had protective roles with HRs ≃ 0.70 ($p = 0.001$ and $p < 0.05$, respectively), while *SLIT3* had a strongly hazardous role with HR ≃ 2.20 ($p < 0.001$). On the other hand, in KIRC, *ROBO3* and *ROBO4* were found to have prominently opposing roles on survival, with *ROBO3* expression negatively ($p < 0.001$) and *ROBO4* expression positively ($p < 0.001$) affecting prognosis. KIRC's best subset identified with glmnet for DSS through SmulTCan corroborates this result.

The protective effect of *ROBO4* was also observed in KICH ($p < 0.05$), a feature which was supported in previous studies with endothelial cells (Jones et al., 2008; Park et al., 2003). The involvement of Slit-Robo in kidney cancers was expectable, since the pathway is known to regulate organ morphogenesis including kidneys (Tong et al., 2019; Wu et al., 2017). Antagonistic and synergistic Slit-Robo pairs, which were identified in both best subsets of several TCGA datasets and the clustering heatmap for DSS, could be investigated further experimentally for their prognostic roles in their respective cancer datasets.

4.4.4 NETRINS-RECEPTORS

Through SmulTCan, HR GSs have been identified for OS in COAD for netrins-receptors; in which the *NTN3-UN5C* pair was antagonistically involved in prognosis. The protective role of axon repulsive *UNC5C* receptor in prognosis was previously observed in colorectal cancer (Shin et al., 2007). More recently, a study involving thirty-two TCGA datasets confirmed the tumor suppressor status of *UNC5C* in sixteen of them, including COAD (Xing et al., 2021).

Clustering heatmap for DSS with netrins-receptors also indicates that this gene set, like Slit-Robo, is involved in kidney and neuronal cancers. A protective role was observed for *NTN4* ($p < 0.05$) and *NTN5* ($p < 0.1$) in KICH, which constituted novel findings. Several members of this gene set were found to be involved in prognosis of LGG, including the netrins *NTN4*, *NTNG1*, *NTNG2*, as well as receptors *UNC5A*, *UNC5B*, and *UNC5C*. *NTNG1*'s protective role identified in the kidney cancer datasets KIRP and KIRC corroborates the results in Hao et al. (2020), however, *NTN1* was not associated with survival rates in pan-renal cell carcinoma (pan-RCC, i.e. KIRC, KIRP, and KICH) in this study using multivariable survival analysis with a gene set composed of members constituting an interdependent pathway.

Differences in HR GSs across pan-RCC may arise from the different cells of origin for each kidney cancer, while clear cell (KIRC) and papillary cell (KIRP) carcinomas arise from cells in the proximal convoluted tubules, chromophobe (KICH) carcinoma arises from intercalated cells in the distal convoluted tubules of the nephron (Tabibu et al., 2019).

The *NTN4* gene was also positively associated with prognosis in the neuronal cancer dataset LGG. While the gene has been previously implicated in metastasis of breast cancer (Xu et al., 2017), it was not associated with BRCA prognosis in this study. In LGG, the *NTNG2-UNC5A* pair was newly associated with DSS using ClusterHR.

4.4.5 SEMAS-RECEPTORS

Using complete lasso with Semas-receptors, it was possible to differentiate between high and low-risk prognostic outcomes for DSS in LGG with a best subset consisting of only two genes: *SEMA4G* and *SEMA6B*. SmulTCan also helped identify an HR GS for OS in ACC with the Semas-receptors gene set. Analyses using the apps developed in this chapter indicate that this gene set is also important in the prognoses of pan-RCC and neuronal cancers as the other two investigated axon guidance ligand-receptor gene sets.

ClusterHR identified two biclusters for the DFI from Semas-receptors, with a smaller bicluster composed of the PAAD and KIRC datasets; indicating that the ten genes in this bicluster, *SEMA4C*, *NRP2*, *SEMA3E*, *PLXNB2*, *SEMA4A*, *SEMA5B*, *SEMA3B*, *SEMA7A*, *PLXNC1*, and *SEMA6D* have similar HR profiles for these two datasets which is separate from the rest of the Semas-receptors and remaining TCGA datasets in the analysis. In contrary to its known pro-angiogenic role activating the Rho kinase (Sakurai et al., 2012), *SEMA4D* was found to have a protective role within Semas-receptors for DFI in KIRC. Other protective Semas for the KIRC dataset that came up in the heatmap were *SEMA3A* and *SEMA3F*; as well as *SEMA3C*, *SEMA4G*, *SEMA5A*, and *SEMA6B*. Demonstration of *SEMA3A*'s protective role in this dataset is within Semas-receptors is interesting, since the gene was associated with poor survival across TCGA-PANCAN (Zhang et al., 2020).

From the first larger bicluster of Semas-receptors, we can observe *PLXNB2*'s opposing effects on survival for HNSC and LGG, on which it has beneficial and hazardous effects, respectively. *PLXNB2* is a known receptor of angiogenin, which as its name implies, a regulator of angiogenesis in both normal and tumor cells (Yu et al., 2017). *PLXND1*, on the other hand, seems to have a protective role in both HNSC and LGG datasets, which is another interesting finding as the gene has been implicated in poor prognosis in several cancer types including prostate cancer as a Notch signaling target (Rehman et al., 2016).

The *SEMA4D-NRP2* pair appears to have a hazardous role on HNSC in this first bicluster for the DFI. The overall protective role of the Sema signaling pathway is evident in this bicluster for members across several cancer datasets; as different subsets of *SEMA3C*, *SEMA6B*, *SEMA4A*, *SEMA4B*, *SEMA4F*, *NRP1*, *PLXNA2*, and *PLXNA3* were found to have protective roles in COAD, ESCA, TGCT, HNSC, and THCA. Biclusters formed from two different input datasets can be compared within ClusterHR, if they are composed of an equal number of datasets.

5 CONCLUSION AND FUTURE PERSPECTIVES

While several genes with transcriptome-wide impact and significance from the exon aggregation analyses were indeed DE in the expected direction in tumor vs. normal samples, expressions of genes including *YTHDC1* in STAD, *CSNK1D* and *AP3M1* in GBM, and *PAFAH1B2* in LIHC did not differentiate between prognostic outcomes based on their expression levels in K-M analyses on UCSC Xena. This might indicate that these

genes that were not DE between tumor and normal cells, but nevertheless univariably significantly associated with survival rates in tumor cells in this study, are important for survival in normal cells also; and in certain cases, might be more crucial for the healthy functioning of cells than they are for tumor viability. Similarly, other genes in the top hazardous/protective gene lists for each dataset could be investigated further to determine whether they have a role on survival rates in their respective cancer datasets. Nevertheless, several genes including *GNA12* in STAD, *AKAP13* in LIHC, and *CARS1* and *RBMXL1* in BRCA which had not been previously implicated in the stated cancers were found to serve as potential biomarkers in each dataset through the analyses in Chapter 2. A possible future direction of this study would be to apply the same exon aggregation pipeline to the remaining TCGA datasets and determine the genes with highest impacts in each.

Methylation β-values of a gene are in the range 0-1 and indicate the extent of the presence of methyl groups on its DNA sequence; where a value of 0 indicates the gene is unmethylated and a value of 1 indicates the gene is fully methylated. Methylation of DNA is catalyzed by DNA methyltransferases (Dnmts), primarily on cytosines preceding a guanine nucleotide, namely CpG sites (Moore et al., 2013). Gene promoters lie mainly in DNA regions of about 1000 bp long with higher CpG density than the rest of the genome, which are known as CpG islands. Methylation of its CpG islands is thereby associated with the silencing of a gene. Hypomethylation (or undermethylation) in transcriptional control sequences and hypermethylation (or overmethylation) on CpG islands of gene regions have been both frequently observed in cancers (Ehrlich, 2002). Therefore, HR β-value signatures of gene sets in specific cancers might be useful in understanding which genes are more/less expressed within the set and the comparative role of each gene within the set with respect to the cancer's survival rate. Both SmulTCan and ClusterHR would aid researchers towards this goal by providing analysis with gene-level methylation β-values×100.

Somatic CNVs are frequently observed in cancer cells with respect to normal and can affect the expression levels of several genes at the same time, that are critical for cellular growth and survival. How and why CNVs arise in tumors, and how these alterations affect oncogenes and tumor suppressors remains an area of ongoing research. In a recent TCGA-PANCAN study, the most frequently amplified genes were determined as *CSMD3* and *FAM135B*, while the most frequently deleted gene were *PTBRD* and *CDKN2A* (Harbers et al., 2021). Additionally, chr8 which contains the MYC gene, was found to be the chromosome on which CNVs occur most frequently and ch17 was found to be amplified in $\simeq$ 20% of BRCA samples (Harbers et al., 2021). GDC employs the GISTIC2.0 algorithm for the detection of somatic CNVs (Mermel et al., 2011). Building multivariable CPH models with CNVs of gene sets using SmulTCan or ClusterHR created in this study would help understand how each gene's presence/absence or its dosage affects survival in a cancer and may shed light onto its molecular role within the gene set and guide researchers in therapeutics studies for the cancer.

Axon guidance ligand-receptor pairs with prognostic significance in cancer have been identified through clustering analyses with ClusterHR, which contribute to understanding their involvement in disease. Slit-Robo ligand-receptor pairs had been previously

identified in various aspects of neural development. For instance, the *SLIT2-ROBO3* pair is known to be involved in gonadotropin-releasing hormone (GnRH) neuron migration from the nasal placode to the forebrain (Cariboni et al., 2012), while the *SLIT1-ROBO2* pair mediates trigeminal ganglion assembly (Shiau et al., 2008), the largest of the cranial sensory ganglia responsible for facial sensation of touch, pain and taste. A non-neuronal function for a Slit-Robo pair was also identified in a study involving mesenchymal stem cells (MSCs), where the *SLIT3-ROBO4* pair was found to promote endothelial vascularization (Paul et al., 2013). Using ClusterHR, these three Slit-Robo pairs were found to be involved in cancer prognosis, including the protective role of *SLIT2-ROBO3* in PAAD for DSS.

Several netrin-receptor pairs, which had not been investigated in detail in cancer prognosis, were also identified in DSS using ClusterHR multivariably, such as the protective role of the *NTNG2-UNC5A* pair in LGG. Additionally, the identification of the *SEMA4D-NRP2* pair to have a hazardous role in HNSC for the DFI is an interesting outcome of the biclustering analysis with Semas-receptors. While Sema4s are known to promote angiogenesis and invasive growth, they are known not to interact with neuropilins. However, they can signal through similar downstream pathways; Sema4D can signal through PlexinB1 to Rac1 and R-Ras, while neuropilins can also signal to the same pathways through different mechanisms via plexinAs. Another interesting finding from this same bicluster of the DFI was the antagonistic pair *SEMA4A-PLXND1* in TGCT, where *SEMA4A* had a protective role while *PLXND1* was hazardous. The signaling mechanism of this relationship in TGCT could be further investigated, since the pair is known to inhibit angiogenesis in endothelial cells via VEGFR2 (Ito & Kumanogoh, 2016; Zhou et al., 2008).

Both ClusterHR and SmulTCan are powerful tools for understanding survival profiles of sets of genes/miRNAs/CNVs/β-values with datasets of TCGA-PANCAN. They can either be used as individual apps separately or complementarily to each other to build CPH models of input sets and analyze them; either in detail with a single dataset in SmulTCan or with heatmaps across TCGA-PANCAN in ClusterHR, once users have uploaded the TSV file they obtained from UCSC Xena onto the apps for their genes/miRNAs of interest. For SmulTCan, it is possible to select best subsets also of exon expressions from its “BeSS” and “glmnet” tabs. A possible improvement on the apps requires the annotation of TCGA-PANCAN samples with respect to their cancer and sample types and the embedding of UCSC Xena gene expression values. This would allow users to determine their input gene sets directly from the app interface and would eliminate the need to download expression files from UCSC Xena prior to analysis with the apps. Hopefully, both apps are going to aid cancer researchers in their analyses and ultimately experiments as interactive and user-friendly tools, to determine the comparative effect of each gene on survival within their gene sets.

References

Abdul Aziz, N. A., Mokhtar, N. M., Harun, R., Mollah, M. M., Mohamed Rose, I., Sagap, I., Mohd Tamil, A., Wan Ngah, W. Z., & Jamal, R. (2016). A 19-Gene expression signature as a predictor of survival in colorectal cancer. *BMC Med Genomics*, *9*(1), 58. https://doi.org/10.1186/s12920-016-0218-1

Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., Tamez-Pena, J. G., & Trevino, V. (2013). SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One*, *8*(9), e74250. https://doi.org/10.1371/journal.pone.0074250

Aguirre-Gamboa, R., & Trevino, V. (2014). SurvMicro: assessment of miRNA-based prognostic signatures for cancer clinical outcomes by multivariate survival analysis. *Bioinformatics*, *30*(11), 1630-1632. https://doi.org/10.1093/bioinformatics/btu087

Ajani, J. A., Lee, J., Sano, T., Janjigian, Y. Y., Fan, D., & Song, S. (2017). Gastric adenocarcinoma. *Nat Rev Dis Primers*, *3*, 17036. https://doi.org/10.1038/nrdp.2017.36

Alto, L. T., & Terman, J. R. (2017). Semaphorins and their Signaling Mechanisms. *Methods Mol Biol*, *1493*, 1-25. https://doi.org/10.1007/978-1-4939-6448-2_1

Backlund, L. M., Nilsson, B. R., Liu, L., Ichimura, K., & Collins, V. P. (2005). Mutations in Rb1 pathway-related genes are associated with poor prognosis in anaplastic astrocytomas. *Br J Cancer*, *93*(1), 124-130. https://doi.org/10.1038/sj.bjc.6602661

Badour, K., McGavin, M. K., Zhang, J., Freeman, S., Vieira, C., Filipp, D., Julius, M., Mills, G. B., & Siminovitch, K. A. (2007). Interaction of the Wiskott-Aldrich syndrome protein with sorting nexin 9 is required for CD28 endocytosis and cosignaling in T cells. *Proc Natl Acad Sci U S A*, *104*(5), 1593-1598. https://doi.org/10.1073/pnas.0610543104

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Kwok-Shing Ng, P., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., Rubio-Perez, C., Nagarajan, N., Cortes-Ciriano, I., Zhou, D. C., Liang, W. W., Hess, J. M., Yellapantula, V. D., Tamborero, D., Gonzalez-Perez, A., Suphavilai, C., Ko, J. Y., Khurana, E., Park, P. J., Van Allen, E. M., Liang, H., Group, M. C. W., Cancer Genome Atlas Research, N., Lawrence, M. S., Godzik, A., Lopez-Bigas, N., Stuart, J., Wheeler, D., Getz, G., Chen, K., Lazar, A. J., Mills, G. B., Karchin, R., & Ding, L. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *174*(4), 1034-1035. https://doi.org/10.1016/j.cell.2018.07.034

Bankhead, A., 3rd, McMaster, T., Wang, Y., Boonstra, P. S., & Palmbos, P. L. (2020). TP63 isoform expression is linked with distinct clinical outcomes in cancer. *EBioMedicine*, *51*, 102561. https://doi.org/10.1016/j.ebiom.2019.11.022

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*, *41*(Database issue), D991-995. https://doi.org/10.1093/nar/gks1193

Bend, R., Cohen, L., Carter, M. T., Lyons, M. J., Niyazov, D., Mikati, M. A., Rojas, S. K., Person, R. E., Si, Y., Wentzensen, I. M., Regeneron Genetics, C., Torti, E., Lee, J. A., Boycott, K. M., Basel-Salmon, L., Ferreira, C. R., & Gonzaga-Jauregui, C. (2020). Phenotype and mutation expansion of the PTPN23 associated disorder characterized by neurodevelopmental delay and structural brain abnormalities. *Eur J Hum Genet*, *28*(1), 76-87. https://doi.org/10.1038/s41431-019-0487-1

Beuchel, C., Kirsten, H., Ceglarek, U., & Scholz, M. (2021). Metabolite-Investigator: an integrated user-friendly workflow for metabolomics multi-study analysis. *Bioinformatics*, *37*(15), 2218-2220. https://doi.org/10.1093/bioinformatics/btaa967

Bischof, K., Knappskog, S., Hjelle, S. M., Stefansson, I., Woie, K., Salvesen, H. B., Gjertsen, B. T., & Bjorge, L. (2019). Influence of p53 Isoform Expression on Survival in High-Grade Serous Ovarian Cancers. *Sci Rep*, *9*(1), 5244. https://doi.org/10.1038/s41598-019-41706-z

Bjorklund, S. S., Panda, A., Kumar, S., Seiler, M., Robinson, D., Gheeya, J., Yao, M., Alnaes, G. I. G., Toppmeyer, D., Riis, M., Naume, B., Borresen-Dale, A. L., Kristensen, V. N., Ganesan, S., & Bhanot, G. (2017). Widespread alternative exon usage in clinically distinct subtypes of Invasive Ductal Carcinoma. *Sci Rep*, *7*(1), 5568. https://doi.org/10.1038/s41598-017-05537-0

Bogerd, H. P., Fridell, R. A., Madore, S., & Cullen, B. R. (1995). Identification of a novel cellular cofactor for the Rev/Rex class of retroviral regulatory proteins. *Cell*, *82*(3), 485-494. https://doi.org/10.1016/0092-8674(95)90437-9

Bonnal, S. C., Lopez-Oreja, I., & Valcarcel, J. (2020). Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol*, *17*(8), 457-474. https://doi.org/10.1038/s41571-020-0350-x

Borcherding, N., Bormann, N. L., Voigt, A. P., & Zhang, W. (2018). TRGAted: A web tool for survival analysis using protein data in the Cancer Genome Atlas. *F1000Res*, *7*, 1235. https://doi.org/10.12688/f1000research.15789.2

Brembilla, A., Olland, A., Puyraveau, M., Massard, G., Mauny, F., & Falcoz, P. E. (2018). Use of the Cox regression analysis in thoracic surgical research. *J Thorac Dis*, *10*(6), 3891-3896. https://doi.org/10.21037/jtd.2018.06.15

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*, 94. https://doi.org/10.1186/1471-2105-11-94

Cariboni, A., Andrews, W. D., Memi, F., Ypsilanti, A. R., Zelina, P., Chedotal, A., & Parnavelas, J. G. (2012). Slit2 and Robo3 modulate the migration of GnRH-secreting neurons. *Development*, *139*(18), 3326-3331. https://doi.org/10.1242/dev.079418

Carlson, M. (2020). *org.Hs.eg.db: Genome wide annotation for Human*. In https://bioconductor.org/packages/3.11/data/annotation/html/org.Hs.eg.db.html

Carlson, M., & Maintainer, B. P. (2015a). *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. In https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg19.knownGene.html

Carlson, M., & Maintainer, B. P. (2015b). *TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts: Annotation package for TxDb object(s)*. In https://bioconductor.org/packages/3.13/data/annotation/html/TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts.html

Caunt, M., Mak, J., Liang, W. C., Stawicki, S., Pan, Q., Tong, R. K., Kowalski, J., Ho, C., Reslan, H. B., Ross, J., Berry, L., Kasman, I., Zlot, C., Cheng, Z., Le Couter, J., Filvaroff, E. H., Plowman, G., Peale, F., French, D., Carano, R., Koch, A. W., Wu, Y., Watts, R. J., Tessier-Lavigne, M., & Bagri, A. (2008). Blocking neuropilin-2 function inhibits tumor cell metastasis. *Cancer Cell*, *13*(4), 331-342. https://doi.org/10.1016/j.ccr.2008.01.029

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web Application Framework for r*. In https://CRAN.R-project.org/package=shiny

Chen, J., & Weiss, W. A. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, *34*(1), 1-14. https://doi.org/10.1038/onc.2013.570

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, *8*, 93-103. https://www.ncbi.nlm.nih.gov/pubmed/10977070

Cinar, O., & Viechtbauer, W. (2021). *poolr: Methods for Pooling P-Values from (Dependent) Tests*. In https://CRAN.R-project.org/package=poolr

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*, *44*(8), e71. https://doi.org/10.1093/nar/gkv1507

Cools, J., Wlodarska, I., Somers, R., Mentens, N., Pedeutour, F., Maes, B., De Wolf-Peeters, C., Pauwels, P., Hagemeijer, A., & Marynen, P. (2002). Identification of novel fusion partners of ALK, the anaplastic lymphoma kinase, in anaplastic large-cell lymphoma and inflammatory myofibroblastic tumor. *Genes Chromosomes Cancer*, *34*(4), 354-362. https://doi.org/10.1002/gcc.10033

Datta, S., & Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, *19*(4), 459-466. https://doi.org/10.1093/bioinformatics/btg025

Denkert, C., Budczies, J., Darb-Esfahani, S., Gyorffy, B., Sehouli, J., Konsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A. C., Muller, B. M., Dietel, M., & Lage, H. (2009). A prognostic gene expression index in ovarian cancer - validation across different independent data sets. *J Pathol*, *218*(2), 273-280. https://doi.org/10.1002/path.2547

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., Jaffrezic, F., & French StatOmique, C. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, *14*(6), 671-683. https://doi.org/10.1093/bib/bbs046

Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. *Oncogene*, *21*(35), 5400-5413. https://doi.org/10.1038/sj.onc.1205651

Ehsani, R., & Drablos, F. (2020). Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Inform*, *19*, 1176935120965542. https://doi.org/10.1177/1176935120965542

Evans, C., Hardin, J., & Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform*, *19*(5), 776-792. https://doi.org/10.1093/bib/bbx008

Finci, L. I., Kruger, N., Sun, X., Zhang, J., Chegkazi, M., Wu, Y., Schenk, G., Mertens, H. D. T., Svergun, D. I., Zhang, Y., Wang, J. H., & Meijers, R. (2014). The crystal structure of netrin-1 in complex with DCC reveals the bifunctionality of netrin-1 as a guidance cue. *Neuron*, *83*(4), 839-849. https://doi.org/10.1016/j.neuron.2014.07.010

Flower, C. T., Chen, L., Jung, H. J., Raghuram, V., Knepper, M. A., & Yang, C. R. (2020). An integrative proteogenomics approach reveals peptides encoded by annotated lincRNA in the mouse kidney inner medulla. *Physiol Genomics*, *52*(10), 485-491. https://doi.org/10.1152/physiolgenomics.00048.2020

Fujiwara, N., Liu, P. H., Athuluri-Divakar, S. K., Zhu, S., & Hoshida, Y. (2019). Risk Factors of Hepatocellular Carcinoma for Precision Personalized Care. In Y. Hoshida (Ed.), *Hepatocellular Carcinoma: Translational Precision Medicine Approaches* (pp. 3-25). https://doi.org/10.1007/978-3-030-21540-8_1

Galili, T., O'Callaghan, A., Sidi, J., & Sievert, C. (2018). heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*, *34*(9), 1600-1602. https://doi.org/10.1093/bioinformatics/btx657

Gao, B., Zhu, J., Negi, S., Zhang, X., Gyoneva, S., Casey, F., Wei, R., & Zhang, B. (2021). Quickomics: exploring omics data in an intuitive, interactive and informative manner. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab255

Gentles, A. J., Bratman, S. V., Lee, L. J., Harris, J. P., Feng, W., Nair, R. V., Shultz, D. B., Nair, V. S., Hoang, C. D., West, R. B., Plevritis, S. K., Alizadeh, A. A., & Diehn, M. (2015). Integrating Tumor and Stromal Gene Expression Signatures With Clinical Indices for Survival Stratification of Early-Stage Non-Small Cell Lung Cancer. *J Natl Cancer Inst*, *107*(10). https://doi.org/10.1093/jnci/djv211

Gerds, T. A., & Ozenne, B. (2020). *riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks.* In https://CRAN.R-project.org/package=riskRegression

Goldman, M. J., Craft, B., Hastie, M., Repecka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., Zhu, J., & Haussler, D. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*, *38*(6), 675-678. https://doi.org/10.1038/s41587-020-0546-8

Goode, E. L., Maurer, M. J., Sellers, T. A., Phelan, C. M., Kalli, K. R., Fridley, B. L., Vierkant, R. A., Armasu, S. M., White, K. L., Keeney, G. L., Cliby, W. A., Rider, D. N., Kelemen, L. E., Jones, M. B., Peethambaram, P. P., Lancaster, J. M., Olson, J. E., Schildkraut, J. M., Cunningham, J. M., & Hartmann, L. C. (2010). Inherited determinants of ovarian cancer survival. *Clin Cancer Res*, *16*(3), 995-1007. https://doi.org/10.1158/1078-0432.CCR-09-2553

Goswami, C. P., & Nakshatri, H. (2014). PROGgeneV2: enhancements on the existing database. *BMC Cancer*, *14*, 970. https://doi.org/10.1186/1471-2407-14-970

Grone, J., Doebler, O., Loddenkemper, C., Hotz, B., Buhr, H. J., & Bhargava, S. (2006). Robo1/Robo4: differential expression of angiogenic markers in colorectal cancer. *Oncol Rep*, *15*(6), 1437-1443. https://www.ncbi.nlm.nih.gov/pubmed/16685377

Grupe, A., Li, Y., Rowland, C., Nowotny, P., Hinrichs, A. L., Smemo, S., Kauwe, J. S., Maxwell, T. J., Cherny, S., Doil, L., Tacey, K., van Luchene, R., Myers, A., Wavrant-De Vrieze, F., Kaleem, M., Hollingworth, P., Jehu, L., Foy, C., Archer, N., Hamilton, G., Holmans, P., Morris, C. M., Catanese, J., Sninsky, J., White, T. J., Powell, J., Hardy, J., O'Donovan, M., Lovestone, S., Jones, L., Morris, J. C., Thal, L., Owen, M., Williams, J., & Goate, A. (2006). A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease. *Am J Hum Genet*, *78*(1), 78-88. https://doi.org/10.1086/498851

Gudenas, B. L., & Wang, L. (2018). Prediction of LncRNA Subcellular Localization with Deep Learning from Sequence Features. *Sci Rep*, *8*(1), 16385. https://doi.org/10.1038/s41598-018-34708-w

Gyorffy, B. (2021). Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput Struct Biotechnol J*, *19*, 4101-4109. https://doi.org/10.1016/j.csbj.2021.07.014

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, *250*(4988), 1684-1689. https://doi.org/10.1126/science.2270482

Han, S., Liu, Y., Cai, S. J., Qian, M., Ding, J., Larion, M., Gilbert, M. R., & Yang, C. (2020). IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. *Br J Cancer*, *122*(11), 1580-1589. https://doi.org/10.1038/s41416-020-0814-x

Hao, W., Yu, M., Lin, J., Liu, B., Xing, H., Yang, J., Sun, D., Chen, F., Jiang, M., Tang, C., Zhang, X., Zhao, Y., & Zhu, Y. (2020). The pan-cancer landscape of netrin family reveals potential oncogenic biomarkers. *Sci Rep*, *10*(1), 5224. https://doi.org/10.1038/s41598-020-62117-5

Harbeck, N. (2020). Breast cancer is a systemic disease optimally treated by a multidisciplinary team. *Nat Rev Dis Primers*, *6*(1), 30. https://doi.org/10.1038/s41572-020-0167-z
Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast cancer. *Nat Rev Dis Primers*, *5*(1), 66. https://doi.org/10.1038/s41572-019-0111-2
Harbers, L., Agostini, F., Nicos, M., Poddighe, D., Bienko, M., & Crosetto, N. (2021). Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Front Oncol*, *11*, 700568. https://doi.org/10.3389/fonc.2021.700568
Harrell Jr, F. E. (2021). *Rms: Regression Modeling Strategies*. In https://CRAN.R-project.org/package=rms
Hashimoto, R., Ohi, K., Okada, T., Yasuda, Y., Yamamori, H., Hori, H., Hikita, T., Taya, S., Saitoh, O., Kosuga, A., Tatsumi, M., Kamijima, K., Kaibuchi, K., Takeda, M., & Kunugi, H. (2009). Association analysis between schizophrenia and the AP-3 complex genes. *Neurosci Res*, *65*(1), 113-115. https://doi.org/10.1016/j.neures.2009.05.008
Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholson, R. I., & Griffiths, K. (1982). A prognostic index in primary breast cancer. *Br J Cancer*, *45*(3), 361-366. https://doi.org/10.1038/bjc.1982.62
He, Z., Bammann, H., Han, D., Xie, G., & Khaitovich, P. (2014). Conserved expression of lincRNA during human and macaque prefrontal cortex development and maturation. *RNA*, *20*(7), 1103-1111. https://doi.org/10.1261/rna.043075.113
Herrero, J., Valencia, A., & Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, *17*(2), 126-136. https://doi.org/10.1093/bioinformatics/17.2.126
Ito, D., & Kumanogoh, A. (2016). The role of Sema4A in angiogenesis, immune responses, carcinogenesis, and retinal systems. *Cell Adh Migr*, *10*(6), 692-699. https://doi.org/10.1080/19336918.2016.1215785
Jian, S. L., Hsieh, H. Y., Liao, C. T., Yen, T. C., Nien, S. W., Cheng, A. J., & Juang, J. L. (2013). Galpha(1)(2) drives invasion of oral squamous cell carcinoma through up-regulation of proinflammatory cytokines. *PLoS One*, *8*(6), e66133. https://doi.org/10.1371/journal.pone.0066133
Jiang, M. C., Ni, J. J., Cui, W. Y., Wang, B. Y., & Zhuo, W. (2019). Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res*, *9*(7), 1354-1366. https://www.ncbi.nlm.nih.gov/pubmed/31392074
Jones, C. A., London, N. R., Chen, H., Park, K. W., Sauvaget, D., Stockton, R. A., Wythe, J. D., Suh, W., Larrieu-Lahargue, F., Mukouyama, Y. S., Lindblom, P., Seth, P., Frias, A., Nishiya, N., Ginsberg, M. H., Gerhardt, H., Zhang, K., & Li, D. Y. (2008). Robo4 stabilizes the vascular network by inhibiting pathologic angiogenesis and endothelial hyperpermeability. *Nat Med*, *14*(4), 448-453. https://doi.org/10.1038/nm1742
Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch, F., De Troyer, E., & Leon, S. (2021). *biclust: BiCluster Algorithms*. In https://CRAN.R-project.org/package=biclust
Karatzas, E., Gkonta, M., Hotova, J., Baltoumas, F. A., Kontou, P. I., Bobotsis, C. J., Bagos, P. G., & Pavlopoulos, G. A. (2021). VICTOR: A visual analytics web application for comparing cluster sets. *Comput Biol Med*, *135*, 104557. https://doi.org/10.1016/j.compbiomed.2021.104557
Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Brief Bioinform*, *22*(1), 393-415. https://doi.org/10.1093/bib/bbz170
Kassambara, A., Kosinski, M., & Biecek , P. (2020). *Survminer: Drawing Survival Curves Using 'ggplot2'*. In https://CRAN.R-project.org/package=survminer/
Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res*, *47*(D1), D155-D162. https://doi.org/10.1093/nar/gky1141
Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J. P., Heese, O., Reifenberger, G., Weller, M., Schackert, G., & German Glioma, N. (2007). Long-term survival with glioblastoma multiforme. *Brain*, *130*(Pt 10), 2596-2606. https://doi.org/10.1093/brain/awm204
Kuhn, M. (2021). *caret: Classification and Regression Training*. In https://CRAN.R-project.org/package=caret
Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*, *9*(1), 1366. https://doi.org/10.1038/s41467-018-03751-6
Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet*, *19*(5), 325. https://doi.org/10.1038/nrg.2018.8
Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lee, Y. J., Seo, H. W., Baek, J. H., Lim, S. H., Hwang, S. G., & Kim, E. H. (2020). Gene expression profiling of glioblastoma cell lines depending on TP53 status after tumor-treating fields (TTFields) treatment. *Sci Rep*, *10*(1), 12272. https://doi.org/10.1038/s41598-020-68473-6

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research, N., & Hu, H. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, *173*(2), 400-416 e411. https://doi.org/10.1016/j.cell.2018.02.052

Liu, K., Gao, L., Ma, X., Huang, J. J., Chen, J., Zeng, L., Ashby, C. R., Jr., Zou, C., & Chen, Z. S. (2020). Long non-coding RNAs regulate drug resistance in cancer. *Mol Cancer*, *19*(1), 54. https://doi.org/10.1186/s12943-020-01162-0

Liu, R., Holik, A. Z., Su, S., Jansz, N., Chen, K., Leong, H. S., Blewitt, M. E., Asselin-Labat, M. L., Smyth, G. K., & Ritchie, M. E. (2015). Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res*, *43*(15), e97. https://doi.org/10.1093/nar/gkv412

Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., Lencioni, R., Koike, K., Zucman-Rossi, J., & Finn, R. S. (2021). Hepatocellular carcinoma. *Nat Rev Dis Primers*, *7*(1), 6. https://doi.org/10.1038/s41572-020-00240-3

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Ma, C., Guo, Y., Zhang, Y., Duo, A., Jia, Y., Liu, C., & Li, B. (2018). PAFAH1B2 is a HIF1a target gene and promotes metastasis in pancreatic cancer. *Biochem Biophys Res Commun*, *501*(3), 654-660. https://doi.org/10.1016/j.bbrc.2018.05.039

Madigan, D. (2013). *StatW2025/W4025: methods in applied statistics*. Columbia University. http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf

Mapes, D. (2014). *Living with Stage 4: The breast cancer no one understands*. FRED HUTCH NEWS SERVICE. https://www.fredhutch.org/en/news/center-news/2014/10/stage-4-metastatic-misunderstood-breast-cancer.html

Marden, J. H. (2008). Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. *Heredity (Edinb)*, *100*(2), 111-120. https://doi.org/10.1038/sj.hdy.6800904

Mehmood, R., El-Ashram, S., Bie, R., Dawood, H., & Kos, A. (2017). Clustering by fast search and merge of local density peaks for gene expression microarray data. *Sci Rep*, *7*, 45602. https://doi.org/10.1038/srep45602

Meijers, R., Smock, R. G., Zhang, Y., & Wang, J. H. (2020). Netrin Synergizes Signaling and Adhesion through DCC. *Trends Biochem Sci*, *45*(1), 6-12. https://doi.org/10.1016/j.tibs.2019.10.005

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Mogilyansky, E., & Rigoutsos, I. (2013). The miR-17/92 cluster: a comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death Differ*, *20*(12), 1603-1614. https://doi.org/10.1038/cdd.2013.125

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, *38*(1), 23-38. https://doi.org/10.1038/npp.2012.112

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, *5*(7), 621-628. https://doi.org/10.1038/nmeth.1226

Murali, T. M., & Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput*, 77-88. https://www.ncbi.nlm.nih.gov/pubmed/12603019

NCI. (2021). *TCGA Timeline & Milestones*. National Cancer Institute at the National Institutes of Health. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/timeline

Oliveira, C., de Bruin, J., Nabais, S., Ligtenberg, M., Moutinho, C., Nagengast, F. M., Seruca, R., van Krieken, H., & Carneiro, F. (2004). Intragenic deletion of CDH1 as the inactivating mechanism of the wild-type allele in an HDGC tumour. *Oncogene*, *23*(12), 2236-2240. https://doi.org/10.1038/sj.onc.1207335

Ortiz, A. B., Garcia, D., Vicente, Y., Palka, M., Bellas, C., & Martin, P. (2017). Prognostic significance of cyclin D1 protein expression and gene amplification in invasive breast carcinoma. *PLoS One*, *12*(11), e0188068. https://doi.org/10.1371/journal.pone.0188068

Ozhan, A., Tombaz, M., & Konu, O. (2021). SmulTCan: A Shiny application for multivariable survival analysis of TCGA data with gene sets. *Comput Biol Med*, *137*, 104793. https://doi.org/10.1016/j.compbiomed.2021.104793

Park, K. W., Morrison, C. M., Sorensen, L. K., Jones, C. A., Rao, Y., Chien, C. B., Wu, J. Y., Urness, L. D., & Li, D. Y. (2003). Robo4 is a vascular-specific receptor that inhibits endothelial migration. *Dev Biol*, *261*(1), 251-267. https://doi.org/10.1016/s0012-1606(03)00258-6

Paul, J. D., Coulombe, K. L. K., Toth, P. T., Zhang, Y., Marsboom, G., Bindokas, V. P., Smith, D. W., Murry, C. E., & Rehman, J. (2013). SLIT3-ROBO4 activation promotes vascular network formation in human engineered tissue and angiogenesis in vivo. *J Mol Cell Cardiol*, *64*, 124-131. https://doi.org/10.1016/j.yjmcc.2013.09.005

Peng, Y., & Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*, *1*, 15004. https://doi.org/10.1038/sigtrans.2015.4

Pizzuti, C., & Rombo, S. E. (2014). Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, *30*(10), 1343-1352. https://doi.org/10.1093/bioinformatics/btu034

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., & Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, *22*(9), 1122-1129. https://doi.org/10.1093/bioinformatics/btl060

Prieto, C., Nguyen, D. T. T., Liu, Z., Wheat, J., Perez, A., Gourkanti, S., Chou, T., Barin, E., Velleca, A., Rohwetter, T., Chow, A., Taggart, J., Savino, A. M., Hoskova, K., Dhodapkar, M., Schurer, A., Barlowe, T. S., Vu, L. P., Leslie, C., Steidl, U., Rabadan, R., & Kharas, M. G. (2021). Transcriptional control of CBX5 by the RNA binding proteins RBMX and RBMXL1 maintains chromatin state in myeloid leukemia. *Nat Cancer*, *2*, 741-757. https://doi.org/10.1038/s43018-021-00220-w

Rajasekharan, S., & Kennedy, T. E. (2009). The netrin protein family. *Genome Biol*, *10*(9), 239. https://doi.org/10.1186/gb-2009-10-9-239

Ransohoff, J. D., Wei, Y., & Khavari, P. A. (2018). The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol*, *19*(3), 143-157. https://doi.org/10.1038/nrm.2017.104

Rehman, M., Gurrapu, S., Cagnoni, G., Capparuccia, L., & Tamagnone, L. (2016). PlexinD1 Is a Novel Transcriptional Target and Effector of Notch Signaling in Cancer Cells. *PLoS One*, *11*(10), e0164660. https://doi.org/10.1371/journal.pone.0164660

Rezniczek, G. A., Grunwald, C., Hilal, Z., Scheich, J., Reifenberger, G., Tannapfel, A., & Tempfer, C. B. (2019). ROBO1 Expression in Metastasizing Breast and Ovarian Cancer: SLIT2-induced Chemotaxis Requires Heparan Sulfates (Heparin). *Anticancer Res*, *39*(3), 1267-1273. https://doi.org/10.21873/anticanres.13237

Rhee, J., Mahfooz, N. S., Arregui, C., Lilien, J., Balsamo, J., & VanBerkum, M. F. (2002). Activation of the repulsive receptor Roundabout inhibits N-cadherin-mediated cell adhesion. *Nat Cell Biol*, *4*(10), 798-805. https://doi.org/10.1038/ncb858

Riddick, G., Kotliarova, S., Rodriguez, V., Kim, H. S., Linkous, A., Storaska, A. J., Ahn, S., Walling, J., Belova, G., & Fine, H. A. (2017). A Core Regulatory Circuit in Glioblastoma Stem Cells Links MAPK Activation to a Transcriptional Program of Neural Stem Cell Identity. *Sci Rep*, *7*, 43605. https://doi.org/10.1038/srep43605

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

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

Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. https://data.princeton.edu/wws509/notes/

Roundtree, I. A., Luo, G. Z., Zhang, Z., Wang, X., Zhou, T., Cui, Y., Sha, J., Huang, X., Guerrero, L., Xie, P., He, E., Shen, B., & He, C. (2017). YTHDC1 mediates nuclear export of N(6)-methyladenosine methylated mRNAs. *Elife*, *6*. https://doi.org/10.7554/eLife.31311

Sadeque, A., Serao, N. V., Southey, B. R., Delfino, K. R., & Rodriguez-Zas, S. L. (2012). Identification and characterization of alternative exon usage linked glioblastoma multiforme survival. *BMC Med Genomics*, *5*, 59. https://doi.org/10.1186/1755-8794-5-59

Saeed, S., Bonnefond, A., Tamanini, F., Mirza, M. U., Manzoor, J., Janjua, Q. M., Din, S. M., Gaitan, J., Milochau, A., Durand, E., Vaillant, E., Haseeb, A., De Graeve, F., Rabearivelo, I., Sand, O., Queniat, G., Boutry, R., Schott, D. A., Ayesha, H., Ali, M., Khan, W. I., Butt, T. A., Rinne, T., Stumpel, C., Abderrahmani, A., Lang, J., Arslan, M., & Froguel, P. (2018). Loss-of-function mutations in ADCY3 cause monogenic severe obesity. *Nat Genet*, *50*(2), 175-179. https://doi.org/10.1038/s41588-017-0023-6

Sakurai, A., Doci, C. L., & Gutkind, J. S. (2012). Semaphorin signaling in angiogenesis, lymphangiogenesis and cancer. *Cell Res*, *22*(1), 23-32. https://doi.org/10.1038/cr.2011.198

Schramm, A., Schowe, B., Fielitz, K., Heilmann, M., Martin, M., Marschall, T., Koster, J., Vandesompele, J., Vermeulen, J., de Preter, K., Koster, J., Versteeg, R., Noguera, R., Speleman, F., Rahmann, S., Eggert, A., Morik, K., & Schulte, J. H. (2012). Exon-level expression analyses identify MYCN and NTRK1 as major determinants of alternative exon usage and robustly predict primary neuroblastoma outcome. *Br J Cancer*, *107*(8), 1409-1417. https://doi.org/10.1038/bjc.2012.391

Sha, Y., Phan, J. H., & Wang, M. D. (2015). Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,

Shen, S., Wang, Y., Wang, C., Wu, Y. N., & Xing, Y. (2016). SURVIV for survival analysis of mRNA isoform variation. *Nat Commun*, *7*, 11548. https://doi.org/10.1038/ncomms11548

Shi, D., Liang, L., Zheng, H., Cai, G., Li, X., Xu, Y., & Cai, S. (2017). Silencing of long non-coding RNA SBDSP1 suppresses tumor growth and invasion in colorectal cancer. *Biomed Pharmacother*, *85*, 355-361. https://doi.org/10.1016/j.biopha.2016.11.036

Shiau, C. E., Lwigale, P. Y., Das, R. M., Wilson, S. A., & Bronner-Fraser, M. (2008). Robo2-Slit1 dependent cell-cell interactions mediate assembly of the trigeminal ganglion. *Nat Neurosci*, *11*(3), 269-276. https://doi.org/10.1038/nn2051

Shin, S. K., Nagasaka, T., Jung, B. H., Matsubara, N., Kim, W. H., Carethers, J. M., Boland, C. R., & Goel, A. (2007). Epigenetic and genetic alterations in Netrin-1 receptors UNC5C and DCC in human colon cancer. *Gastroenterology*, *133*(6), 1849-1857. https://doi.org/10.1053/j.gastro.2007.08.074

Shinmura, K., Goto, M., Tao, H., Shimizu, S., Otsuki, Y., Kobayashi, H., Ushida, S., Suzuki, K., Tsuneyoshi, T., & Sugimura, H. (2005). A novel STK11 germline mutation in two siblings with Peutz-Jeghers syndrome complicated by primary gastric cancer. *Clin Genet*, *67*(1), 81-86. https://doi.org/10.1111/j.1399-0004.2005.00380.x

Si, W., Shen, J., Zheng, H., & Fan, W. (2019). The role and mechanisms of action of microRNAs in cancer drug resistance. *Clin Epigenetics*, *11*(1), 25. https://doi.org/10.1186/s13148-018-0587-8

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, *39*(5), 1-13. https://www.jstatsoft.org/v39/i05/

Squire, L. R. (2013). *Fundamental neuroscience* (4th ed.). Elsevier/Academic Press.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat Rev Genet*, *20*(11), 631-656. https://doi.org/10.1038/s41576-019-0150-2

Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*, *22*(2), 96-118. https://doi.org/10.1038/s41580-020-00315-9

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics*, *54*, 1 30 31-31 30 33. https://doi.org/10.1002/cpbi.5

Stoter, M., Bamberger, A. M., Aslan, B., Kurth, M., Speidel, D., Loning, T., Frank, H. G., Kaufmann, P., Lohler, J., Henne-Bruns, D., Deppert, W., & Knippschild, U. (2005). Inhibition of casein kinase I delta alters mitotic spindle formation and induces apoptosis in trophoblast cells. *Oncogene*, *24*(54), 7964-7975. https://doi.org/10.1038/sj.onc.1208941

Tabibu, S., Vinod, P. K., & Jawahar, C. V. (2019). Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Sci Rep*, *9*(1), 10509. https://doi.org/10.1038/s41598-019-46718-3

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*, *45*(W1), W98-W102. https://doi.org/10.1093/nar/gkx247

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data : extending the Cox model*. Springer,. https://doi.org/10.1007/978-1-4757-3294-8

Tian, T., Zhang, L., Tang, K., Wang, A., Wang, J., Wang, J., Wang, F., Wang, W., & Ma, X. (2020). SEMA3A Exon 9 Expression Is a Potential Prognostic Marker of Unfavorable Recurrence-Free Survival in Patients with Tongue Squamous Cell Carcinoma. *DNA Cell Biol*, *39*(4), 555-562. https://doi.org/10.1089/dna.2019.5109

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J R Statist Soc B (Statistical Methodology)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tijsterman, M., & Plasterk, R. H. (2004). Dicers at RISC; the mechanism of RNAi. *Cell*, *117*(1), 1-3. https://doi.org/10.1016/s0092-8674(04)00293-4

Tong, M., Jun, T., Nie, Y., Hao, J., & Fan, D. (2019). The Role of the Slit/Robo Signaling Pathway. *J Cancer*, *10*(12), 2694-2705. https://doi.org/10.7150/jca.31877
Trisciuoglio, D., Tupone, M. G., Desideri, M., Di Martile, M., Gabellini, C., Buglioni, S., Pallocca, M., Alessandrini, G., D'Aguanno, S., & Del Bufalo, D. (2017). BCL-XL overexpression promotes tumor progression-associated properties. *Cell Death Dis*, *8*(12), 3216. https://doi.org/10.1038/s41419-017-0055-y
Turner, H. L., Bailey, T. C., Krzanowski, W. J., & Hemingway, C. A. (2005). Biclustering models for structured microarray data. *IEEE/ACM Trans Comput Biol Bioinform*, *2*(4), 316-329. https://doi.org/10.1109/TCBB.2005.49
Unterluggauer, J. J., Prochazka, K., Tomazic, P. V., Huber, H. J., Seeboeck, R., Fechter, K., Steinbauer, E., Gruber, V., Feichtinger, J., Pichler, M., Weniger, M. A., Kuppers, R., Sill, H., Schicho, R., Neumeister, P., Beham-Schmid, C., Deutsch, A. J. A., & Haybaeck, J. (2018). Expression profile of translation initiation factor eIF2B5 in diffuse large B-cell lymphoma and its correlation to clinical outcome. *Blood Cancer J*, *8*(9), 79. https://doi.org/10.1038/s41408-018-0112-5
Valentini, G. (2009). *Bioinformatics methods*. University of Milan. https://homes.di.unimi.it/valentini/MB0910.html
Wang, H., Wang, Y., Xie, S., Liu, Y., & Xie, Z. (2017). Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Res*, *45*(5), 2786-2796. https://doi.org/10.1093/nar/gkw909
Wang, J., Keusters, W. R., Wen, L., & Leeflang, M. M. G. (2021). IPDmada: An R Shiny tool for analyzing and visualizing individual patient data meta-analyses of diagnostic test accuracy. *Res Synth Methods*, *12*(1), 45-54. https://doi.org/10.1002/jrsm.1444
Wang, S. L., X. (2019). The UCSCXenaTools r Package: A Toolkit for Accessing Genomics Data from UCSC Xena Platform, from Cancer Multi-Omics to Single-Cell RNA-Seq. *Journal of Open Source Software*, *4*(40), 4750. https://doi.org/https://doi.org/10.21105/joss.01627
Wang, Y., Goodison, S., Li, X., & Hu, H. (2017). Prognostic cancer gene signatures share common regulatory motifs. *Sci Rep*, *7*(1), 4750. https://doi.org/10.1038/s41598-017-05035-3
Wang, Y., Liu, J., Huang, B. O., Xu, Y. M., Li, J., Huang, L. F., Lin, J., Zhang, J., Min, Q. H., Yang, W. M., & Wang, X. Z. (2015). Mechanism of alternative splicing and its regulation. *Biomed Rep*, *3*(2), 152-158. https://doi.org/10.3892/br.2014.407
Weller, M., Wick, W., Aldape, K., Brada, M., Berger, M., Pfister, S. M., Nishikawa, R., Rosenthal, M., Wen, P. Y., Stupp, R., & Reifenberger, G. (2015). Glioma. *Nat Rev Dis Primers*, *1*, 15017. https://doi.org/10.1038/nrdp.2015.17
Wen, C., Zhang, A., Quan, S., & Wang, X. (2020). BeSS: An r Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. *Journal of Statistical Software*, *94*(4), 1627. https://doi.org/10.21105/joss.01627
Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., & et al. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, *265*(5181), 2088-2090. https://doi.org/10.1126/science.8091231
Wu, M. F., Liao, C. Y., Wang, L. Y., & Chang, J. T. (2017). The role of Slit-Robo signaling in the regulation of tissue barriers. *Tissue Barriers*, *5*(2), e1331155. https://doi.org/10.1080/21688370.2017.1331155
Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141
Xing, H., Wang, P., Liu, S., Jing, S., Lin, J., Yang, J., Zhu, Y., & Yu, M. (2021). A global integrated analysis of UNC5C down-regulation in cancers: insights from mechanism and combined treatment strategy. *Biomed Pharmacother*, *138*, 111355. https://doi.org/10.1016/j.biopha.2021.111355
Xing, Y., Yu, T., Wu, Y. N., Roy, M., Kim, J., & Lee, C. (2006). An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*, *34*(10), 3150-3160. https://doi.org/10.1093/nar/gkl396
Xu, X., Yan, Q., Wang, Y., & Dong, X. (2017). NTN4 is associated with breast cancer metastasis via regulation of EMT-related biomarkers. *Oncol Rep*, *37*(1), 449-457. https://doi.org/10.3892/or.2016.5239
Xu, Y., Olman, V., & Xu, D. (2002). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, *18*(4), 536-545. https://doi.org/10.1093/bioinformatics/18.4.536
Xue, M., Shang, J., Chen, B., Yang, Z., Song, Q., Sun, X., Chen, J., & Yang, J. (2019). Identification of Prognostic Signatures for Predicting the Overall Survival of Uveal Melanoma Patients. *J Cancer*, *10*(20), 4921-4931. https://doi.org/10.7150/jca.30618
Yan, X., Fu, X., Guo, Z. X., Liu, X. P., Liu, T. Z., & Li, S. (2020). Construction and validation of an eight-gene signature with great prognostic value in bladder cancer. *J Cancer*, *11*(7), 1768-1779. https://doi.org/10.7150/jca.38741

Yang, B., Qu, M., Wang, R., Chatterton, J. E., Liu, X. B., Zhu, B., Narisawa, S., Millan, J. L., Nakanishi, N., Swoboda, K., Lipton, S. A., & Zhang, D. (2015). The critical role of membralin in postnatal motor neuron survival and disease. *Elife*, *4*. https://doi.org/10.7554/eLife.06500
Yang, L., Tan, W., Yang, X., You, Y., Wang, J., Wen, G., & Zhong, J. (2020). Sorting nexins: A novel promising therapy target for cancerous/neoplastic diseases. *J Cell Physiol*. https://doi.org/10.1002/jcp.30093
Yao, R. W., Wang, Y., & Chen, L. L. (2019). Cellular functions of long noncoding RNAs. *Nat Cell Biol*, *21*(5), 542-551. https://doi.org/10.1038/s41556-019-0311-8
You, K. (2021). mclustcomp: Measures for Comparing Clusters. https://CRAN.R-project.org/package=mclustcomp
Yu, G. (2021). *enrichplot: Visualization of Functional Enrichment Result*. In https://bioconductor.org/packages/release/bioc/html/enrichplot.html
Yu, H., Peters, J. M., King, R. W., Page, A. M., Hieter, P., & Kirschner, M. W. (1998). Identification of a cullin homology region in a subunit of the anaphase-promoting complex. *Science*, *279*(5354), 1219-1222. https://doi.org/10.1126/science.279.5354.1219
Yu, W., Goncalves, K. A., Li, S., Kishikawa, H., Sun, G., Yang, H., Vanli, N., Wu, Y., Jiang, Y., Hu, M. G., Friedel, R. H., & Hu, G. F. (2017). Plexin-B2 Mediates Physiologic and Pathologic Functions of Angiogenin. *Cell*, *171*(4), 849-864 e825. https://doi.org/10.1016/j.cell.2017.10.005
Yuan, J. H., Yang, F., Wang, F., Ma, J. Z., Guo, Y. J., Tao, Q. F., Liu, F., Pan, W., Wang, T. T., Zhou, C. C., Wang, S. B., Wang, Y. Z., Yang, Y., Yang, N., Zhou, W. P., Yang, G. S., & Sun, S. H. (2014). A long noncoding RNA activated by TGF-beta promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell*, *25*(5), 666-681. https://doi.org/10.1016/j.ccr.2014.03.010
Zhang, E., He, X., Zhang, C., Su, J., Lu, X., Si, X., Chen, J., Yin, D., Han, L., & De, W. (2018). A novel long noncoding RNA HOXC-AS3 mediates tumorigenesis of gastric cancer by binding to YBX1. *Genome Biol*, *19*(1), 154. https://doi.org/10.1186/s13059-018-1523-0
Zhang, F., Prahst, C., Mathivet, T., Pibouin-Fragner, L., Zhang, J., Genet, G., Tong, R., Dubrac, A., & Eichmann, A. (2016). The Robo4 cytoplasmic domain is dispensable for vascular permeability and neovascularization. *Nat Commun*, *7*, 13517. https://doi.org/10.1038/ncomms13517
Zhang, X., Klamer, B., Li, J., Fernandez, S., & Li, L. (2020). A pan-cancer study of class-3 semaphorins as therapeutic targets in cancer. *BMC Med Genomics*, *13*(Suppl 5), 45. https://doi.org/10.1186/s12920-020-0682-5
Zhang, Z., Turer, E., Li, X., Zhan, X., Choi, M., Tang, M., Press, A., Smith, S. R., Divoux, A., Moresco, E. M., & Beutler, B. (2016). Insulin resistance and diabetes caused by genetic or diet-induced KBTBD2 deficiency in mice. *Proc Natl Acad Sci U S A*, *113*(42), E6418-E6426. https://doi.org/10.1073/pnas.1614467113
Zhao, H. F., Wang, J., Shao, W., Wu, C. P., Chen, Z. P., To, S. T., & Li, W. P. (2017). Recent advances in the use of PI3K inhibitors for glioblastoma multiforme: current preclinical and clinical development. *Mol Cancer*, *16*(1), 100. https://doi.org/10.1186/s12943-017-0670-3
Zhao, S. J., Shen, Y. F., Li, Q., He, Y. J., Zhang, Y. K., Hu, L. P., Jiang, Y. Q., Xu, N. W., Wang, Y. J., Li, J., Wang, Y. H., Liu, F., Zhang, R., Yin, G. Y., Tang, J. H., Zhou, D., & Zhang, Z. G. (2018). SLIT2/ROBO1 axis contributes to the Warburg effect in osteosarcoma through activation of SRC/ERK/c-MYC/PFKFB2 pathway. *Cell Death Dis*, *9*(3), 390. https://doi.org/10.1038/s41419-018-0419-y
Zheng, H., Brennan, K., Hernaez, M., & Gevaert, O. (2019). Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *Gigascience*, *8*(12). https://doi.org/10.1093/gigascience/giz145
Zhou, W. J., Geng, Z. H., Chi, S., Zhang, W., Niu, X. F., Lan, S. J., Ma, L., Yang, X., Wang, L. J., Ding, Y. Q., & Geng, J. G. (2011). Slit-Robo signaling induces malignant transformation through Hakai-mediated E-cadherin degradation during colorectal epithelial cell carcinogenesis. *Cell Res*, *21*(4), 609-626. https://doi.org/10.1038/cr.2011.17
Zhou, Y., Gunput, R. A., & Pasterkamp, R. J. (2008). Semaphorin signaling: progress made and promises ahead. *Trends Biochem Sci*, *33*(4), 161-170. https://doi.org/10.1016/j.tibs.2008.01.006
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(5), 768-768.

Permissions

For Figures 29-30, permissions were obtained from Elsevier with license #5170380416279. Figures 31-33 did not require permissions for reproduction. See below for permission forms from the Copyright Clearance Center (http://www.copyright.com/) and the Creative Commons Attribution (https://creativecommons.org/licenses/by-nc/4.0/).

For Zhou et al. (2008):

ELSEVIER LICENSE
TERMS AND CONDITIONS

Nov 01, 2021

This Agreement between Ayse Ozhan ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5170380416279
License date	Oct 15, 2021
Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Biochemical Sciences
Licensed Content Title	Semaphorin signaling: progress made and promises ahead
Licensed Content Author	Yeping Zhou,Rou-Afza F. Gunput,R. Jeroen Pasterkamp
Licensed Content Date	Apr 1, 2008
Licensed Content Volume	33
Licensed Content Issue	4
Licensed Content Pages	10
Start Page	161
End Page	170
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	2
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Miss
Institution name	Bilkent University
Expected presentation date	Oct 2021
Portions	Figures 2 & 3

Requestor Location	Ankara, other 06800 Turkey Attn: Ayse Ozhan
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD

For Meijers et al. (2020):

For Tong et al. (2019):

For Ozhan et al. (2021):