Syrian Arab Republic

Ministry of Higher Education

Syrian Virtual University

Master of Bioinformatics

Investigating the role of aberrant alternative splicing in Cholangiocarcinoma via Integrated Bioinformatics

A project submitted for Master's degree of Bioinformatics

Presented by:

Dima Swaid / dima_162261

Supervised by:

Dr. Yasser khadra

2023

Table of Contents

Table of tables

Table of figures

Table of abbreviations	
Abbreviation	**The meaning**
A3SS	Alternative 3’ splice site
A5SS	Alternative 5’ splice site
AFE	Alternative first exon
ALE	Alternative last exon
AS	Alternative Splicing
BP	Biological Process
CC	Cellular Component
CHOL	Cholangiocarcinoma
CRISPR	clustered regularly interspaced short palindromic repeats
ES	Exon Skipping
GO	Gene Ontology
IR	Intron Retention
KEGG	Kyoto Encyclopedia of Genes and Genomes
MF	Molecular Function
MXE	Mutually exclusives Exons
OS	Overall-Survival
PCA	Principle Component Analysis
PAM	Protospacer Adjacent Motif
TCGA	The Cancer Genome Atlas

Abstract:

Background: Cholangiocarcinoma is one of the rare and fatal cancers, early diagnosis is important and essential in treatment. Bearing in mind that alternative splicing has a major role and is mentioned in many researches related to cancer. Studying alternative splicing patterns and comparing them between healthy and tumoral samples may have a major role in finding new biomarkers for the diagnosis and prediction of cholangiocarcinoma before reaching advanced stages of cancer.

Aim of the study: This study was conducted to examine the impact of alternative splicing on cholangiocarcinoma. It aimed to analyze the differences in splicing patterns between tumoral and normal samples using bioinformatics-based alternative splicing detection tools. Furthermore, the study aimed to discover new biomarkers and investigate genetic modification techniques.

Tools and Methods: A series of analyses were conducted using Psichomics R package to explore downregulated and upregulated genes associated with CHOL. To gain further insights, we performed GO enrichment analysis (including Molecular Function, Biological Process, and Cellular Component) and KEGG pathway enrichment analysis using the ShinyGO 0.76 tool. We also used the STRING software to construct a network and exported it to Cytoscape for further exploration and gene clustering. Afterwards basic information related to the main genes cluster were checked and retrieved from AScancer atlas**.** Finally, we performed targeted genome editing using CRISPOR tool

Results: we identified 283 events that were differentially spliced between tumoral and normal samples, by networking the resulted genes we found the main cluster of genes identified in the network including 10 genes. These genes were found to be enriched in biological processes directly associated with alternative splicing. Although these genes have been reported to be involved in other conditions such as relapsing-remitting multiple sclerosis, Myotonic dystrophy type 2, Localized scleroderma, Liposarcoma, and Sjogren's syndrome, their association with cholangiocarcinoma had not been previously reported. Therefore, these genes may open up new avenues for research in the context of cholangiocarcinoma.

By analyzing the differential splicing and PSI distributions of 10 AS events across tumoral and normal samples we found that 6 events were upregulated in tumoral samples, while 4 events were downregulated compared to normal samples. Additionally, Kaplan-Meier survival analysis revealed that two alternative splicing events (TIAL1 and SNRNP70) did not have a significant p-value and therefore cannot be used as tumor biomarkers for cholangiocarcinoma. However, the other alternative splicing events showed a significant p-value and may serve as novel biomarkers with high prognostic value. Finally, by using CRISPOR web tool we were able to design gRNA for the AS event on EIF4A2 gene (one of the genes that has significant AS event correlated with CHOL patients) with the suitable PAM and restriction enzymes in order to make targeted genome editing**.**

In conclusion, the results improve our understanding of the association between AS events and CHOL and might be a starting point for further research to confirm the importance of the splicing events studied in this research in CHOL and to identify new prognosis biomarkers.

1. Introduction

Cholangiocarcinoma is a type of cancer that forms in the bile ducts, which are a series of tubes that transport bile from the liver and gallbladder to the small intestine. This cancer can occur in different parts of the bile ducts, and it is classified into three main types based on its location: intrahepatic (within the liver), perihilar (at the junction where the bile ducts leave the liver), and distal (further down the bile ducts closer to the small intestine). Although cholangiocarcinoma is rare, its global incidence has been increasing in recent decades. [1, 2]

CHOL is asymptomatic in the early stages, which leads to a delay in diagnosis until the disease reaches advanced stages. The 5-year survival rate for patients with cholangiocarcinoma is estimated to be 7-20%, and the recurrence rates of tumors after surgical removal are high. [3]

For these reasons, there is an urgent need to find new biomarkers for the diagnosis and prediction of cholangiocarcinoma before reaching advanced stages of cancer.

Most genes in eukaryotes contain sequences called introns that located between the exons which are considered the expressed sequences. After transcription, the pre-mRNA from any multi-exon gene has to undergo extensive processing to remove the introns, in order for the expressed transcript (pre-mRNA) to become a suitable message for downstream processes such as translation of the encoded protein.[4]

The process of removing introns from pre-mRNAs and connecting the remaining exons to each other to produce mature mRNA is called splicing, while the different combination of exons in the mRNA producing diversified mature mRNA is called alternative splicing (AS) , so AS is a posttranscriptional process occurs in about 90% of all genes in eukaryotes

by which identical pre-mRNA molecules are spliced in different ways so it is the main source of protein diversity and complexity in >90% of human genes. [5, 6]

The process is carried out by a special molecule called the spliceosome, which identifies the junction between introns and exons. To differentiate introns from exonic sequences, the spliceosome typically follows the "GU-AG" rule using four specific consensus sequences: The 5′splice-site (SS) characterized by a GU dinucleotide at the 5′ end of the intron, the 3′ SS with AG at the 3′ end of the intron, the branch point sequence (BPS) located upstream of the 3′SS, and the polypyrimidine tract located between the BPS and the 3′ SS.[7]

There are five main forms of alternative splicing, including: ES exon skipping (also called cassette exon), IR intron retention, MXE mutually exclusive exons (only some exons appear in mature mRNA), A5SS (the change of the splicing site causes the position of the 3′ end of the exon to change) and A3SS (the change of the splicing site causes the position of the 5′ end of the exon to change). [8]

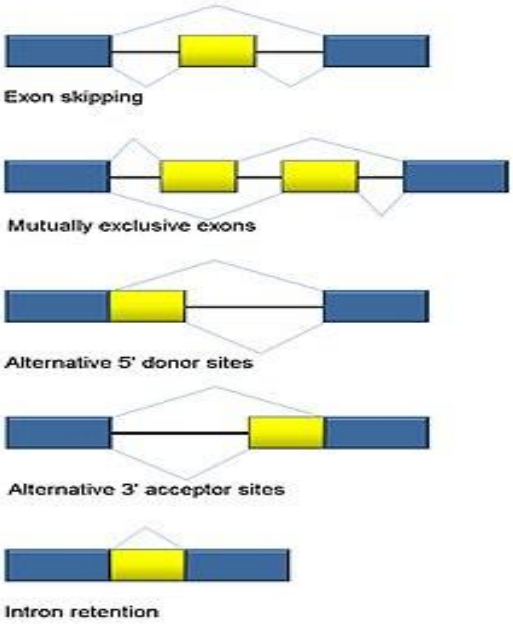


Fig (1) The main types of alternative splicing

All data to date indicate that alternative splicing is a well-designed process tightly regulated to produce a network of alternatively spliced transcripts. While normal alternative splicing generates diverse and multifunctional proteins, it has been shown in several previous studies that alterations in AS affect basic biological processes and many pathological conditions, and it has been shown that aberrant AS can contribute to both the original and emerging hallmarks of cancer. [9, 10]

Furthermore, in the past two decades many studies have shown that splicing changes dramatically in disease states, especially tumors due to somatic mutations in components of the human splicing machinery. Cancer cells are characterized by their long-life span and their ability to divide greatly. In general, some tumor suppressor genes prevent normal cells from becoming cancerous by acting on the cell cycle and promoting epigenetic changes; It is therefore common for cancer cells to exhibit aberrant splicing activity with an increased frequency of splicing isoforms that maintain the abnormal rhythm of proliferation and apoptosis. [4, 8]

High-throughput sequencing of short cDNA fragments, also known as RNA-seq, is a method used to study and analyze alternative splicing and other post-transcriptional processes. By analyzing RNA-seq reads, researchers can identify the usage of known or unknown splice sites and determine the expression level of exons.

Split alignments can never cover the constitutive exons, whereas alternative exonic parts are located within highly expressed splicing junctions. The ratio of the relative abundance of all isoforms containing a certain exon over the relative abundance of all isoforms of the gene containing the exon, AKA percent spliced in (PSI) is a popular statistic that have been widely used in differential expression analysis as it indicates how efficiently sequences of interest are spliced into transcripts. [11, 12]

2. *Research problem*

The question that defines the research problem is:

Is there a role for alternative splicing in the emergence and progression of cholangiocarcinoma?

3. *Research hypothesis*

To answer the question that defines the research problem we hypothesize that alternative splicing has a role in cholangiocarcinoma development and progression.

4. *Aim of the study*

This study was conducted to examine the impact of alternative splicing on cholangiocarcinoma. It aimed to analyze the differences in splicing patterns between tumoral and normal samples using bioinformatics-based alternative splicing detection tools. Furthermore, the study aimed to discover new biomarkers and investigate genetic modification techniques.

5. *Literature Review:*

The role of aberrant alternative splicing in cholangiocarcinoma has been studied to understand its impact on the development and progression of this cancer. Aberrant alternative splicing events have been found to contribute to the dysregulation of various cellular processes involved in cholangiocarcinoma, including cell proliferation, apoptosis, invasion, and metastasis.

Several studies have identified specific genes and splicing factors that are involved in aberrant alternative splicing in cholangiocarcinoma, however the methodology of each research differed according to the nature of the study. Some of the studies were literature reviews and relied on summarizing the theoretical aspect of important genes in order to clarify the current understanding of the impact of AS on CHOL. For example, the study conducted by Yosudjai et al [13] and the study conducted by Marin et al [14] listed a set of genes involved in cholangiocarcinoma, their splicing variants, and their consequences.

Table 1 provides a summary of 9 genes and their splicing variants from the study conducted by Yosudjai et al. It includes information on the gene, splicing variant, and its function

Table 2 in the paragraph summarizes 8 genes, their splicing variants, and their consequences from the study conducted by Marin et al. It provides information on the gene, splicing variant, and its consequence. For instance, CD44 with exon 6 skipping is associated with decreased overall survival, and AGR2 with several combinations of exons 2 to 7 skipping affects cancer cell survival and migration.

Table 1: Summary of 9 genes and their splicing variants in Yosudjai. J et al study

Gene	Splicing variant	Function
CD44	Retained exon v6	Proliferation
WISP1	Skipping exon 3	Neural and lymphatic invasion
NEK2B	Skipping exon 8	Unknown

TFF2	Skipping exon 2	Independent prognostic marker
FOXP3	Skipping exon 3	Unknown
P53	Exon 1-4 skipping	Independent prognostic marker
PKM2	Mutually exclusive exon; exon 9 skipping and exon 10 retention	Neural invasion
EP3-4	Exon 2b,3,4,6 and 8 skipping	Proliferation migration and invasion
AGR2vH	Alternative 3' and 5' splice site and exon 4-7 skipping	Migration, invasion and adhesion

Table 2: Summary of 8 genes, their splicing variants and its consequences in Marin, J.J.G et al study

Gene	Splicing variant	Consequence
CD44	Exon 6 skipping	Decreased OS
AGR2	Several combinations of exons 2 to 7 skipping	Affect cancer cell survival and migration
BAP1	Multiple SV lacking exon 14-17	Promote tumorigenesis
TFF2	Skipping exon 2	Increased OS
FOXP3	Skipping exon 2-4	Immune suppression
EP3-4	Contains exon 1,2a,5 and 10	Enhanced cell proliferation, migration and invasion
PKM2	Mutually exclusive exon; exon 9 skipping and exon 10 retention	Decreased OS, Enhanced risk of metastasis
WISP1	Exon 3 skipping	Decreased OS, Induction of invasion

Some other researches took the direction of clinical (in-vitro) studies, which involved studying cell lines and comparing alternative splicing events in certain genes between healthy subjects and cholangiocarcinoma patients. An example is the study conducted by Herraez et al [15], which investigated the expression of aberrant OCT1 variants in hepatocellular carcinoma and cholangiocarcinoma. They evaluated the potential impact of these variants on the sensitivity of tumors to sorafenib by sequencing DNA samples from surgically removed tumors.

In addition, computational (in-silico) studies were conducted by researchers like Lin et al [3] and Wu et al [16]. Lin et al used bioinformatics tools to develop prognostic signatures for overall survival in cholangiocarcinoma patients based on cancer pathway-related alternative splicing events. Wu et al relied on bioinformatics tools to construct a model for predicting cholangiocarcinoma prognosis based on alternative splicing events.

6. *Tools and data collection*

Several tools are currently available to analyze alternative splicing data, we reviewed multiple tools and found that many of them have certain shortcomings. These include:

1. Some tools require expensive computational resources that take BAM or FastQ files as input (e.g., SplAdder [17] an alternative splicing toolbox, that takes RNA-Seq alignments and an annotation file as input)
2. Certain tools have a limited set of statistical options for differential splicing analysis (e.g., TCGA SpliceSeq [18]: A tool for investigating alternative mRNA splicing in TCGA tumor and adjacent normal samples, it contains various statistical summaries, but the user cannot perform different statistical tests that serve different purposes for research and analysis)
3. Some tools lack a user-friendly interactive graphical interface (e.g., SUPPA [19] a tool to study splicing at the transcript isoform or at the local alternative splicing event level has been developed in Python, the user must be familiar with Python and requires direct involvement in every step of the analysis).

In order to avoid these drawbacks, we used ***Psichomics*** [20]. An interactive R package with an intuitive Shiny-based graphical interface for alternative splicing quantification and integrative analyses based on The Cancer Genome Atlas (TCGA) [21], the Genotype-Tissue Expression project (GTEx) [22] and Sequence Read Archive (SRA) [23].

However, it should be noted that Psichomics only detects exon skipping events and has this as its main limitation.

The package *is* available in Bioconductor at (http://bioconductor.org/packages/psichomics) [24]

psichomics uses PSI values to quantify AS, as shown in figure 2.

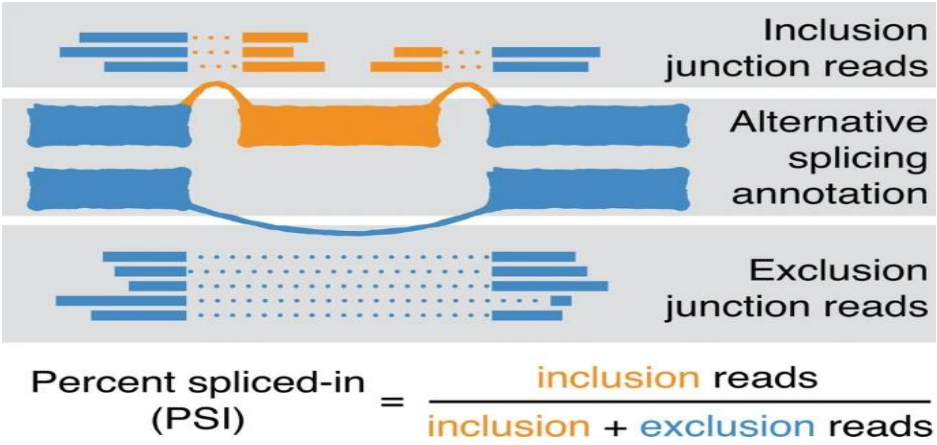


Fig (2): PSI calculation method in Psichomics; each AS event is quantified based on the ratio of the number of reads that contain a given alternative splicing sequence (in orange) to the total number of reads that contain and do not contain that sequence (in blue).

We utilized multiple tools for further analysis, including:

ShinyGO 0.76 [25]

This graphical tool serves as a large annotation and pathway database compiled from various sources for gene enrichment analysis. It is developed based on several R packages and provides access to KEGG and STRING for pathway diagrams and protein-protein interaction network retrieval.

STRING [26]

It is a database that contains known protein-protein interactions, including direct physical and indirect functional associations, as well as predicted interactions based on computational prediction and knowledge

transfer between organisms. Users can explore networks of their set of genes of interest and export them to Cytoscape for further analysis.

Cytoscape [27]

This open-source software platform is used for visualizing molecular interaction networks and biological pathways. It offers various tools and sub-programs to integrate networks with annotations, gene expression profiles, and other state data.

ASCancer Atlas [28]

Alternative **S**plicing in human **Cancer**s **A**tlas is a comprehensive knowledgebase of alternative splicing in human cancer, houses about 2 million Computationally Putative Splicing Events (CPSE), covering 33 TCGA cancer types and 31 GTEx normal tissues and A total of 2,006 high-confidence Cancer Associated Splicing Events (CASE) with established oncogenic roles summarized from extensive functional studies. It has several toolkits that allow the user to explore, visualize and analyze AS events associated with different types of cancer.

CRISPOR [29]

This web tool is specifically designed for genome editing experiments using the CRISPR-Cas9 system. It identifies guide RNAs in an input sequence and ranks them based on different scores that evaluate potential off-targets in the genome of interest and predict on-target activity.

7. Workflow for alternative splicing analysis:

The flowchart in figure 3 summarize the workflow:

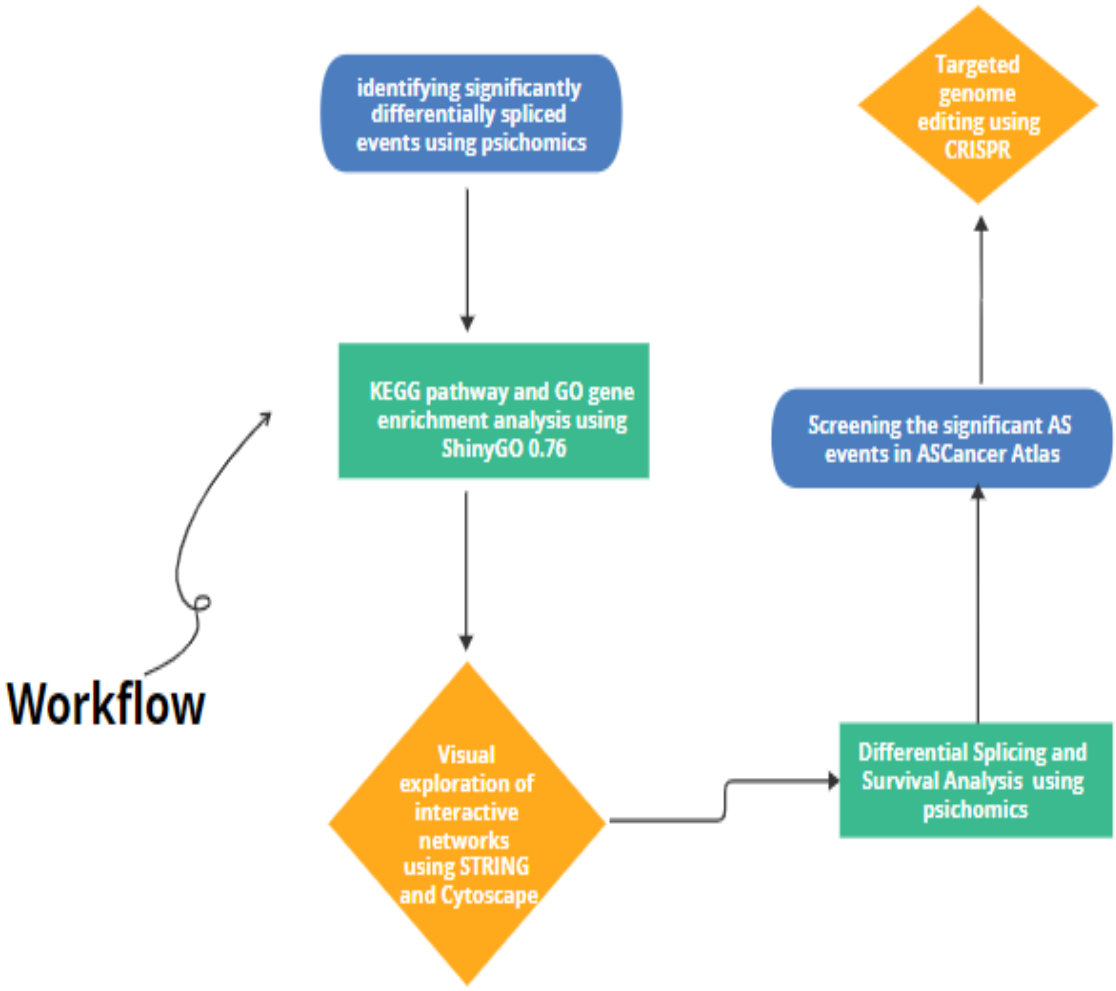


Fig (3): Flowchart of the methods workflow

7.1. Identifying significantly deferentially spliced events

First of all, we obtained cholangiocarcinoma data from The Cancer Genome Atlas TCGA Database and conducted an initial analysis to determine the number of splicing events and the most common types of alternative splicing (AS). Subsequently, we utilized the Psichomics R package to perform several steps including:

7.1.1 Exon-exon junction quantification, gene expression and sample-associated data retrieval

Junction quantification, gene expression quantifications (obtained from pre-processed RNA-seq data), clinical and sample metadata for cholangiocarcinoma were loaded from TCGA database.

7.1.2 Gene expression pre-processing

Gene expression quantifications was normalized by raw library size scaling which is one of the three types of normalization. The other two methods are normalization by gene length and normalization by known or unknown technical artifacts across samples. Library size, which refers to the total number of reads generated for a specific sample, can vary due to various factors such as differences in gene expression levels.

The purpose of library size normalization is to make the library sizes comparable by scaling the raw read counts in each sample using a single sample-specific factor that reflects its library size. This process aims to ensure that the gene expression measurements are adjusted for differences in library sizes among samples. [30]

7.1.3 Alternative splicing annotation and quantification

Sequencing of DNA produces sequences of unknown function identifying functional elements along the sequence of a genome is called genome annotation [31], the hg19 and hg38 annotation of alternative splicing events are available in psichomics.

Quantification of AS events is based on PSI index which is calculated differently for each type of alternative splicing, as depicted in figure 4.

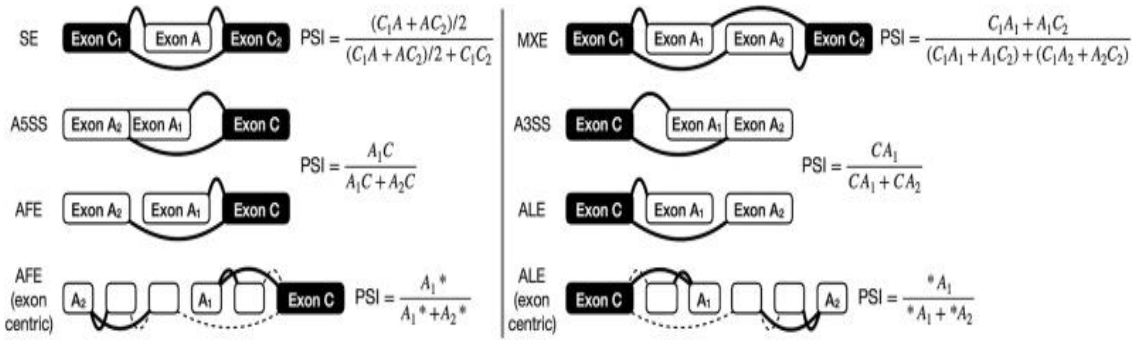


Fig (4): PSI calculation method for each AS type.

To be more specific, we excluded alternative splicing events that fell below a certain threshold in order to avoid inaccurate quantifications that are based on insufficient evidence.

7.1.4 Data grouping

Psichomics enables the grouping of subjects and their samples or genes and their alternative splicing events based on phenotypic and clinical characteristics. We categorized the data into six groups (as shown in figure 5) based on the clinical condition. The colors assigned to each group will be utilized to represent those same groups in the plots.

Group	Patients	Samples	Colour
stage i	18	18	
stage iv	7	7	
stage ii	11	10	
stage iii	4	1	
Primary solid Tumor	35	36	
Solid Tissue Normal	9	9	

Fig (5): The groups of patients

7.1.5 Dimensionality reduction

Dimensionality reduction is a technique used to simplify high-dimensional data by preserving important characteristics. There are various techniques for dimensionality reduction, and in this case, we utilized Principal Component Analysis (PCA) [32].

PCA identifies the combinations of variables that contribute the most to the variance in the data. We implemented PCA using the singular value decomposition (SVD) algorithm provided by the prcomp function from the R package stats (version 3.4.1). We applied PCA to the inclusion levels of all genes and splicing events in all samples, considering a tolerance of 5% for missing values per event. Additionally, we performed PCA on gene expression using both tumoral and normal samples.

7.1.6 Filtering alternative splicing events:

We conducted a Wilcoxon rank test to identify splicing events that were significantly differentially spliced. Subsequently, we filtered the splicing events by considering a substantial difference in median between the selected groups (Tumor and Normal).

Specifically, we only counted the events with an absolute difference in median PSI (Δ Median PSI) greater than 0.1 and a Wilcoxon q-value less than or equal to 0.01.

7.2. KEGG pathway and Gene Ontology enrichment analysis

The Gene Ontology (GO) is a widely used ontology that specifies the participation of human and model organism genes in Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). It provides

insights into the biological significance of genes within a specific gene set of interest [33]. KEGG mapping is the process of mapping molecular objects (genes, proteins, small molecules, etc.) to molecular interaction/reaction/relation networks [34].

In this study, we utilized ShinyGO 0.76 [35] to perform Gene Ontology (GO) and Kyoto Gene and Genome Encyclopedia (KEGG) enrichment analysis on the set of genes generated from the previous process.

7.3. Visual exploration of interactive networks

We utilized the STRING online database [36] to investigate significant interactions among the genes of interest. Subsequently, the network was exported to Cytoscape for further exploration of interactive networks and clustering of the genes to identify the main gene cluster.

7.4. Differential Splicing and Survival Analysis

Back to Psichomics density plots for every AS event was performed to check whether a significant difference in inclusion (PSI) between tumor and normal TCGA samples for AS events of each gene of the main cluster.

To evaluate the prognostic value of a given alternative splicing event, survival analysis was performed on groups of patients separated based on a given alternative splicing quantification (i.e., PSI) cut-off.

Kaplan-Meier estimators (and illustrating curves) and proportional hazard (PH) models have been applied to groups of patients with survival distributions being compared using the log-rank test.

7.5. Screening for the significant AS events in ASCancer

As previously mentioned, ASCancer Atlas provides various toolkits that enable us to explore, visualize, and analyze alternative splicing (AS) events. We compiled essential information about each gene in the main gene cluster and identified the splicing model of the AS event in protein-coding transcripts. Additionally, we compared the PSI value and expression value for each significant splicing event in CHOL TCGA samples.

7.6. Targeted genome editing using CRISPR

CRISPR-Cas9, also known as clustered regularly interspaced short palindromic repeats and CRISPR-associated system 9, is a naturally occurring genome editing system used by bacteria as a defense mechanism against viral infections. When bacteria are infected by viruses, short DNA sequences from the viruses are integrated into CRISPR loci within the bacterial genome. These sequences act as a "memory" of previous infections, allowing the bacteria to recognize and defend against the viruses in the future.

If the viruses attack again, the bacteria produce RNA segments from the CRISPR arrays to search for a matching sequence. This triggers the CRISPR-associated (Cas) nuclease to create a double-strand break at specific "foreign" DNA sequences."

Researchers adapted modified this immune defense system to edit DNA by designing a small RNA molecule that contains a short "guide" sequence. This guide RNA is capable of binding to a specific target sequence in the DNA of a cell. Additionally, the guide RNA also binds to the Cas9 enzyme. When introduced into cells, the guide RNA identifies the desired DNA

sequence, and the Cas9 enzyme cuts the DNA at the designated location. This process allows for gene editing by removing, adding, or modifying sections of the DNA sequence [37, 38].

In the context of targeting oncogenic alternative splicing events, CRISPR-Cas9 can be used to specifically edit the DNA sequences that are responsible for aberrant alternative splicing in cancer cells. By targeting and modifying these sequences, researchers can potentially correct the abnormal splicing patterns and restore normal gene expression.

We utilized the CRISPOR [39] web tool to identify specific guide RNAs for the significant alternative splicing (AS) event in the EIF4A2 gene.

8. *Results and data analysis:*

8.1. significant deferentially spliced events involved in CHOL:

Through the analysis of the CHOL TCGA dataset, we identified a total of 38,804 alternative splicing (AS) events in 9,673 genes. Among these events, exon skipping was the most common AS event, as depicted in figure (6).

Consequently, we utilized Psichomics to quantify all skipped exon events, as this tool specifically detects exon skipping events.

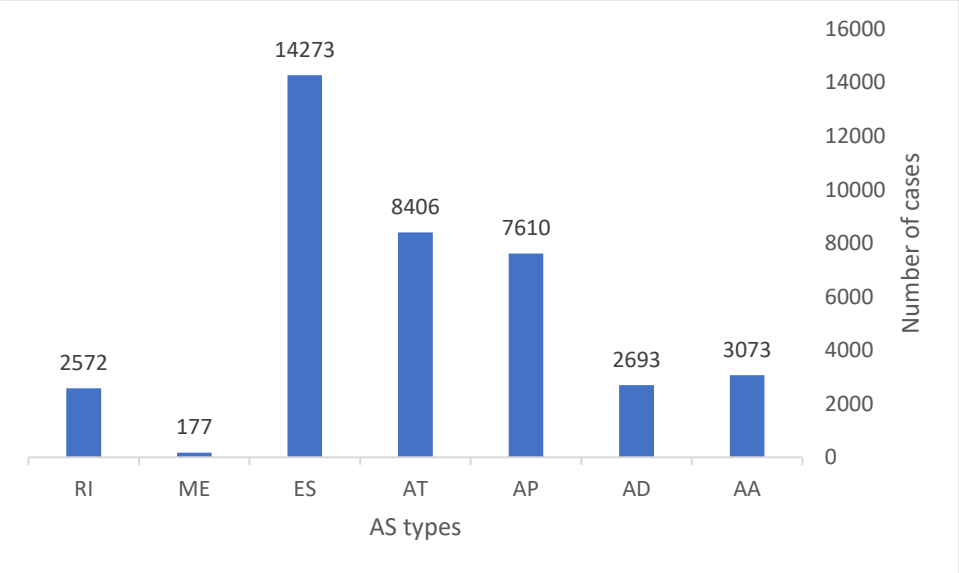


Fig (6): Distribution of AS types across cases in CHOL TCGA dataset

8.1.1 Data retrieval

The clinical data comprised 44 patients, consisting of 25 females and 19 males. All relevant clinical information, such as age, height, race, ethnicity, vital status, medical history, risk factors, age at initial pathologic diagnosis, and tumor stage (as shown in figure 7), was retrieved from TCGA database.

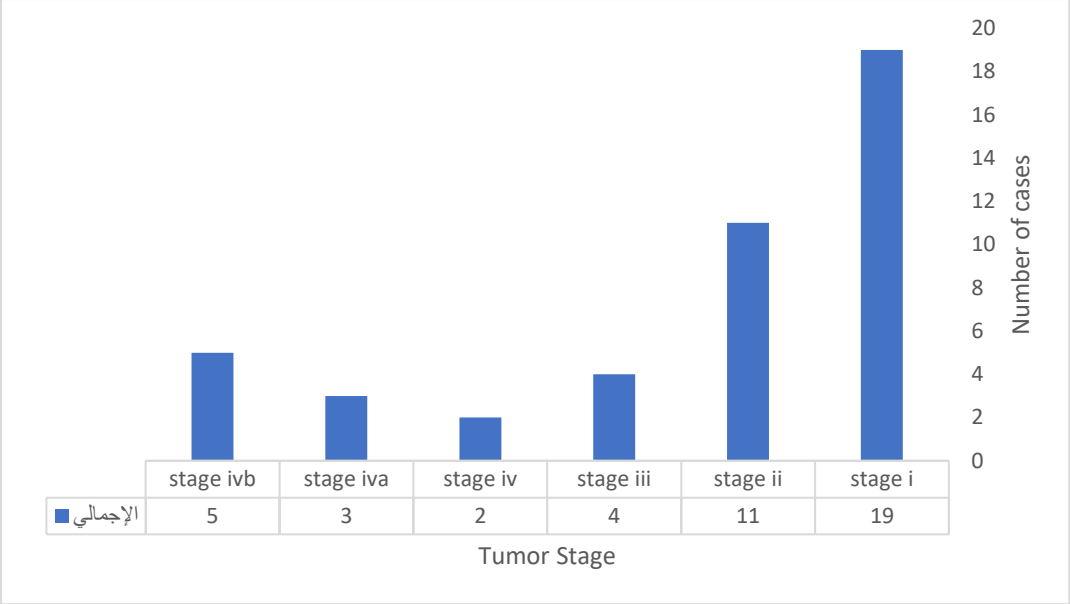


Fig (7): Distribution of tumor stages across cases in CHOL TCGA dataset

The junction quantification data revealed the presence of 249,567 splice junctions in the 45 patient samples. Additionally, gene expression analysis detected 20,531 genes. The distribution of library sizes across the samples is depicted in figure (8).

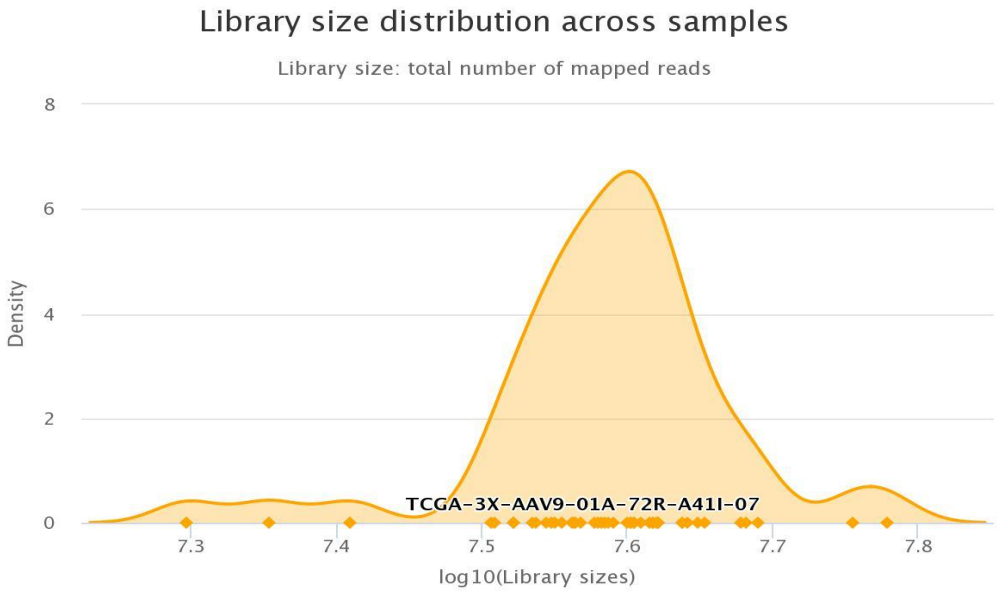


Fig (8): The library size distribution across samples

8.1.2 Gene expression pre-processing

Following the filtration and normalization of gene expression, we obtained a total of 14,036 genes. The distribution of these genes per sample is depicted in figure 9. Additionally, the distribution of the library sizes across the samples is illustrated in figure 10.

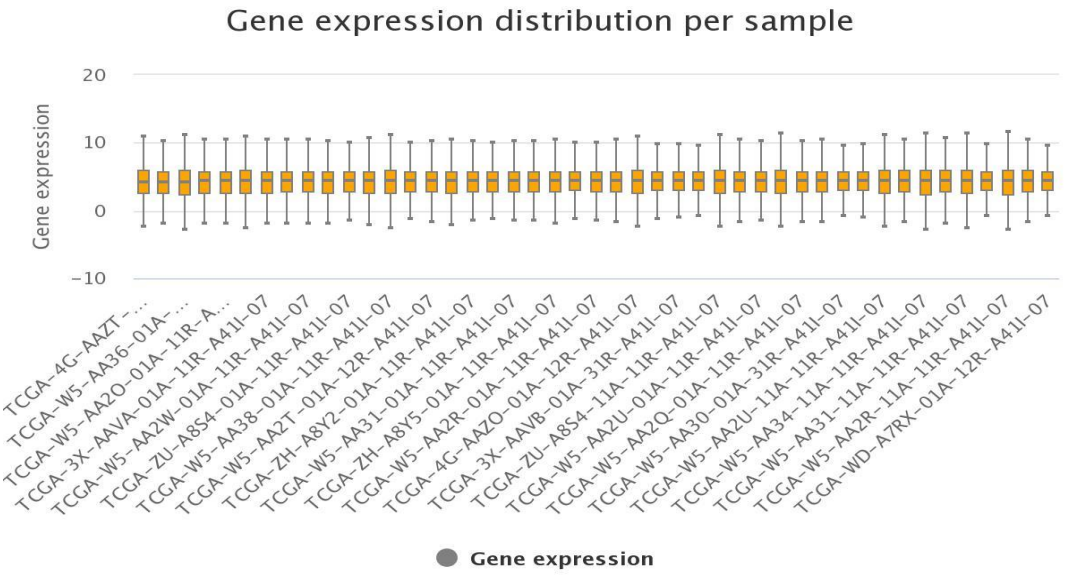


Fig (9): boxplot showing Distribution of genes across samples

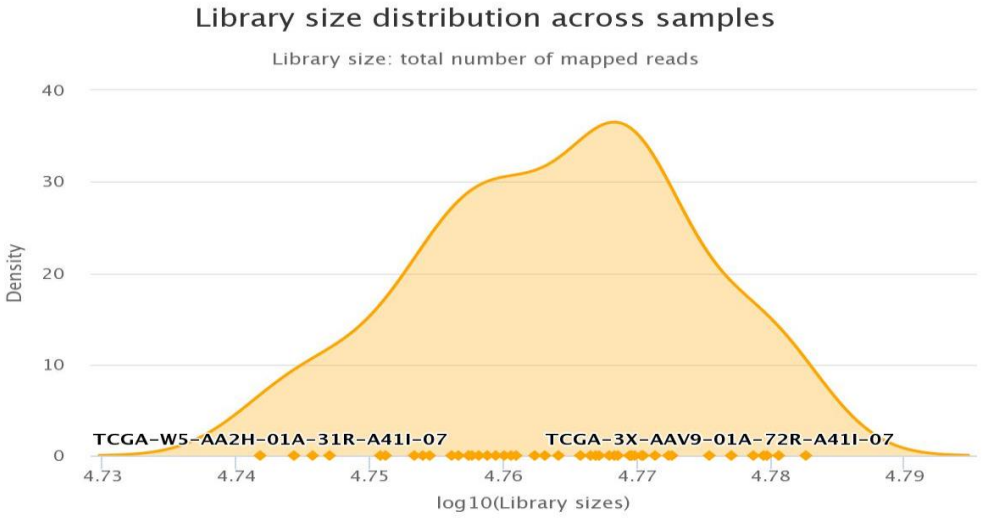


Fig (10): The library distribution across samples after normalization

8.1.3 Alternative splicing quantification

Quantifying alternative splicing was performed by selecting the junction quantification dataset from the loaded data and choosing Human hg19 as the alternative splicing event annotation. We chose Skipped exon (SE), Mutually exclusive exon (MXE), Alternative 5' splice site (A5SS), Alternative 3' splice site (A3SS), Alternative first exon (AFE) and Alternative last exon (ALE) as the event types of interest and set the minimum read counts' threshold to 10. Inclusion levels calculated with total read counts below this threshold are discarded from further analyses.

We identified a total of 30,196 alternative splicing events, each with a PSI value. Scatter plots illustrating the AS quantification per event can be found in figures 11A and 11B.

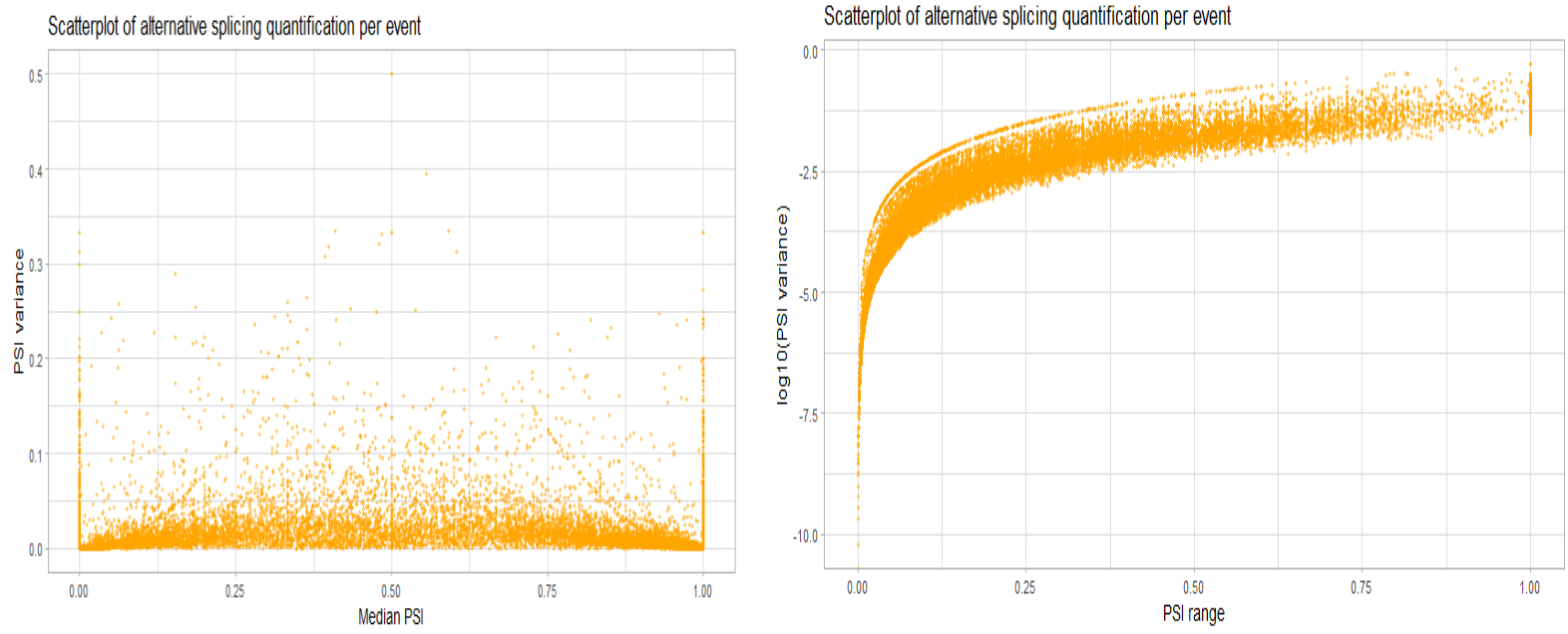


Fig (11A+ 11B) Scatter plots of AS quantification per event

8.1.4 Dimensionality reduction

After conducting PCA on Inclusion levels, we obtained 7,032 splice events. Subsequently, two PCA plots were generated. The initial plot, referred to as figure 12A, is a score plot illustrating the clinical samples. On the other hand, the second plot, known as the loadings plot (figure 12B), showcases the variables, specifically the alternative splicing events.

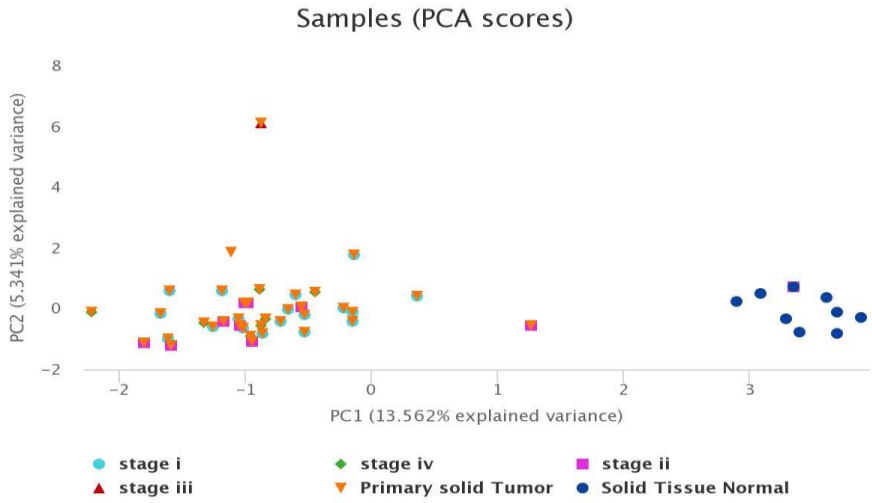


Fig (12A): score plot of PCA scores for the clinical samples

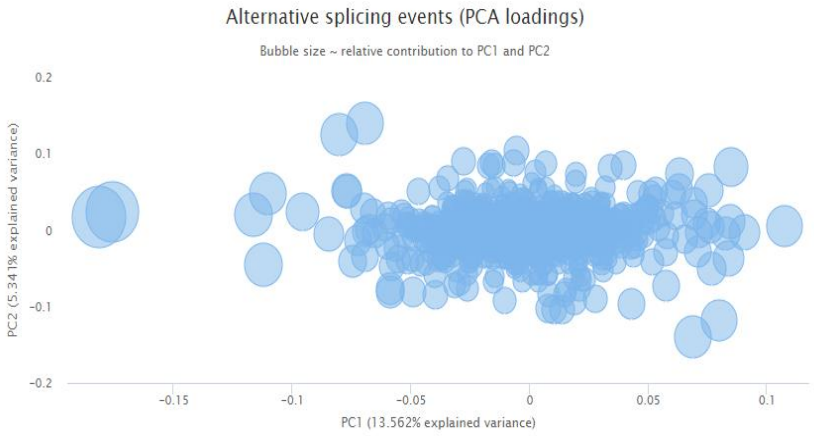


Fig (12B): loadings plot depicts the projection of splicing events on the two first principal components, with selected events labelled with their cognate gene symbol. The bubble size represents the relative contribution of each alternative splicing event to the selected principal components.

Then when we performed PCA for gene expression using Tumor and Normal samples we obtained 14,036 genes, also, two PCA plots are then generated. The first plot (figure13A) is a score plot that shows the clinical samples, while the second plot (figure 13B) displays the variables (in this case, gene expression)

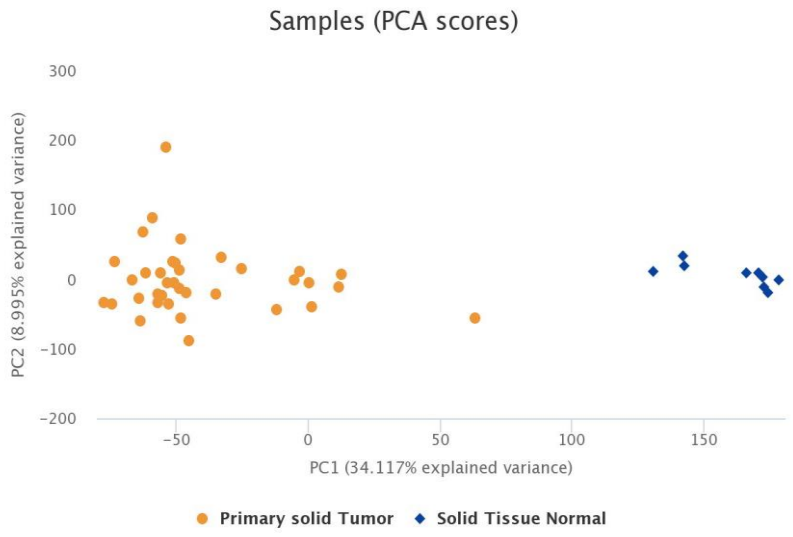


Fig (13A) score plot of PCA scores for the clinical samples

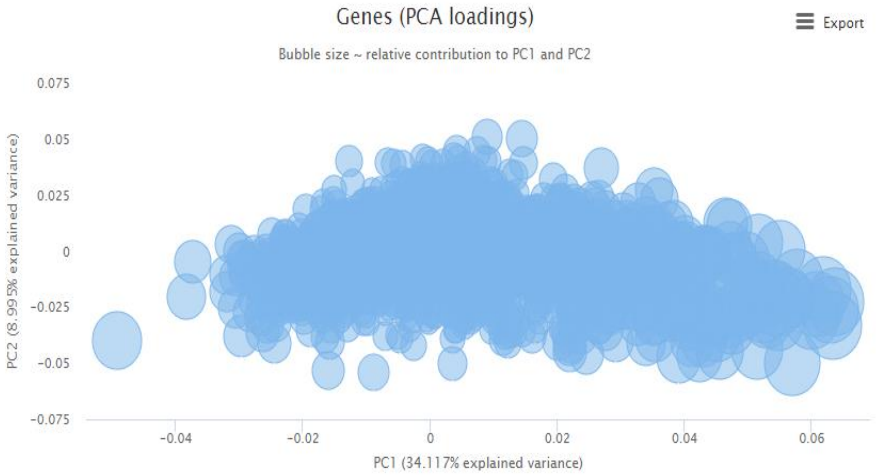


Fig (13B) loadings plot depicts the projection of gene expression on the two first principal components.

8.1.5 Filtering alternative splicing events:

Analyses deemed 283 events to be differentially spliced between tumor stage I and normal samples (with |Δ Median PSI| > 0.1 and Wilcoxon q-value ≤ 0.01) as shown in figure (14).

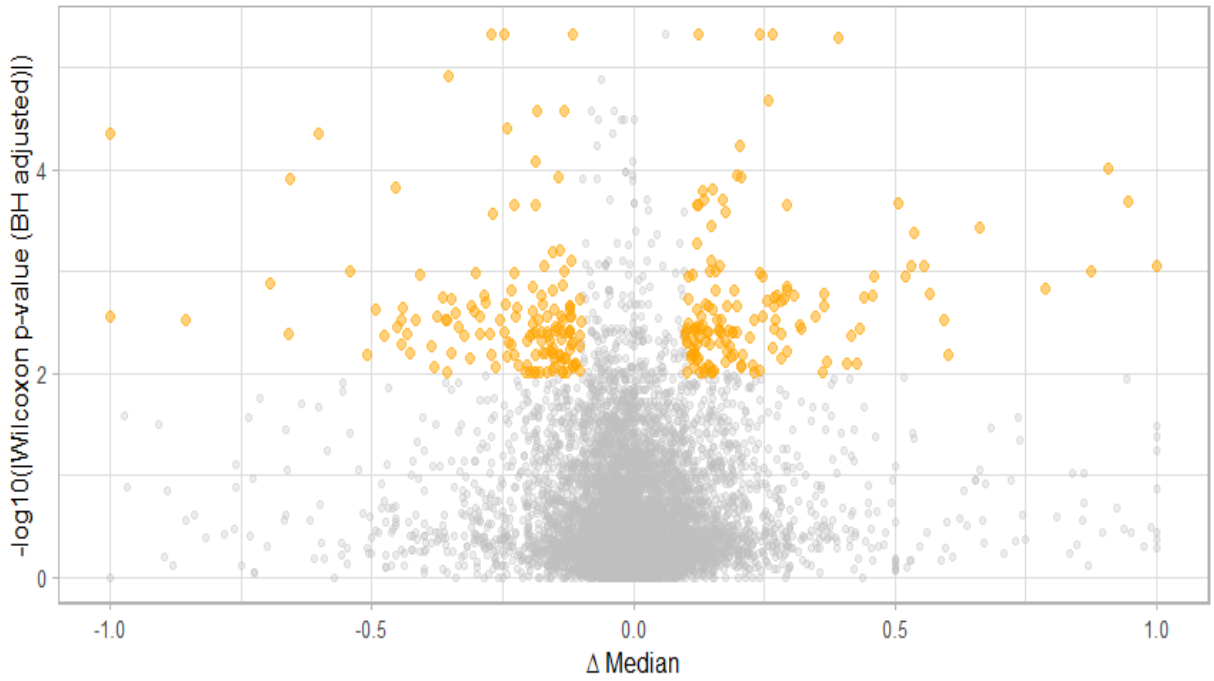


Fig (14): Volcano plot of differential splicing analysis performed between tumor stage I and normal samples using the Wilcoxon rank-sum test. Significantly differentially spliced events (|Δ median PSI| ≥ 0.1 and Wilcoxon rank-sum test with Benjamini–Hochberg (FDR) adjustment ≤ 0.01) are highlighted in orange (283 events).

8.2. KEGG pathway enrichment analysis and GO gene enrichment analysis:

The results of the GO Biological Process enrichment analysis showed that the genes associated with OS-related AS events were involved in various processes, including Peroxisome fission, Hepatocyte proliferation, Epithelial cell proliferation involved in Liver morphogenesis,

Mitochondrial fission, Liver development, and Hepatobiliary system development (Figure 15A+ 15B).

The GO Molecular Function analysis indicated that the genes with OS-related AS events were mainly associated with functions such as Vascular endothelial growth factor receptor 1 binding, L-ascorbic acid binding, Cadherin binding, Carboxylic acid binding, Heparin binding, Sulfur compound binding, and Cell adhesion molecule binding (Figure 16).

In terms of GO Cellular Component analysis, the genes with OS-related AS events were primarily categorized in Cytoplasmic stress granule, Cell cortex, Organelle sub compartment, Organelle envelope, Envelope, and Mitochondrion (Figure 17).

Furthermore, the KEGG pathway enrichment analysis revealed that the genes corresponding to these AS events were primarily enriched in Galactose metabolism, Regulation of lipolysis in adipocytes, VEGF signaling pathway, and PPAR signaling pathway (Figure 18).

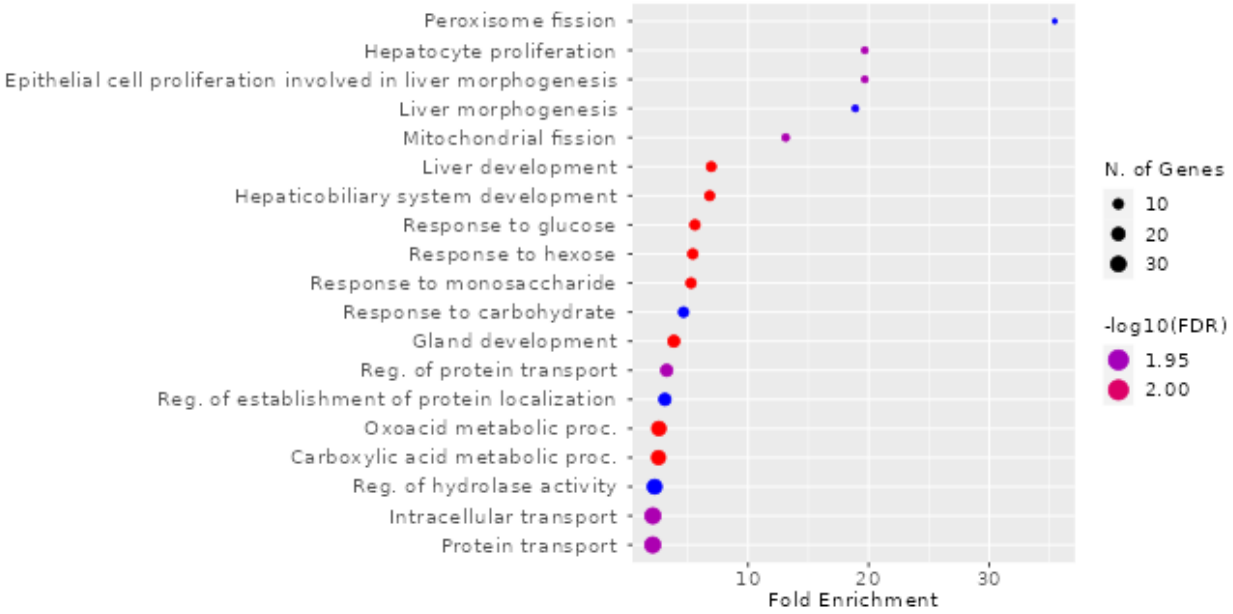


Fig (15A): Dot plot of GO Biological Process enrichment analysis of genes whose events are significantly differentially spliced between normal and tumoral samples

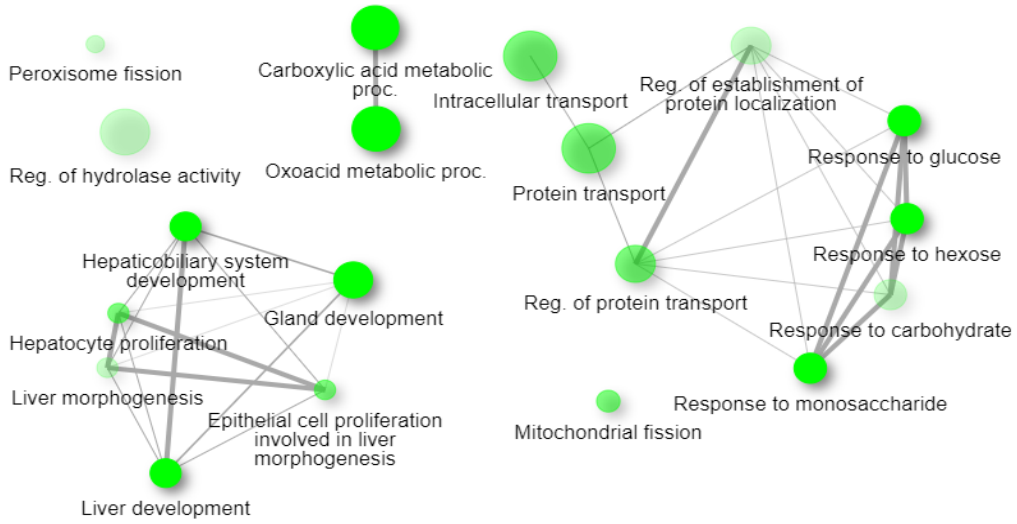


Fig (15B): A network of GO(BP) pathways of genes whose events are significantly differentially spliced between normal and tumoral samples. Two pathways (nodes) are connected if they share 20% or more genes. Darker nodes are more significantly enriched gene sets. Bigger nodes represent larger gene sets. Thicker edges represent more overlapped genes.

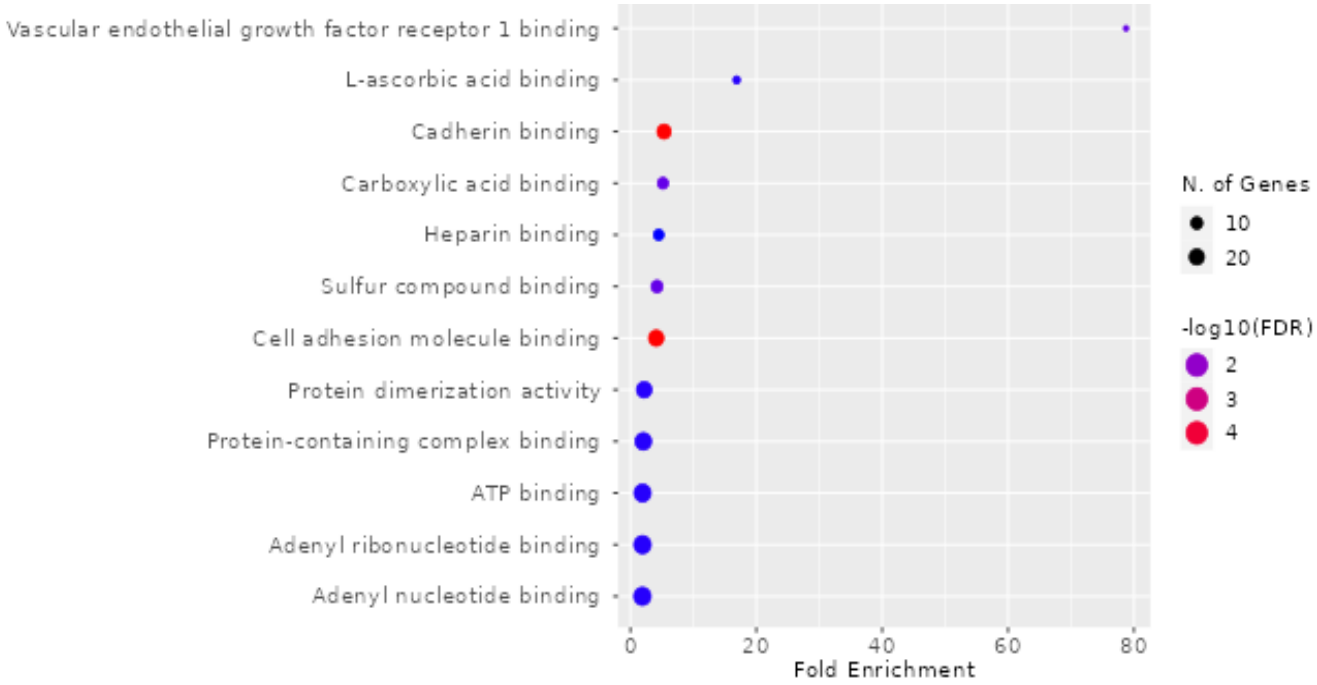


Fig (16): Dot plot of GO Molecular Function analysis of genes whose events are significantly differentially spliced between normal and tumoral samples

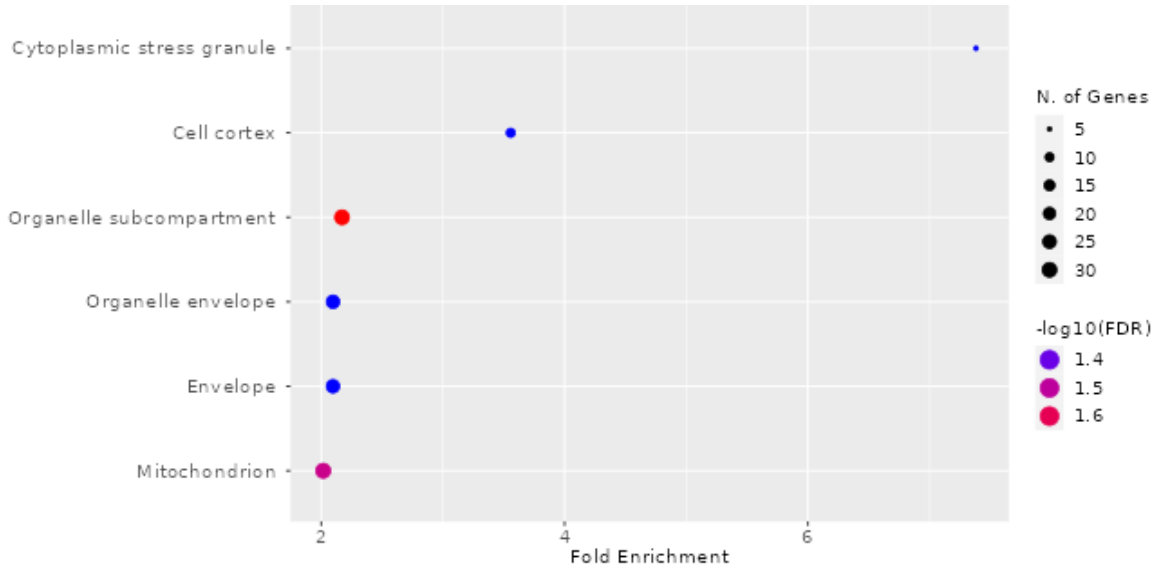


Fig (17): Dot plot of GO Cellular Component analysis of genes whose events are significantly differentially spliced between normal and tumoral samples

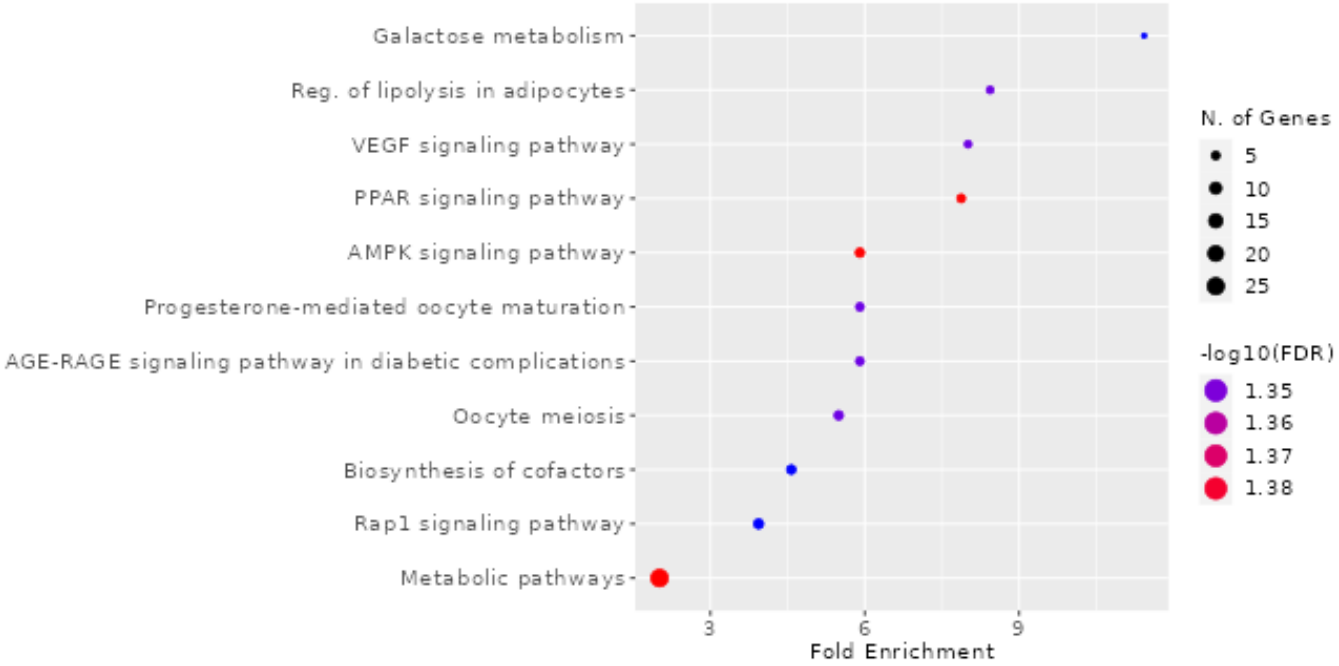


Fig (18): Dot plot of KEGG pathway enrichment analysis of genes whose events are significantly differentially spliced between normal and tumoral samples.

The distribution of events among genome is illustrated in figure (19):

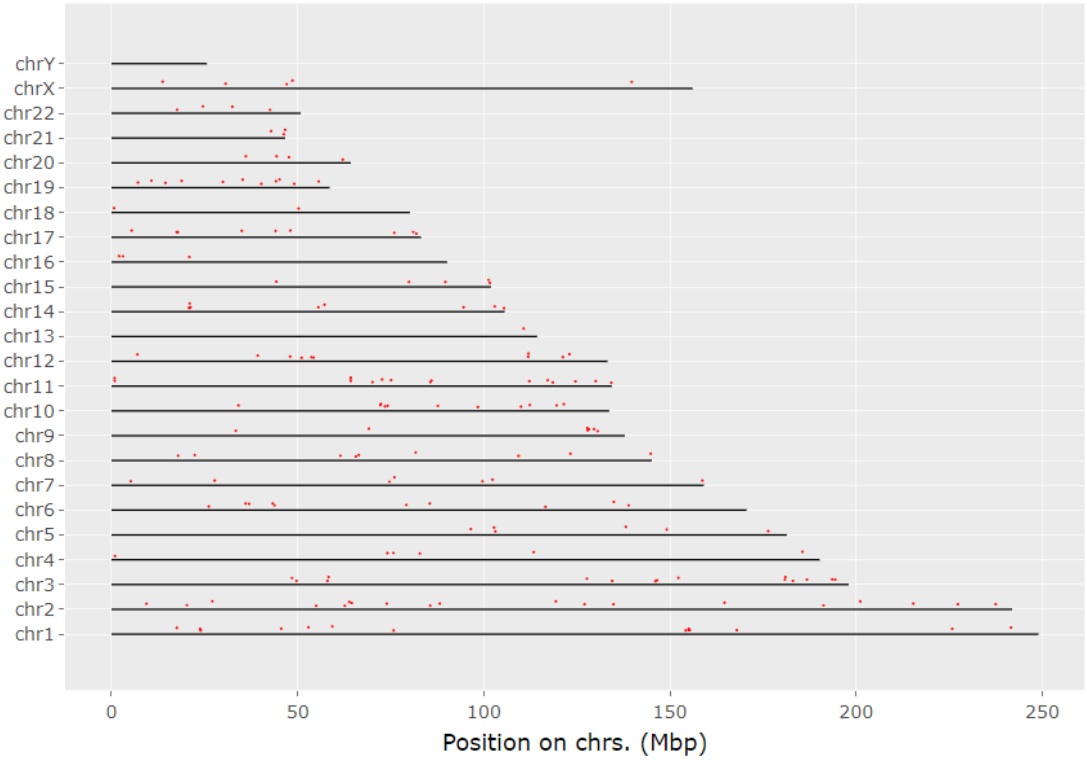


Fig (19): distribution of genes whose events are significantly differentially spliced between normal and tumoral samples among genome.

8.3. Interactive network of genes of interest

The network depicted in figure 20, which was created using STRING software, consists of 209 nodes and 249 edges. The average node degree is 2.38, and the average local clustering coefficient is 0.36PPI. Additionally, the enrichment p-value is 6.36e-08. This enrichment suggests that the proteins in the network are connected as a group, as they exhibit more interactions among themselves than would be anticipated in a random set of proteins with a similar size and degree distribution derived from the genome.

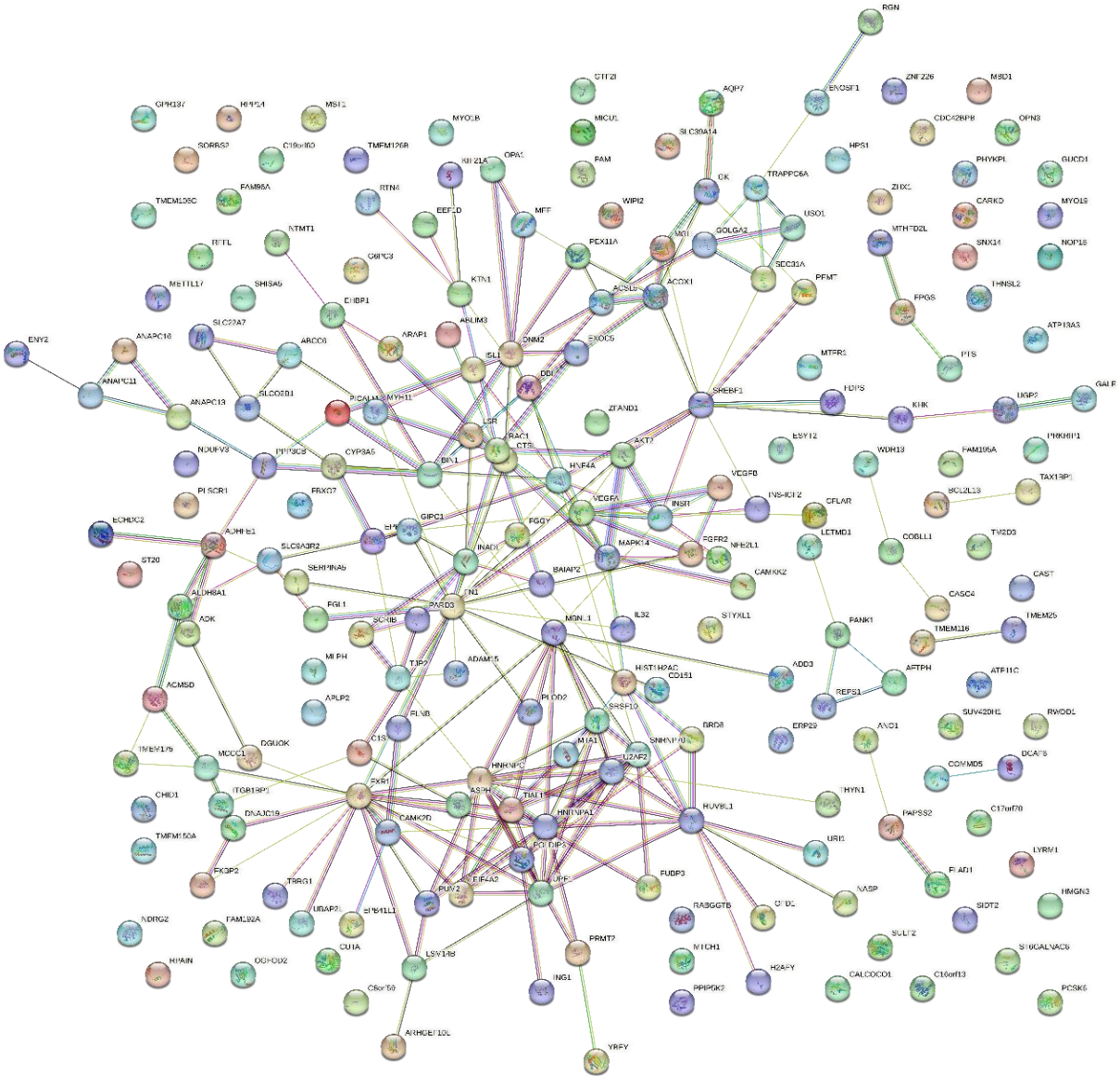


Fig (20): A network of interaction between genes whose events are significantly differentially spliced between normal and tumoral samples, generated in STRING

After exporting the network to Cytoscape, the MCODE function was utilized to identify significant clusters. The MCODE function employed a degree cutoff of 2, a node score cutoff of 0.2, a k score of 100, and a max depth of 100. The first and most crucial cluster comprises 10 nodes and 27 edges. The genes within this cluster are highlighted in yellow and can be found at the center of figure 21.

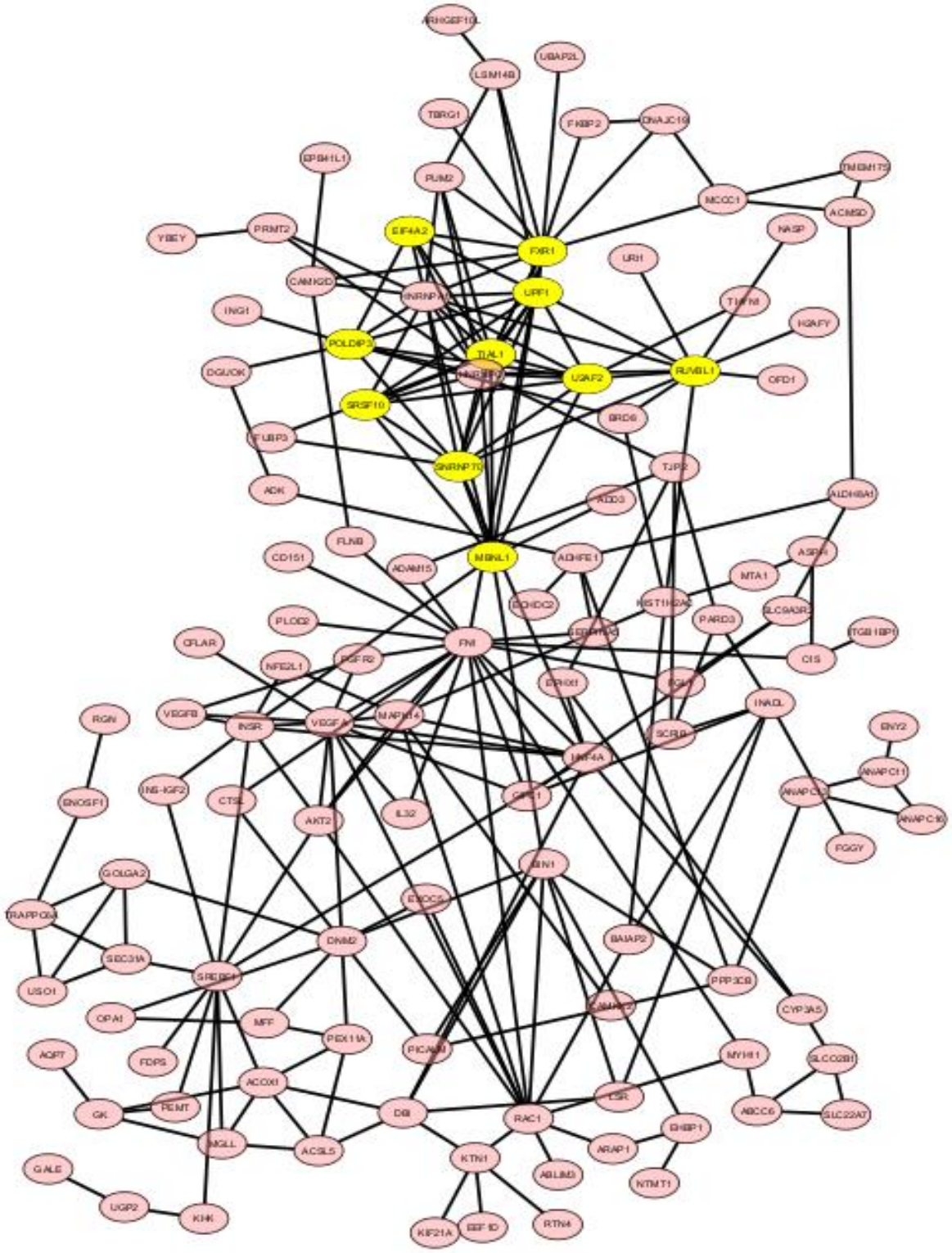


Fig (21): The main cluster of the genes network is highlighted in yellow generated via Cytoscape.

GO Biological Process enrichment analysis demonstrated that the genes in the cluster mainly corresponded to Negative regulation of mRNA splicing via spliceosome, Regulation of mRNA splicing via spliceosome, Regulation of mRNA processing, Regulation of RNA splicing, mRNA export from nucleus, mRNA-containing ribonucleoprotein complex export from nucleus, Ribonucleoprotein complex localization and mRNA transport.(Figure 22A+ 22B)

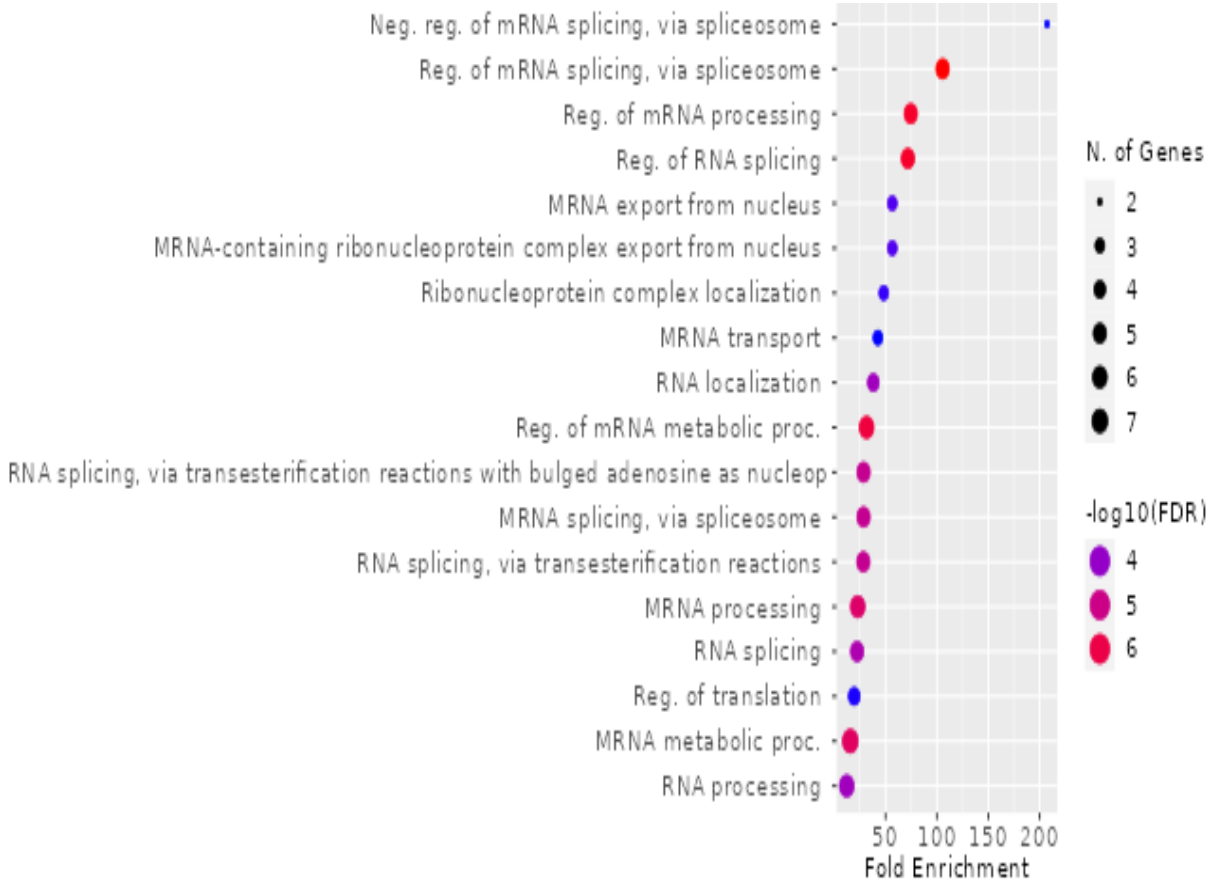


Fig (22A): Dot plot of GO Biological Process enrichment analysis of the gene cluster.

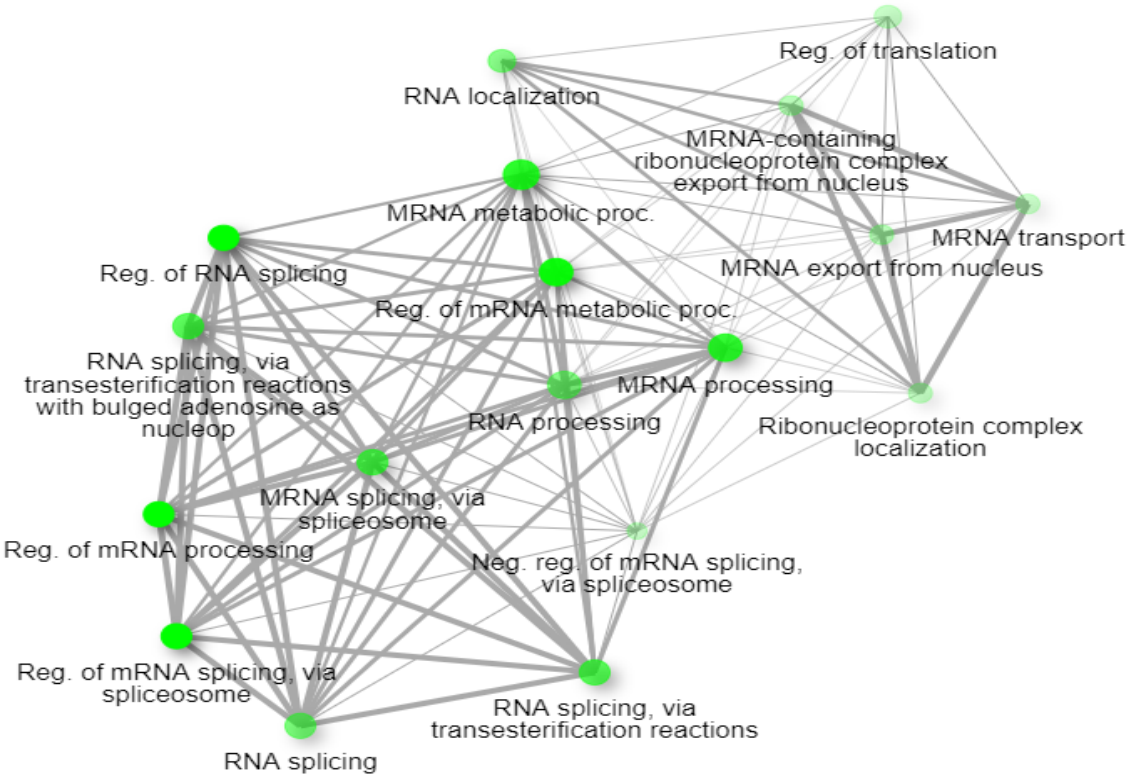


Fig (22B): A network of GO(BP) pathways of the gene cluster.

Through ShinyGO 0.76 access to STRING-db, we also retrieve a protein-protein interaction (PPI) network (figure 23) of the gene cluster that indicated there were 33 interactions between the genes with p value 8.09e-13.

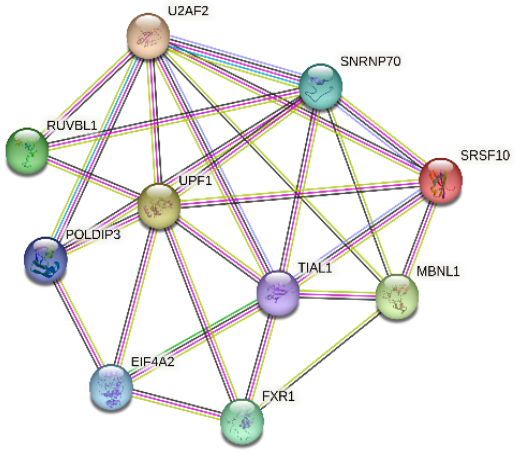


Fig (23): protein-protein interaction (PPI) network of the gene cluster

According to Jensen disease enrichment analysis, these genes were primarily found to be associated with relapsing-remitting multiple sclerosis, Myotonic dystrophy type 2, Localized scleroderma, Liposarcoma, and Sjogren's syndrome. This information is depicted in figure (24).

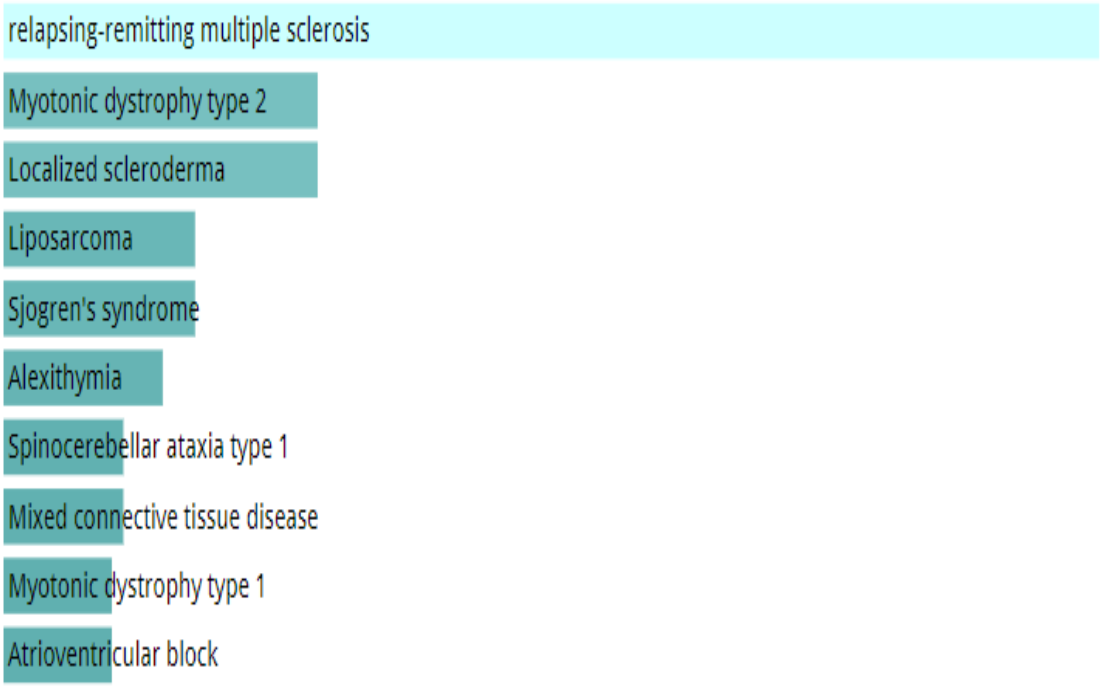


Fig (24): Bar graph of Jensen disease enrichment analysis of the gene cluster.

8.4. Differential Splicing and Survival Analysis

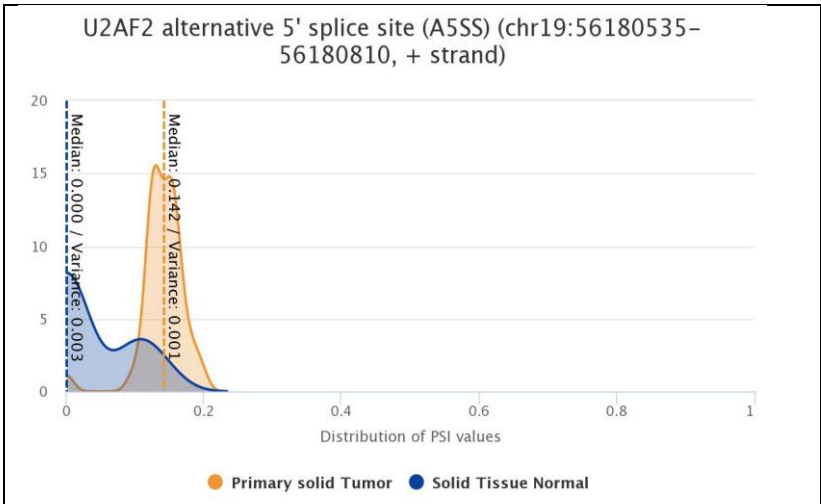
The inclusion levels of the 10 genes (table 3) were analyzed using Psichomics to determine if there was a significant difference in their splicing between normal and tumor TCGA CHOL samples.

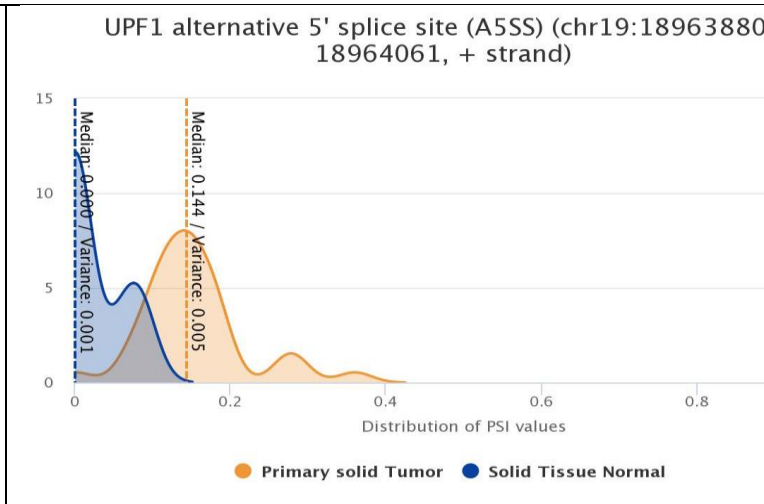
Table 3: Differentially AS events of the genes cluster and their genomic coordinates

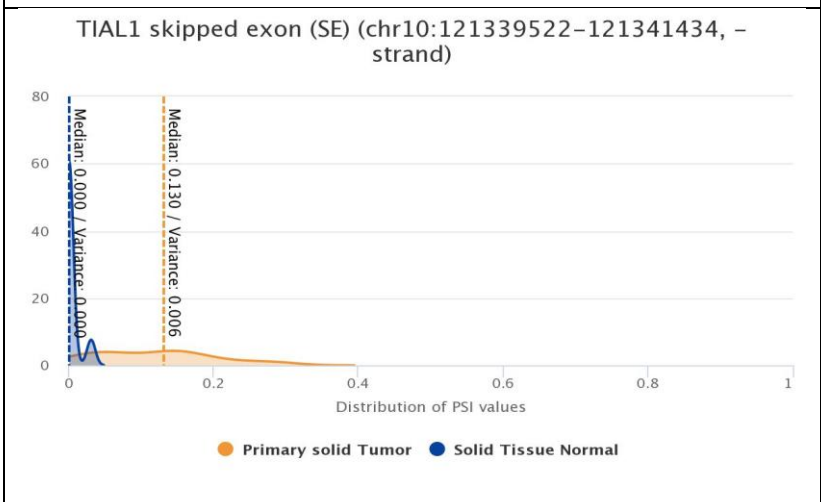
genomic coordinates	Strand	Chromosome	Event type	Gene
18964061_18963880	+	19	Alternative 5' splice site (A5SS)	UPF1
56180810_56180535	+	19	Alternative 5' splice site (A5SS)	U2AF2
121339588_121341434	-	10	Skipped exon (SE)	TIAL1
24294213_24298063	-	1	Alternative last exon (ALE)	SRSF10
49607891_49604728	+	19	Skipped exon (SE)	SNRNP70
127838257_127842427	-	3	Alternative first exon (AFE)	RUVBL1
42995799_42998776	-	22	Skipped exon (SE)	POLDIP3
152165409_152163328	+	3	Skipped exon (SE)	MBNL1
180693910_180688146	+	3	Skipped exon (SE)	FXR1
186503672_186502485	+	3	Skipped exon (SE)	EIF4a2

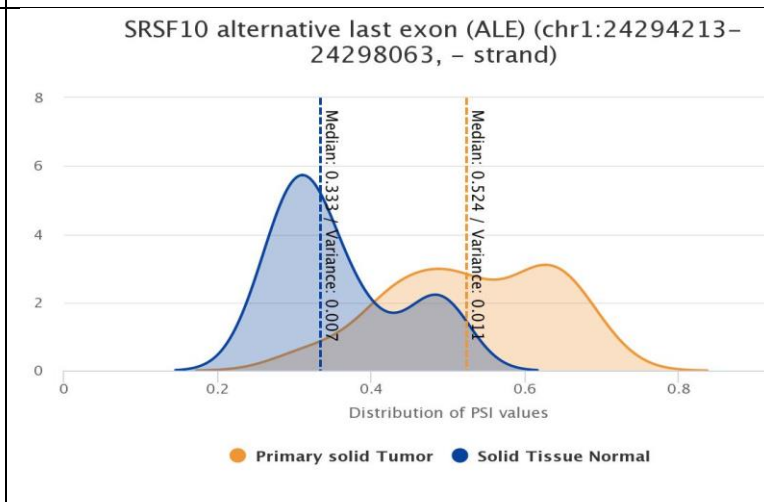
PSI distributions in tumor stage I and normal samples are depicted in the density plots in table 4:

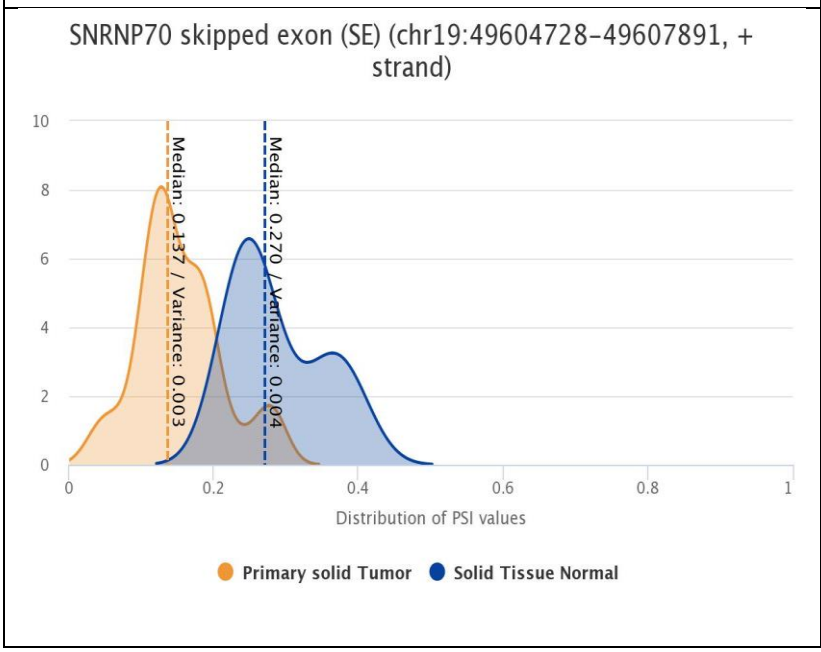
Table 4: Table of density plots of PSI distributions (of the 10 genes cluster) in tumoral and normal samples.

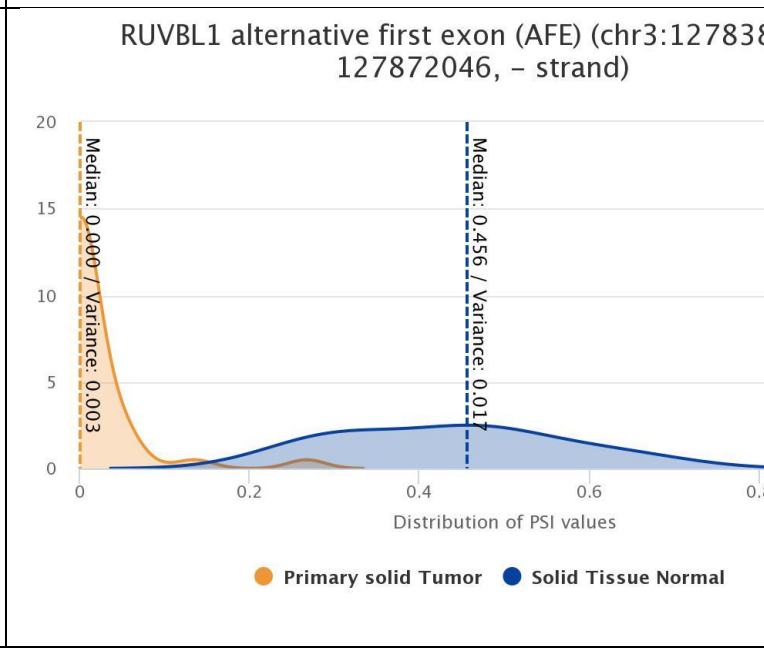


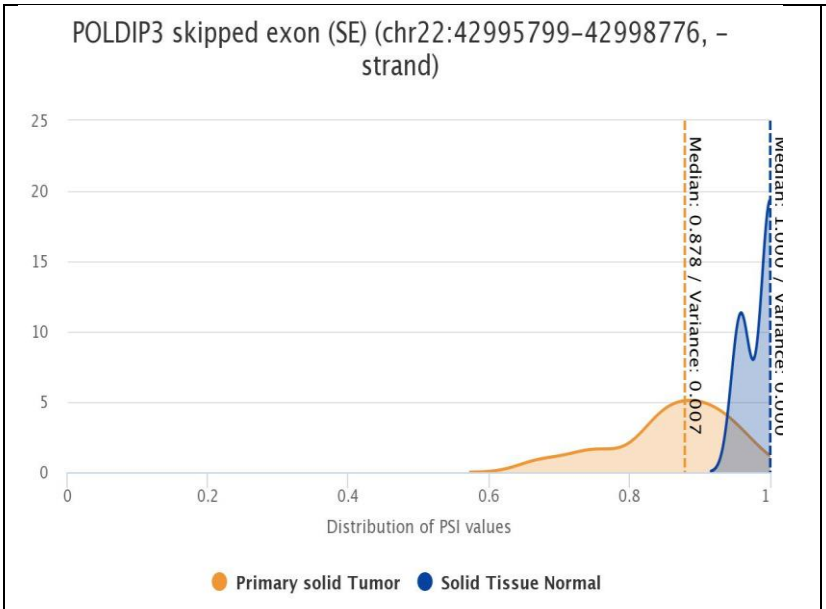


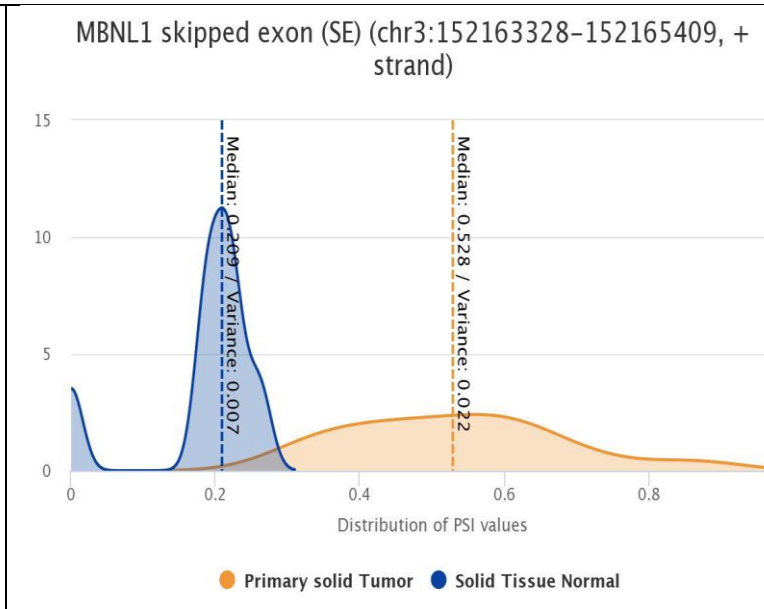


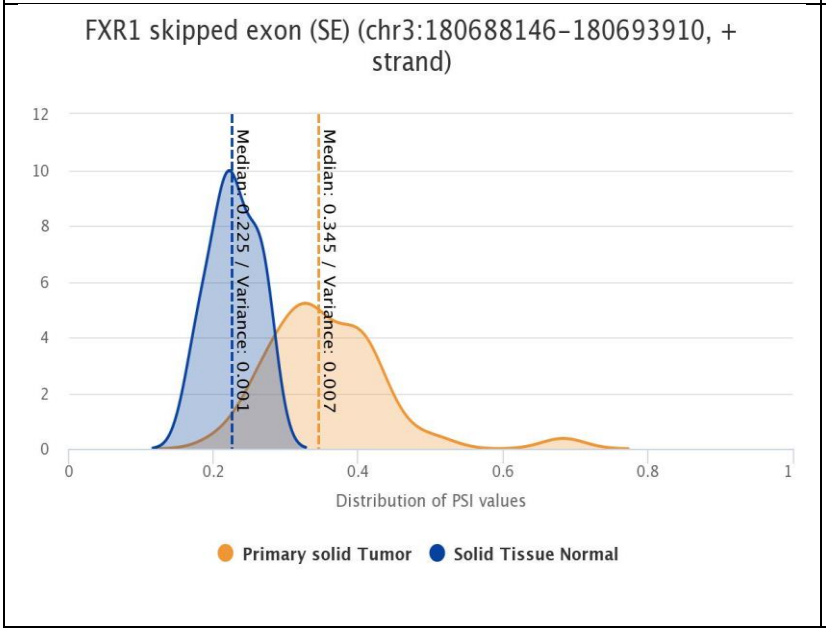


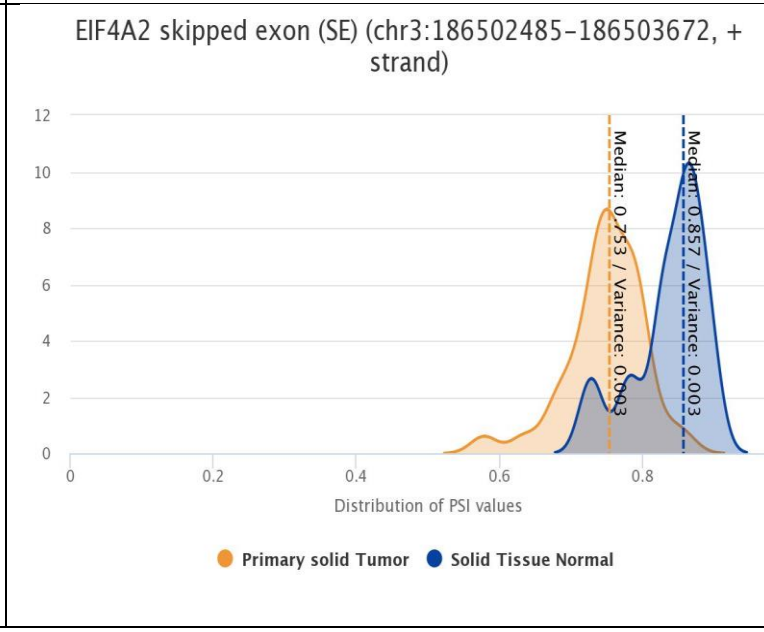






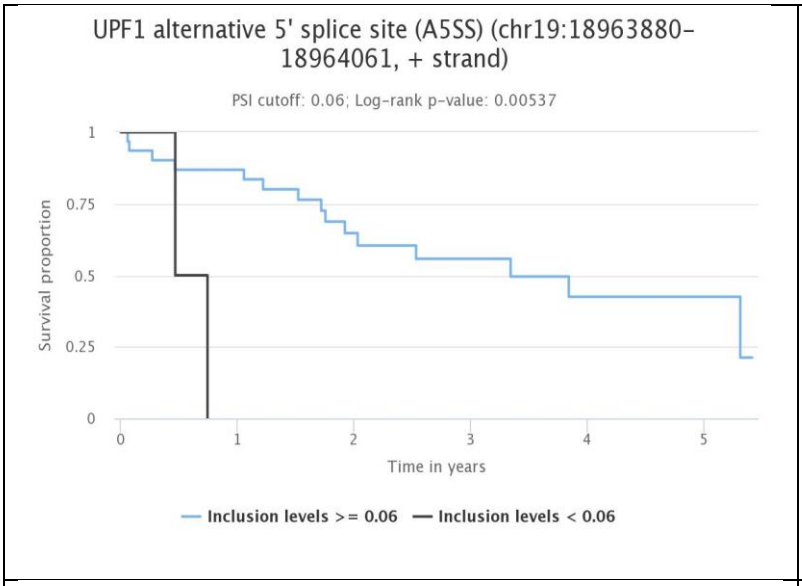


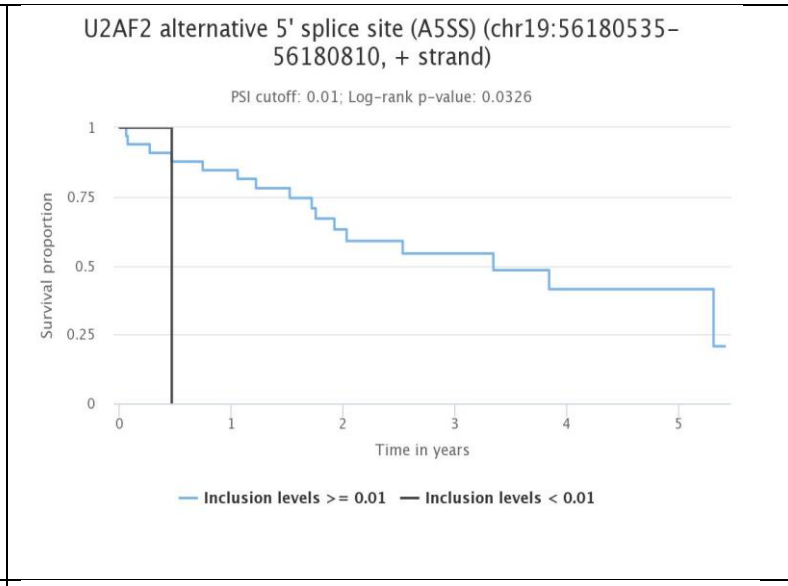


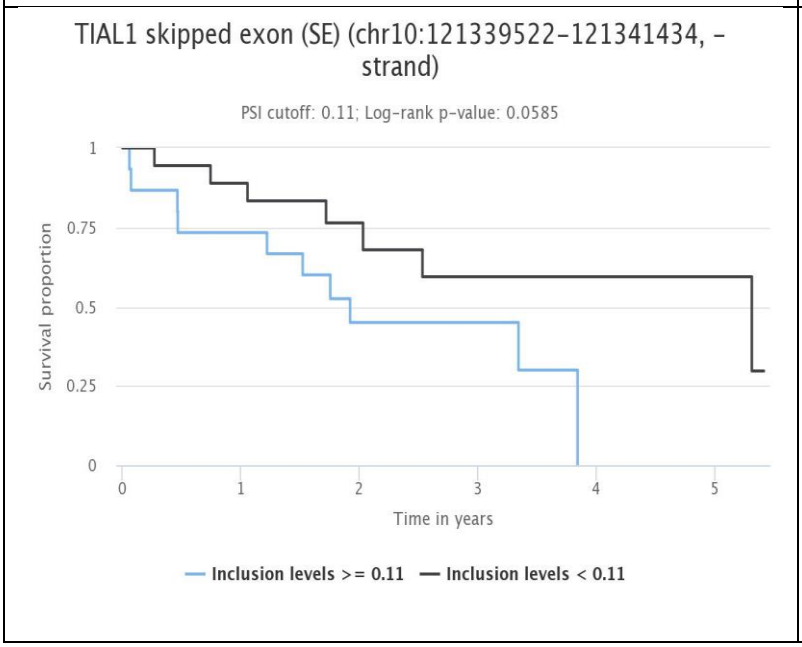


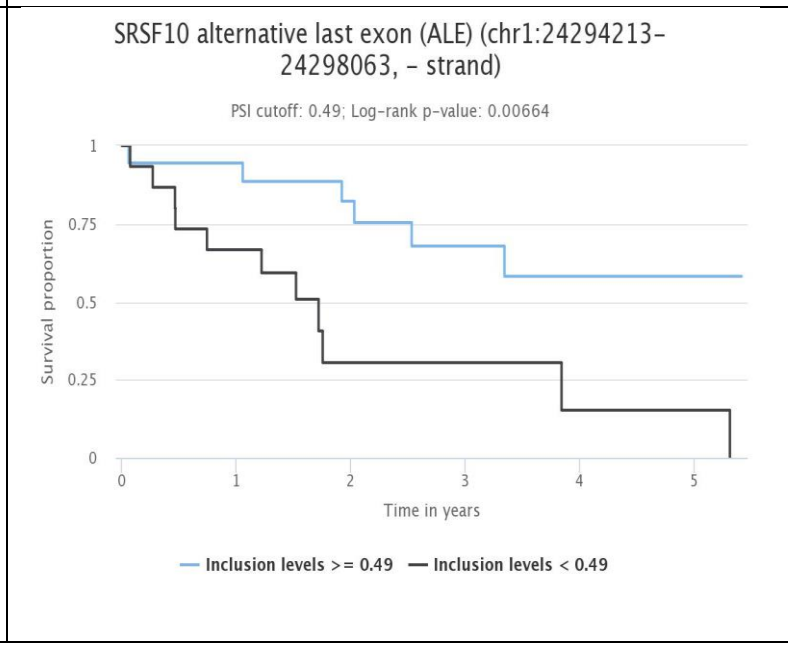
To study the impact of alternative splicing events on prognosis, Kaplan-Meier curves were plotted for groups of patients separated by the optimal PSI cutoff for each alternative splicing event that maximizes the significance of group differences in survival analysis. The purpose was to identify genes that could serve as prognostic biomarkers for the disease. The prognostic significance of the alternative splicing events is demonstrated by the Kaplan-Meier survival curves presented in Table 5.

Table 5: Kaplan–Meier survival curves illustrating the overall survival probability of CHOL patients stratified according to inclusion levels of indicated AS events in the 10 genes

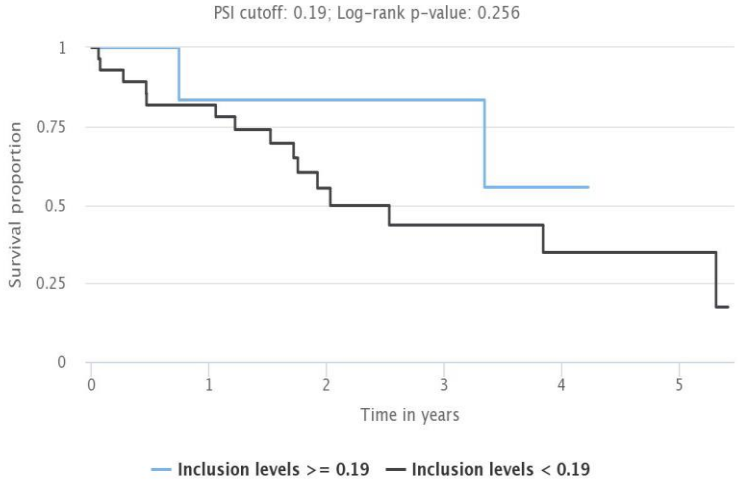




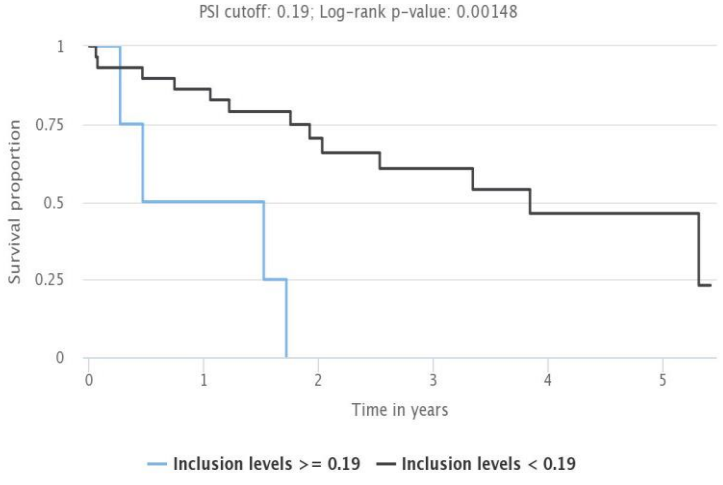




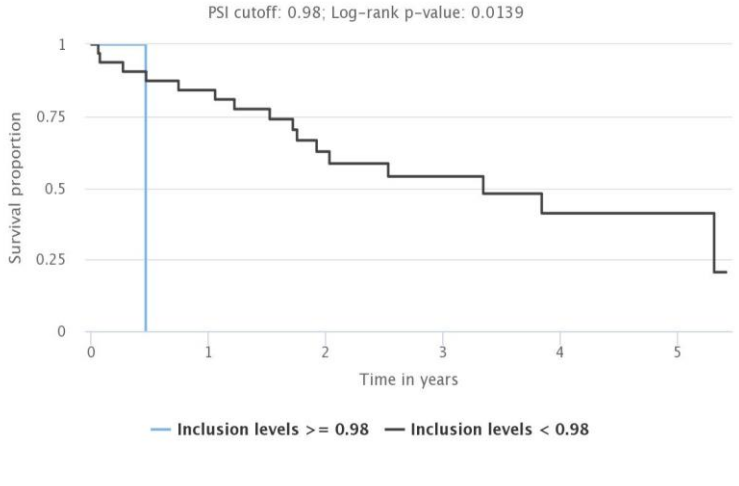
SNRNP70 skipped exon (SE) (chr19:49604728-49607891, + strand)



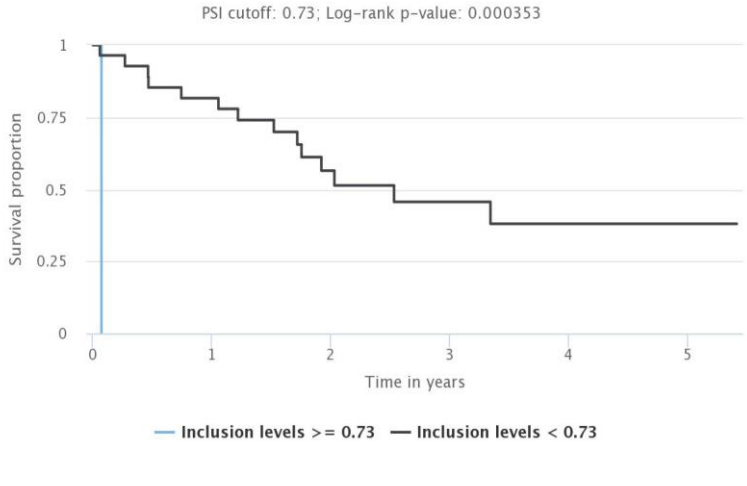
RUVBL1 alternative first exon (AFE) (chr3:127838257-127872046, - strand)



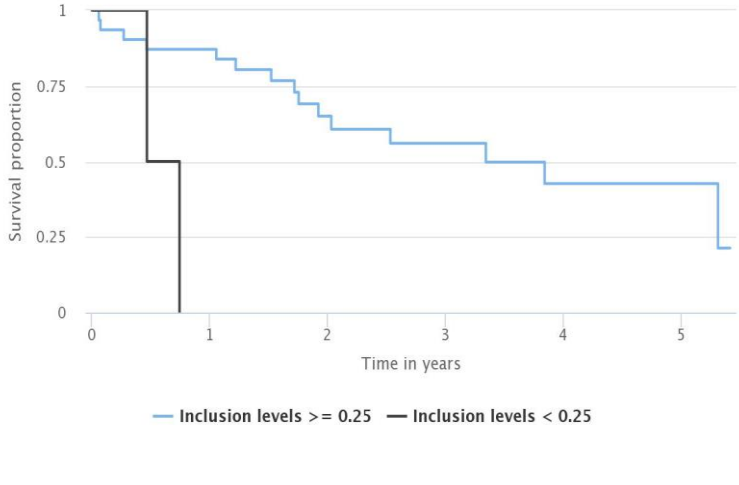
POLDIP3 skipped exon (SE) (chr22:42995799-42998776, - strand)



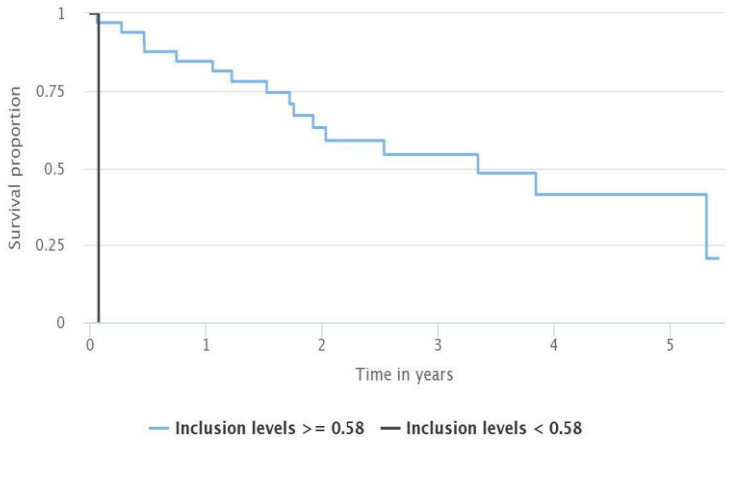
MBNL1 skipped exon (SE) (chr3:152163328-152165409, + strand)



FXR1 skipped exon (SE) (chr3:180688146-180693910, + strand)



EIF4A2 skipped exon (SE) (chr3:186502485-186503672, + strand)



8.5. Basic information related significant AS events

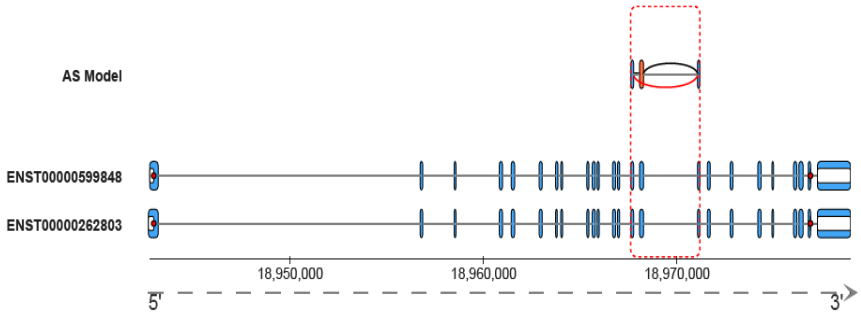
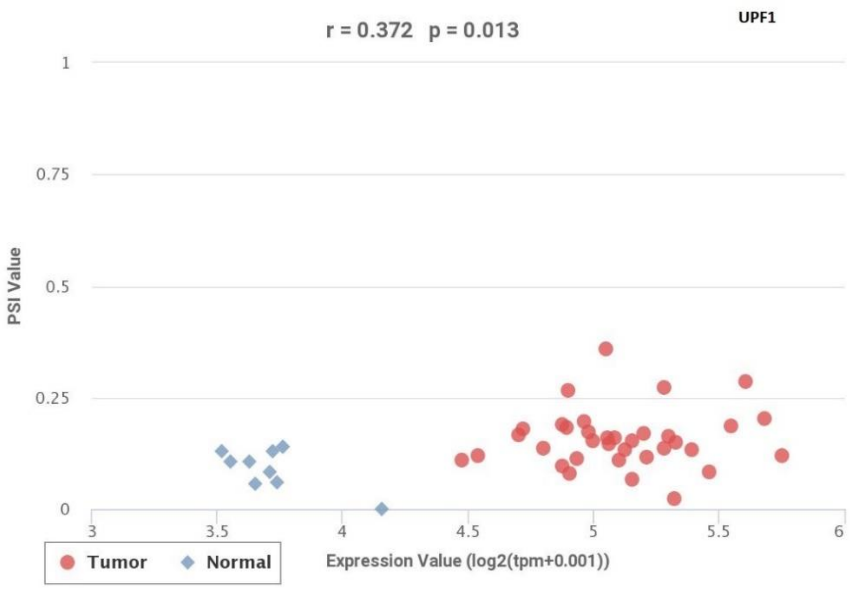
Basic information about the 10 genes obtained from ASCancer Atlas were summarized in tables from 6A to 6J. These tables also include the splicing model of the alternative splicing (AS) event in protein-coding transcripts, and plots that compare the PSI value and expression value for each significant splicing event in CHOL TCGA samples.

Tables 6A to 6J: Basic information related to each of the 10 genes, splicing model, and plots of PSI value and expression in normal and tumoral samples

1. UPF1

Table 6A

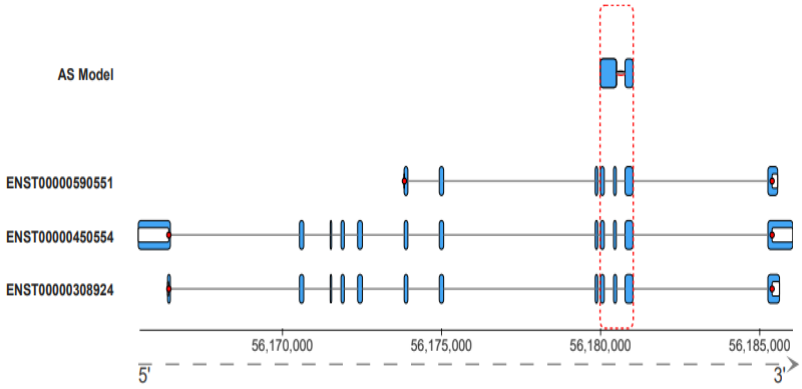
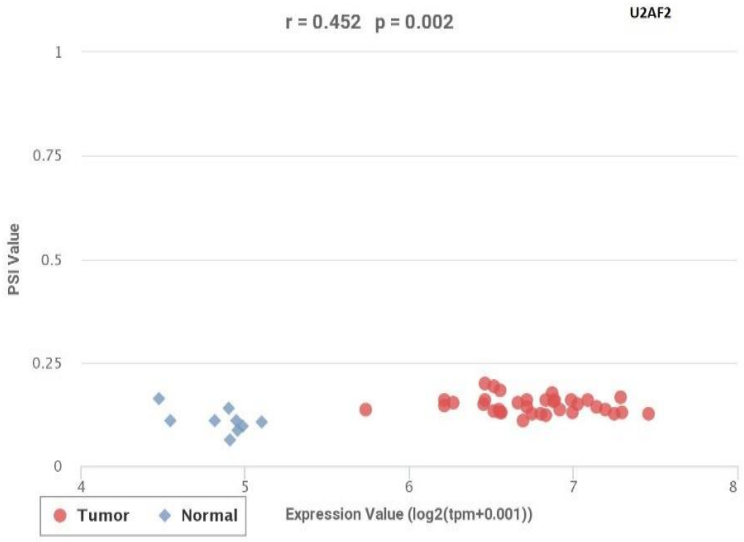
Ensembl Gene ID	ENSG00000005007
Approved Name	UP-Frameshift protein 1
Locus Type	Protein-coding gene
Gene Summary	This gene encodes a protein that is a component of a multiprotein complex responsible for both the export of mRNA from the nucleus and the surveillance of mRNA.

AS model	
Splicing Expression	

2. U2AF2

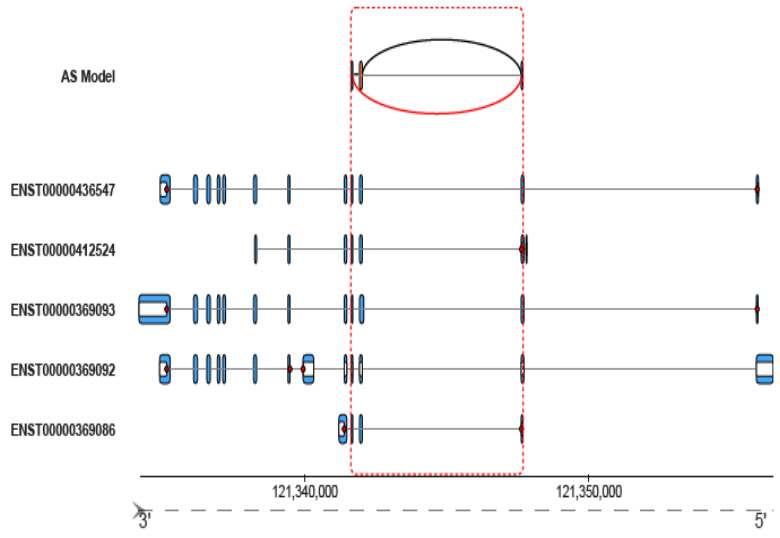
Table 6B

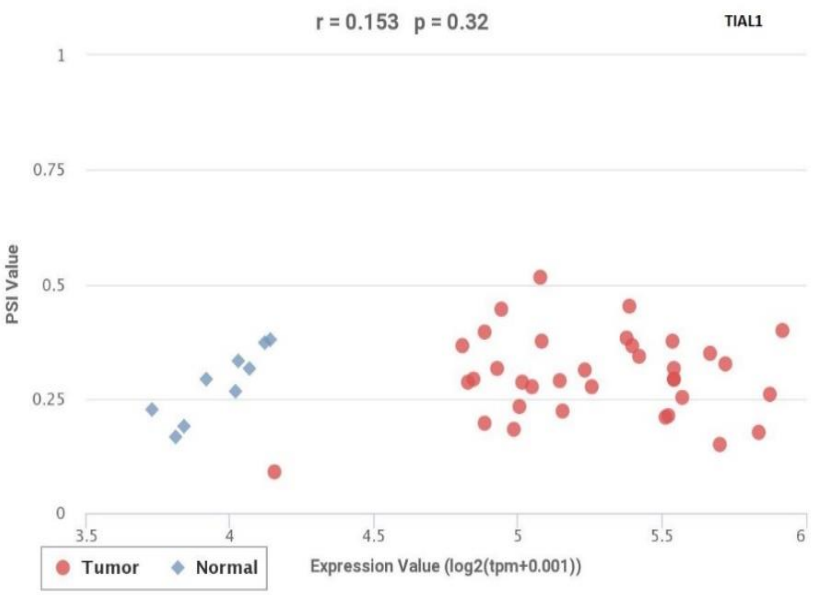
Ensembl Gene ID	ENSG00000063244
Approved Name	U2 small nuclear RNA auxiliary factor 2

Locus Type	Protein-coding gene
Gene Summary	U2 auxiliary factor (U2AF) is a splicing factor that consists of a large and a small subunit. The U2AF2 gene encodes the U2AF large subunit, which contains a region that can specifically bind to RNA with three RNA recognition motifs and an Arg/Ser-rich domain that is essential for splicing. The large subunit of U2AF binds to the polypyrimidine tract of introns at an early stage of spliceosome assembly.
AS model	
Splicing Expression	

3. TIAL1

Table 6C

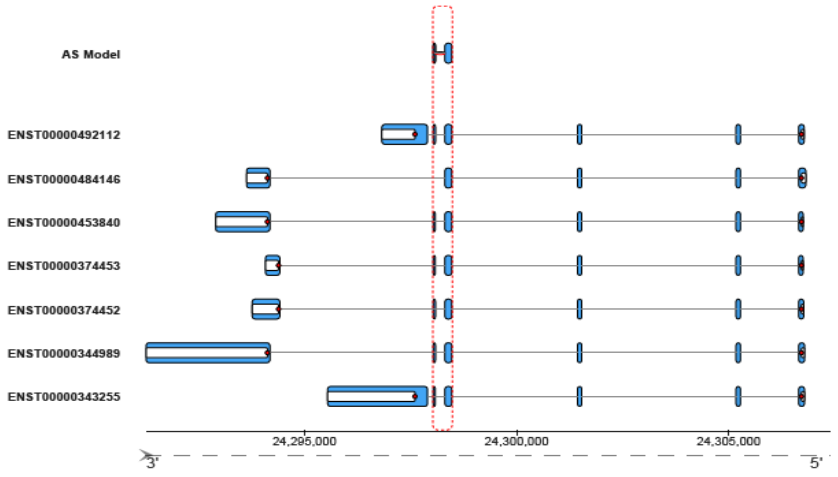
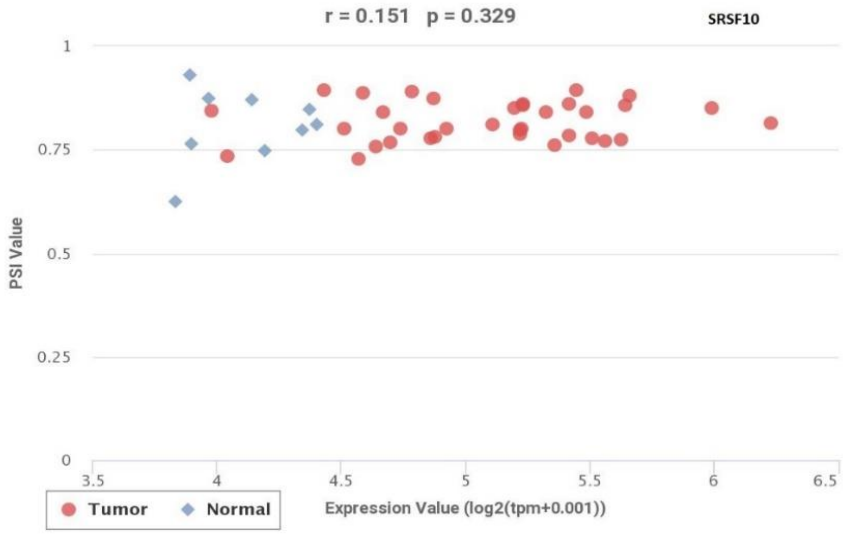
Ensembl Gene ID	ENSG00000151923
Approved Name	TIA1 cytotoxic granule associated RNA binding protein like 1
Locus Type	Protein-coding gene
Gene Summary	The protein encoded by this gene is an RNA-binding protein. This protein plays a role in regulating several activities, including translational control, splicing, and apoptosis. It contains three RNA recognition motifs (RRMs) and is capable of binding to adenine and uridine-rich elements in mRNA and pre-mRNAs of various genes.
AS model	

Splicing Expression	

4. SRSF10

Table 6D

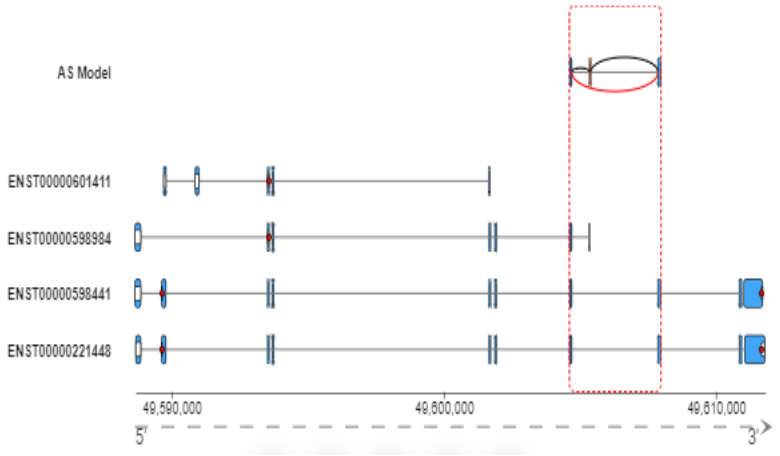
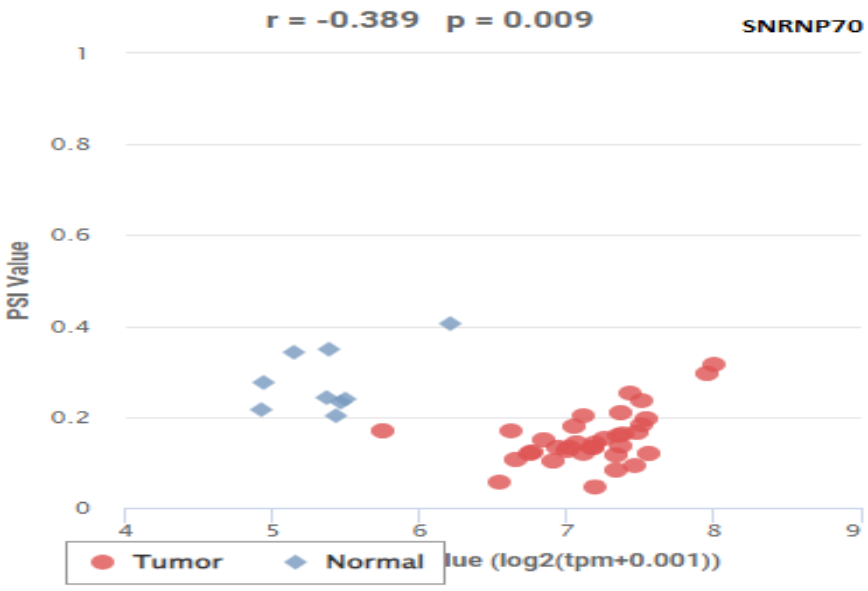
Ensembl Gene ID	ENSG00000188529
Approved Name	Serine and arginine rich splicing factor 10
Locus Type	Protein-coding gene
Gene Summary	This gene product belongs to the serine-arginine family, which plays a role in both constitutive and regulated RNA splicing.

AS model	
Splicing Expression	

5. SNRNP70

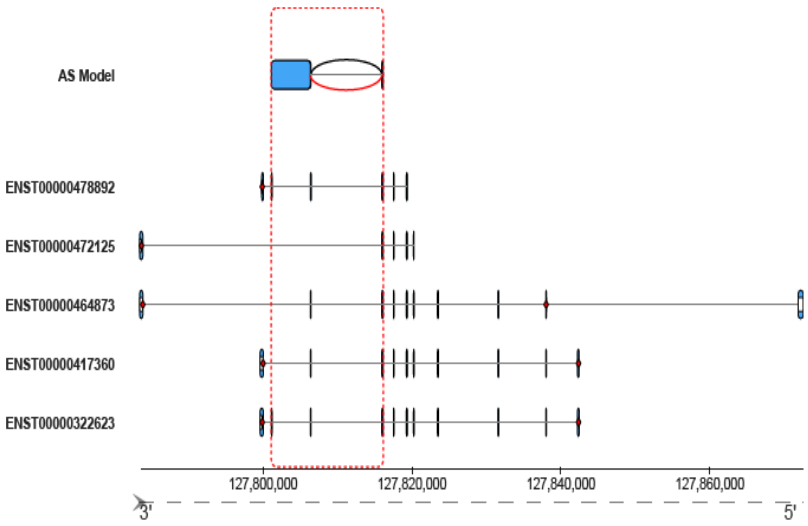
Table 6E

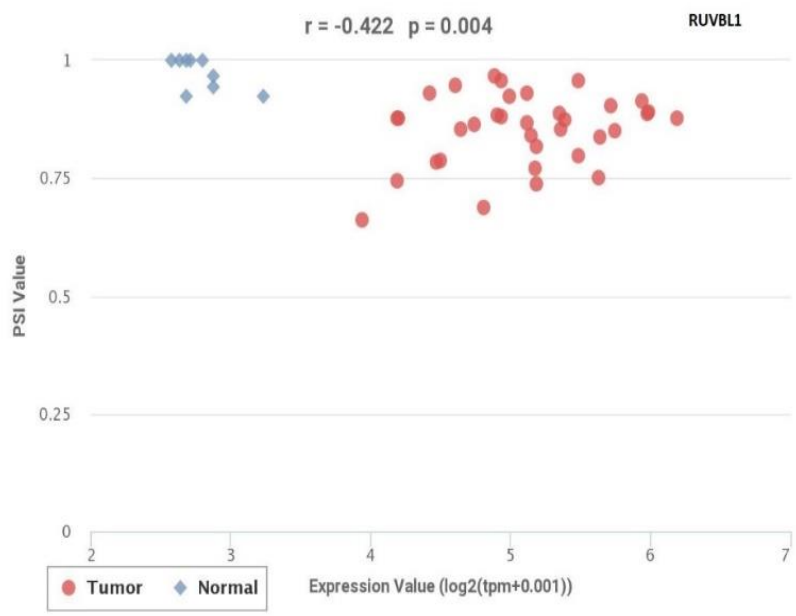
Ensembl Gene ID	ENSG00000104852
Approved Name	Small nuclear ribonucleoprotein U1 subunit 70

Locus Type	Protein-coding gene
Gene Summary	This gene product is a member of the serine-arginine family, which is involved in both constitutive and regulated RNA splicing. It is a small nuclear ribonucleoprotein that associates with U1 spliceosomal RNA, forming the U1snRNP, which is a core component of the spliceosome. Located in the nucleoplasm.
AS model	
Splicing Expression	

6. RUVBL1

Table 6F

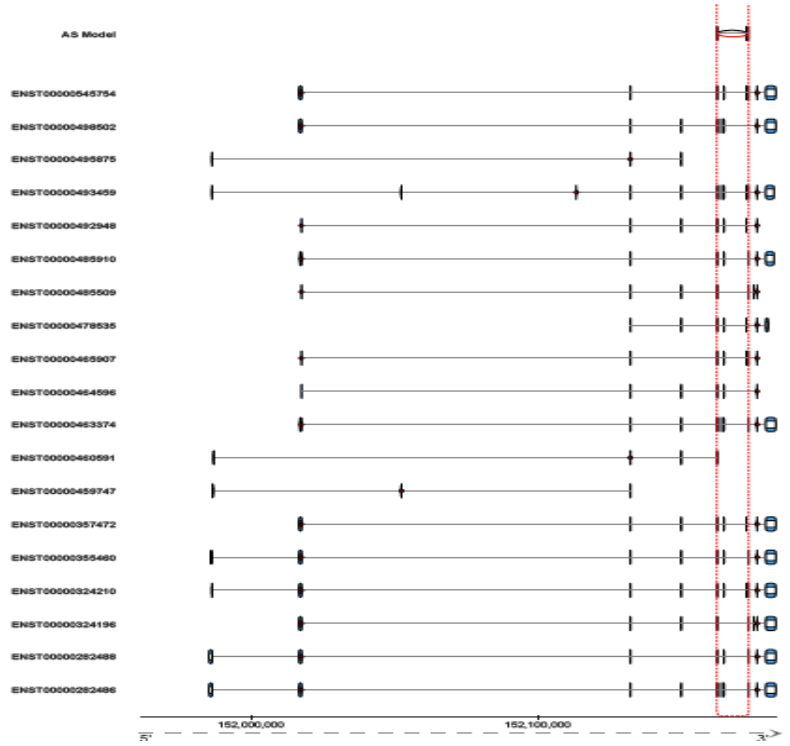
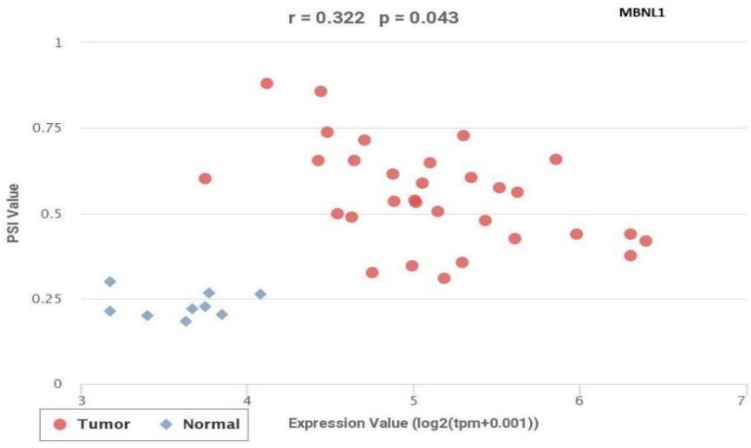
Ensembl Gene ID	ENSG00000175792
Approved Name	RuvB like AAA ATPase 1
Locus Type	protein-coding gene
Gene Summary	This gene encodes a protein that possesses both DNA-dependent ATPase and DNA helicase activities. It is a member of the ATPases associated with diverse cellular activities. The protein interacts with multiple multi subunit transcriptional complexes and protein complexes involved in ATP-dependent remodeling and histone modification.
AS model	

Splicing Expression	

7. MBNL1

Table 6G

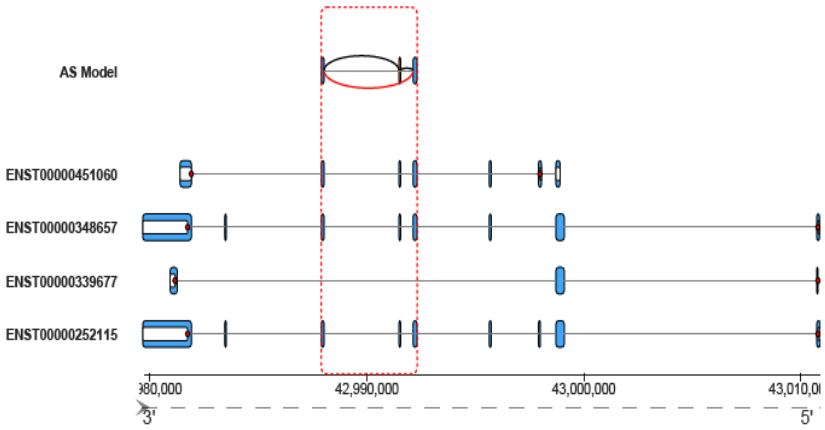
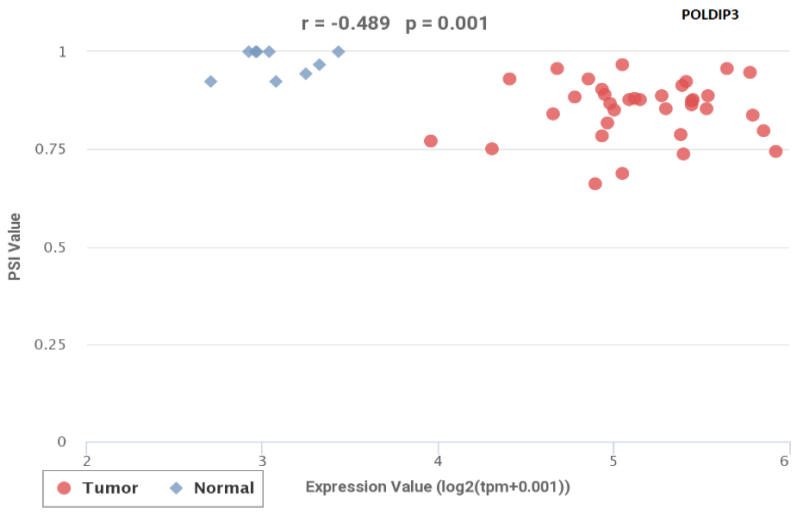
Ensembl Gene ID	ENSG00000152601
Approved Name	Muscle blind like splicing regulator 1
Locus Type	Protein-coding gene
Gene Summary	This gene encodes a C3H-type zinc finger protein that belongs to the muscle blind protein family. It was first identified in Drosophila melanogaster. The protein plays a role in regulating alternative splicing of pre-mRNAs.

AS model	
Splicing Expression	

8. POLDIP3

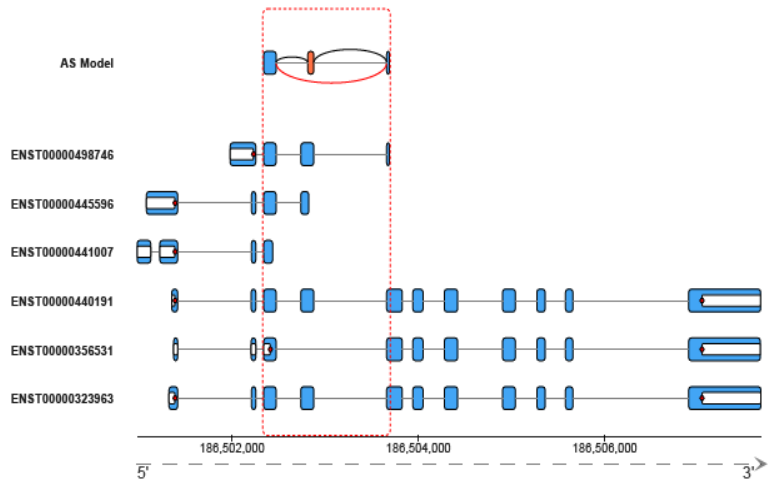
Table 6H

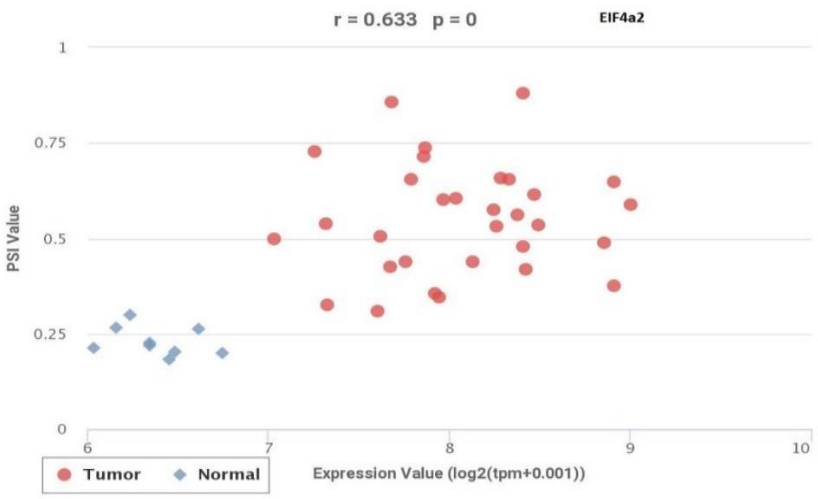
Ensembl Gene ID	ENSG00000100227

Approved Name	DNA polymerase delta interacting protein 3
Locus Type	Protein-coding gene
Gene Summary	This gene encodes a protein that contains an RNA recognition motif (RRM) and is involved in the regulation of translation. It functions by recruiting ribosomal protein S6 kinase beta-1 to mRNAs.
AS model	
Splicing Expression	

9. EIF4A2

Table 6I

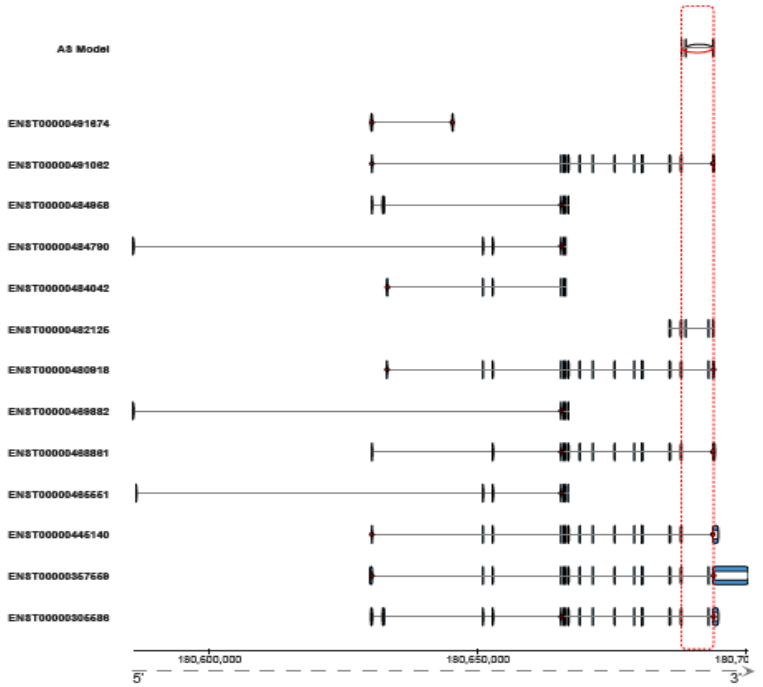
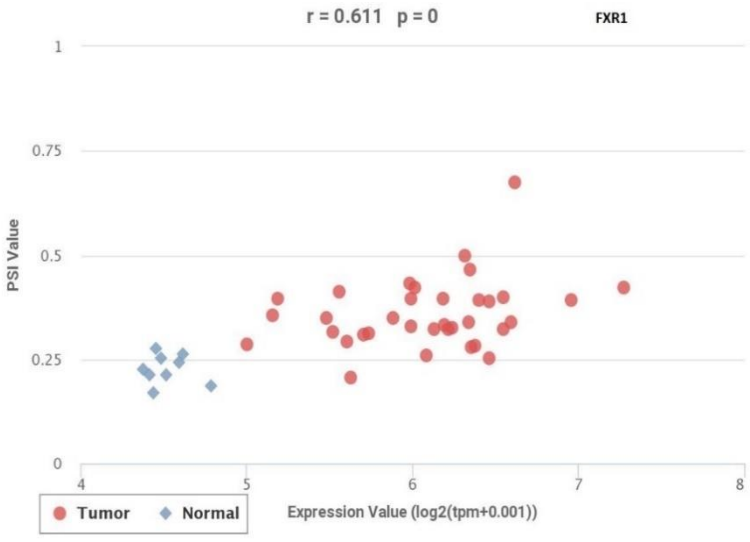
Ensembl Gene ID	ENSG00000156976
Approved Name	Eukaryotic translation initiation factor 4A2
Locus Type	Protein-coding gene
Gene Summary	The translation of mRNA is a complex process that involves initiation, elongation, and termination. Essential factors for translation are members of the eukaryotic initiation factor 4A (eIF4A) family. The different isoforms of eIF4A have been named as eIF4A1 (DDX2A), eIF4A2 (DDX2B), and eIF4A3 (DDX48). Both eIF4A1 and eIF4A2 are involved in the initiation of translation. There is evidence suggesting that various eukaryotic initiation factors are closely associated with the development and prognosis of different types of human cancers.
AS model	

Splicing Expression	

10. FXR1

Table 6J

Ensembl Gene ID	ENSG00000114416
Approved Name	FMR1 autosomal homolog 1
Locus Type	Protein-coding gene
Gene Summary	The gene product is an RNA binding protein that interacts with the proteins FMR1 and FXR2. These proteins move between the nucleus and cytoplasm and bind to polyribosomes, particularly the 60S ribosomal subunit.

AS model	
Splicing Expression	

8.6. Targeted editing on EIF4A2 gene

The CRISPOR web tool was utilized to identify specific guide RNAs for the significant alternative splicing (AS) events associated with CHOL patients. For instance, we chose the EIF4A2 gene as an example. This gene plays a role in the initiation of translation, and most of the evidence has indicated that different eukaryotic initiation factors are closely linked to the development and prognosis of various types of human cancers. From our findings, we observed a significant alternative splicing event in this gene that is associated with patients diagnosed with CHOL. This AS event is located on the third chromosome, specifically on the forward strand, within the genomic coordinates of 186502485-186502751. We obtained the sequence of this genomic location, which is 267 base pairs long. By querying CRISPOR for guide RNAs, we discovered 24 potential guide sequences. The figure (25) displays these sequences along with their corresponding PAM sites and the expected cleavage position.

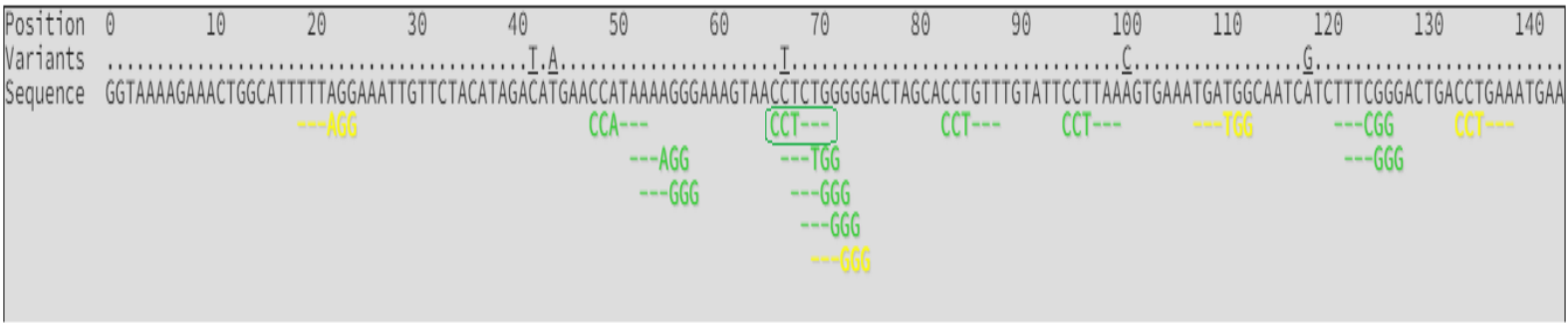


Fig (25): PAM sites and the expected cleavage position in the genome, colors green and yellow indicate high and medium specificity of the PAM's guide sequence in the genome.

We selected the guide with the highest specificity score

Guide Sequence: AACAGGTGCTAGTCCCCCAG

PAM (Protospacer Adjacent Motif): AGG

Restriction Enzymes: LpnPI, MaeIII

The specificity score is 91, which is a prediction of how likely an RNA guide sequence for this target may cause off-target cleavage in other parts of the genome. A higher specificity score (>50) indicates a better guide.

The efficiency score is 79, which predicts how well the target can be cut by its RNA guide sequence.

The out-of-frame score is 69, indicating the likelihood of the guide causing out-of-frame deletions.

The off-target mismatch counts represent the number of possible off-targets in the genome for each number of mismatches. The sequence '0-0-1-6-77' means that the target matches 0 locations in the genome with no mismatches, 0 locations with 1 mismatch, 1 location with 2 mismatches, 6 locations with 3 mismatches, and 77 locations with 4 mismatches.

From the previous results, we can observe that the EIF4A2 gene is expressed at high levels in tumor samples. Additionally, the alternative splicing (AS) event associated with this gene is up-regulated. Therefore, if our goal is to decrease the expression of the EIF4A2 gene, we should design the gRNA to target the promoter elements of the gene. On the other hand, if our desired genetic manipulation is to introduce a specific sequence change, such as modifying the splicing position, the gRNA should be designed to target the region where the desired edit should be made and, in this case, a repair template will be required to facilitate the desired genetic manipulation.

Once the RNA guide and appropriate vector have been selected, the next step involves delivering the gRNA(s) into the cells. This can be done through ex-vivo or in-vivo methods. In ex-vivo therapy, cells are isolated and edited outside of the body before being transplanted back. In in-vivo therapy, genetic materials are directly injected into the body. The final step

is to validate the genetic modification. CRISPR editing can result in various genotypes within the cell population. Some cells may remain wild type due to a lack of gRNA and/or Cas9 expression, or inefficient target cleavage in cells expressing both Cas9 and gRNA. There are several common methods to confirm that the desired edit is present in specific cells, such as PCR amplification and gel electrophoresis, or PCR amplification and next-generation sequencing. [40, 41]

The flowchart in figure (26) summarizes the workflow of Crispr technology.

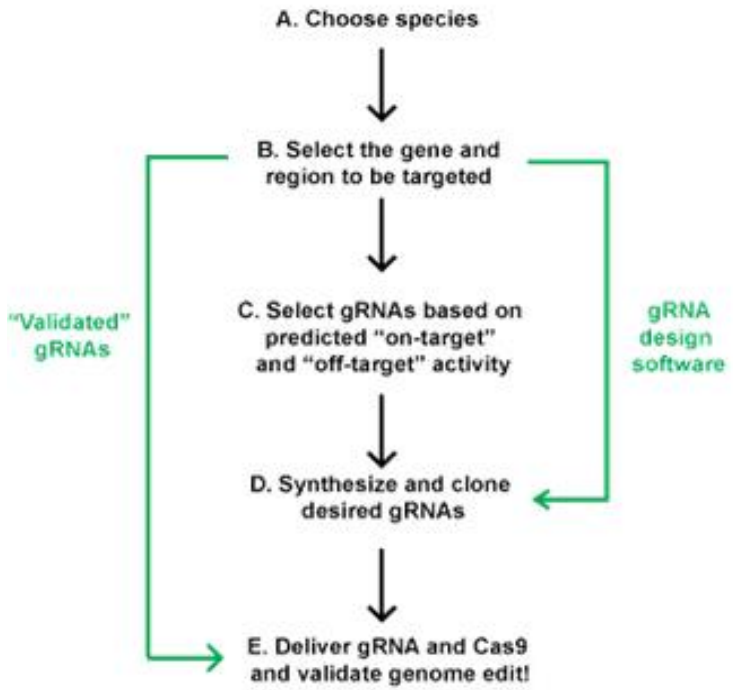


Fig (26): Flowchart of the Crispr technology workflow

9. Discussion

Cholangiocarcinoma, the second most common primary liver cancer worldwide, has a relatively low incidence in most high-income countries (0.3-2 cases per 100,000 people). However, in certain regions like Southeast Asia, where a liver fluke parasite is prevalent, the incidence can be much higher (up to 40-fold greater). In addition, gallstones and inflammatory conditions of the digestive tract can increase the risk of bile duct cancer. [42, 43]

For early-stage tumors, surgery is the only therapeutic option that offers a chance of cure. If imaging tests indicate a good chance of removing the entire tumor, surgery may be performed. However, cholangiocarcinoma is often asymptomatic during the early stages, leading to late-stage diagnosis where surgery cannot completely remove metastatic cancer [44, 45]. Therefore, there is an urgent and growing need to understand the pathogenesis of this tumor and develop prognosis and diagnosis methods.

In the present study, a series of analyses were conducted using Psichomics R package to explore downregulated and upregulated genes associated with cholangiocarcinoma. The normalized RNA-seq data from healthy subjects and cholangiocarcinoma patients were derived from TCGA database, The goal was to identify significantly differentially spliced events between tumoral and normal samples. A total of 283 events were found to be differentially spliced based on specific criteria ($|\Delta \text{median PSI}| \geq 0.1$ and Wilcoxon rank-sum test with Benjamini–Hochberg (FDR) adjustment ≤ 0.01). These events were considered as alternative splicing events associated with overall survival in CHOL and classified into three groups: biological processes, molecular functions, and cellular components, using Gene Ontology (GO) terms. Additionally, KEGG pathway enrichment analysis was conducted using ShinyGO 0.76.

The genes with OS-related AS events in CHOL were found to be mainly involved in Peroxisome fission, Hepatocyte proliferation, Epithelial cell proliferation involved in liver morphogenesis, and were primarily enriched in Cytoplasmic stress granule, Cell cortex, Envelope, and Mitochondrion. They were also associated with functions such as Vascular endothelial growth factor receptor 1 binding, L-ascorbic acid binding, Cadherin binding, and Cell adhesion molecule binding.

Furthermore, pathway analysis revealed that these genes are mainly involved in Galactose metabolism, Regulation of lipolysis in adipocytes, VEGF signaling pathway, and PPAR signaling pathway.

Using bioinformatics tools including STRING and Cytoscape analyses of all genes with significant prognostic values identified the following genes to be located at the center of the gene network: UPF1, U2AF2, TIAL1, SRSF10, SNRNP70, RUVBL1, POLDIP3, MBNL1, FXR1 and EIF4a2. These genes are primarily enriched in biological processes related to the regulation and negative regulation of mRNA splicing via spliceosome, as well as the regulation of mRNA processing and RNA splicing. These processes are directly associated with alternative splicing.

Additionally, these genes as a group have been reported to be involved in various conditions such as relapsing-remitting multiple sclerosis, Myotonic dystrophy type 2, Localized scleroderma, Liposarcoma, and Sjogren's syndrome. However, their association with CHOL has not been previously reported. Therefore, these genes may open up new avenues for research, as their clustering results from a network that accurately screens significant AS events identified through the Wilcoxon rank-sum test with Benjamini–Hochberg (FDR) adjustment and median change.

Differential splicing and PSI distributions of the 10 AS events across tumoral and normal samples revealed that 6 events were up-regulated in tumoral samples and 4 were down- regulated when compared to normal samples, Kaplan–Meier survival analysis for overall survival probability of CHOL patients stratified according to inclusion levels of indicated AS events in the 10 genes revealed 2 AS events (TIAL1 and SNRNP70) had p value >0.05 thus can't be used as tumor biomarkers for this tumor, meanwhile the other 8 AS events had a significant p value thus had a high prognostic value and might be novel biomarkers.

In looking for each of the eight genes individually, we found that there are some genes that have not been previously well studied in the context of cholangiocarcinoma, including eIF4A2 which is one of the initiation factors of mRNA translation. The dysregulation and aberrant expression of eIF4A isoforms have been found in various tumor tissues including gastric, lung, colorectal and breast cancer [46]. U2AF2 Plays a role in pre-mRNA splicing and 3'-end processing its alterations have been associated with a variety of human cancers such as lung cancer [47], and according to reports, alterations in U2AF2 splicing factor have very low frequencies among hematological tumors [8]. Also, FXR1 which is an RNA binding protein that is related to poor prognosis in some cancers, including ovarian cancer, breast cancer, and head and neck squamous carcinoma [48].

In the other hand, we found some genes that have been shown to be involved in cholangiocarcinoma including UPF1, SRSF10, RUVBL1, POLDIP3 and MBNL1. Currently, UPF1 is known to have a crucial role in promoting cell proliferation and differentiation [49], additionally, there is growing evidence suggesting that UPF1 could potentially be used as a biomarker for diagnosing and treating cancer in future clinical applications as UPF1 is dysregulated and plays a significant role in various types of

cancer, including hepatocellular [50], colorectal [51], gastric [52], pancreatic [53], thyroid [54], ovarian [55], and prostate cancer [56]

POLDIP3 (DNA Polymerase Delta Interacting Protein 3) is a binding partner and target of S6 kinase 1, regulates DNA replication, mRNA translation efficiency and cell growth. a study of alternative transcript of POLDIP3 demonstrated that this alternative transcript is upregulated and functions as a critical oncogene in hepatocellular carcinoma.[57]

Serine/arginine splicing factor 10 (SRSF10) is a member of the family of mammalian splicing regulators which have been implicated in the carcinogenesis and progress of a variety of cancers, SRSF10 has been found to have a strong association with overall survival and is significantly related to the prognosis of intrahepatic cholangiocarcinoma.[58]

RUVBL1, also named as Pontin is a protein that is involved in various biological processes, such as regulating gene transcription, modifying chromatin structure, detecting and repairing DNA damage. It has been found to be overexpressed in several types of cancer, interestingly RUVBL1 has been found to have high expression and a significant prognostic value for patients with Hilar Cholangiocarcinoma.[59]

MBNL1 which is one of RNA-binding proteins (RBPs) has previously been linked to several cancers and in the context of cholangiocarcinoma it has been found that MBNL1 was upregulated with differential expression levels in tumoral samples in consistent with the results of our study.[60]

The AScancer atlas was utilized to gather additional information about 10 alternative splicing (AS) events and determine the splicing model in protein-coding transcripts. By comparing the Percent Spliced In (PSI) and expression values for each significant splicing event, statistically

significant differences were observed between tumoral and normal samples in terms of PSI, expression values, or both.

Because variation in AS patterns vary from patient to patient which may make personalized genetic treatment the best choice as synthetic regulation of alternative splicing provides critical molecular tools in biomedicine. While the use of CRISPR-Cas9 for targeting oncogenic AS events is not yet well studied, the specificity of these events to different tumor types and the customizable nature of the CRISPR-Cas9 system suggest that splice variants can be effectively targeted with CRISPR-Cas9. [61, 62]

At the end by using CRISPOR web tool we were able to design gRNA for the AS event on EIF4A2 gene with the suitable PAM (Protospacer Adjacent Motif) and restriction enzymes in order to make targeted genome editing.

10. Conclusion

In conclusion, it was observed that certain genes involved in alternative splicing exhibited abnormal patterns in cholangiocarcinoma samples compared to normal tissues. These aberrant splicing events were found to be associated with specific biological pathways implicated in cholangiocarcinoma development and progression.

Furthermore, the identification and characterization of alternative splicing events contributed to a better understanding of the molecular mechanisms underlying cholangiocarcinoma. Overall, the integrated bioinformatics analysis in this research provided insights into the role of aberrant alternative splicing in the emergence and progression of cholangiocarcinoma by influencing various cellular processes and gene expression patterns., highlighting its potential impact on disease pathogenesis and offering potential avenues for targeted therapies.

However, the study had some limitations. First, it was a pure bioinformatics study and did not include clinical experiments to prove the scientific hypothesis. Second, the patients enrolled in the study were exclusively from a single database, which may limit the generalizability of the findings. Third, the study focused mainly on exon skipping as the most represented splicing type among the alternative splicing events in the cholangiocarcinoma dataset. Fourth, as we mentioned before the use of CRISPR-Cas9 for targeting alternative splicing (AS) events is still an area of active research and is not well studied. CRISPR technology, in general, has many caveats and risks associated with it, as genetic modification can have unintended consequences. Therefore, it is crucial to conduct in-vivo and ex-vivo studies to validate the clinical consequences of genetic modifications before considering their use in a clinical setting.

11. References

1. Golino, J. L., Wang, X., Maeng, H. M., & Xie, C. (2023). **Revealing the Heterogeneity of the Tumor Ecosystem of Cholangiocarcinoma through Single-Cell Transcriptomics**. Cells, 12(6), 862.

2. Van Dyke, A. L., Shiels, M. S., Jones, G. S., Pfeiffer, R. M., Petrick, J. L., Beebe-Dimmer, J. L., & Koshiol, J. (2019). **Biliary tract cancer incidence and trends in the United States by demographic group**, 1999-2013. Cancer, 125(9), 1489–1498. https://doi.org/10.1002/cncr.31942

3. Lin, Z., Gong, J., Zhong, G., Hu, J., Cai, D., Zhao, L., & Zhao, Z. (2021). **Identification of Mutator-Derived Alternative Splicing Signatures of Genomic Instability for Improving the Clinical Outcome of Cholangiocarcinoma**. Frontiers in Oncology, 11. https://www.frontiersin.org/articles/10.3389/fonc.2021.666847

4. El Marabti, E., & Younis, I. (2018). **The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer. Frontiers in Molecular Biosciences**, 5, 80. https://doi.org/10.3389/fmolb.2018.00080

5. Yang, Q., Zhao, J., Zhang, W., Chen, D., & Wang, Y. (2019). **Aberrant alternative splicing in breast cancer. Journal of Molecular Cell Biology**, 11(10), 920–929. https://doi.org/10.1093/jmcb/mjz033

6. Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Gervais-Bird, J., Madden, R., Paquet, E. R., Koh, C., Venables, J. P., Prinos, P., Jilaveanu-Pelmus, M., Wellinger, R., Rancourt, C., Chabot, B., & Abou Elela, S. (2008). **Multiple alternative splicing markers for ovarian cancer.** Cancer Research, 68(3), 657–663. https://doi.org/10.1158/0008-5472.CAN-07-2580

7. Cherry, S., & Lynch, K. W. (2020). **Alternative splicing and cancer: Insights, opportunities, and challenges from an expanding view of the transcriptome**. Genes & Development, 34(15–16), 1005–1016. https://doi.org/10.1101/gad.338962.120

8. Zhang, Y., Qian, J., Gu, C., & Yang, Y. (2021). **Alternative splicing and cancer: A systematic review. Signal Transduction and Targeted Therapy**, 6(1), 78. https://doi.org/10.1038/s41392-021-00486-7

9. Coomer, A. O., Black, F., Greystoke, A., Munkley, J., & Elliott, D. J. (2019). **Alternative splicing in lung cancer**. Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms, 1862(11–12), 194388. https://doi.org/10.1016/j.bbagrm.2019.05.006

10. Pio, R., & Montuenga, L. M. (2009). **Alternative Splicing in Lung Cancer**. Journal of Thoracic Oncology, 4(6), 674–678. https://doi.org/10.1097/JTO.0b013e3181a520dc

11. Schafer, S., Miao, K., Benson, C. C., Heinig, M., Cook, S. A., & Hubner, N. (2015). **Alternative Splicing Signatures in RNA-seq Data: Percent Spliced in (PSI)**. Current Protocols in Human Genetics, 87, 11.16.1-11.16.14. https://doi.org/10.1002/0471142905.hg1116s87

12. Zhou, J., Ma, S., Wang, D., Zeng, J., & Jiang, T. (2018). **FreePSI: An alignment-free approach to estimating exon-inclusion ratios without a reference transcriptome**. Nucleic Acids Research, 46(2), e11. https://doi.org/10.1093/nar/gkx1059

13. Yosudjai, J., Wongkham, S., Jirawatnotai, S., & Kaewkong, W. (2019). **Aberrant mRNA splicing generates oncogenic RNA isoforms and contributes to the development and progression of cholangiocarcinoma.** Biomedical Reports, 10(3), 147–155. https://doi.org/10.3892/br.2019.1188

14. Marin, J. J. G., Reviejo, M., Soto, M., Lozano, E., Asensio, M., Ortiz-Rivero, S., Berasain, C., Avila, M. A., & Herraez, E. (2021). **Impact of Alternative Splicing Variants on Liver Cancer Biology. Cancers**, 14(1), 18. https://doi.org/10.3390/cancers14010018

15. Herraez, E., Lozano, E., Macias, R. I. R., Vaquero, J., Bujanda, L., Banales, J. M., Marin, J. J. G., & Briz, O. (2013). **Expression of SLC22A1 variants may affect the response of hepatocellular carcinoma and cholangiocarcinoma to sorafenib. Hepatology** (Baltimore, Md.), 58(3), 1065–1073. https://doi.org/10.1002/hep.26425

16. Wu, H.-Y., Wei, Y., Liu, L.-M., Chen, Z.-B., Hu, Q.-P., & Pan, S.-L. (2019). **Construction of a model to predict the prognosis of patients with cholangiocarcinoma using alternative splicing events**. Oncology Letters, 18(5), 4677–4690. https://doi.org/10.3892/ol.2019.10838

17. Kahles, A., Ong, C. S., Zhong, Y., & Rätsch, G. (2016). **SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data**. Bioinformatics (Oxford, England), 32(12), 1840–1847. https://doi.org/10.1093/bioinformatics/btw076

18. R Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., & Weinstein, J. N. (2012). **SpliceSeq: A resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts**. Bioinformatics, 28(18), 2385–2387. https://doi.org/10.1093/bioinformatics/bts452

19. Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018**). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biology**, 19(1), 40. https://doi.org/10.1186/s13059-018-1417-1.

20 .Saraiva-Agostinho, N., & Barbosa-Morais, N. L. (2019). **psichomics: Graphical application for alternative splicing quantification and analysis**. Nucleic Acids Research, 47(2), e7. https://doi.org/10.1093/nar/gky888

21. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). **The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology** (Poznan, Poland), 19(1A), A68-77. https://doi.org/10.5114/wo.2014.47136

22. GTEx Consortium. (2013). **The Genotype-Tissue Expression (GTEx) project**. Nature Genetics, 45(6), 580–585. https://doi.org/10.1038/ng.2653

23. Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). **The sequence read archive**. Nucleic Acids Research, 39(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

24. Saraiva-Agostinho, N., Barbosa-Morais, N. L., Falcão, A., Paez, L. G., Bordone, M., Maia, T., Ferreira, M., Leote, A. C., & Almeida, B. de. (2023). **Psichomics: Graphical Interface for Alternative Splicing Quantification, Analysis and Visualisation** (1.26.0). Bioconductor version: Release (3.17). https://doi.org/10.18129/B9.bioc.psichomics

25. Ge, S. X., Jung, D., & Yao, R. (2020a). **ShinyGO: A graphical gene-set enrichment tool for animals and plants**. Bioinformatics, 36(8), 2628–2629. https://doi.org/10.1093/bioinformatics/btz931

26. Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., Jensen, L. J., & von Mering, C. (2021). **The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets**. Nucleic Acids Research, 49(D1), D605–D612. https://doi.org/10.1093/nar/gkaa1074

27. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. Genome Research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

28. Wu, S., Huang, Y., Zhang, M., Gong, Z., Wang, G., Zheng, X., Zong, W., Zhao, W., Xing, P., Li, R., Liu, Z., & Bao, Y. (2023). **ASCancer Atlas: A comprehensive knowledgebase of alternative splicing in human cancers**. Nucleic Acids Research, 51(D1), D1196–D1204. https://doi.org/10.1093/nar/gkac955

29. Concordet, J.-P., & Haeussler, M. (2018). **CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens**. Nucleic Acids Research, 46(Web Server issue), W242–W245. https://doi.org/10.1093/nar/gky354

30. Abbas-Aghababazadeh, F., Li, Q., & Fridley, B. L. (2018). **Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing**. PLoS ONE, 13(10), e0206312. https://doi.org/10.1371/journal.pone.0206312

31. Abril, J. F., & Castellano, S. (2019). **Genome Annotation.** In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology (pp. 195–209). Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20226-4

32. Ramachandran,R., Ravichandran,G. & Raveendran,A.(2020). **Evaluation of Dimensionality Reduction Techniques for Big data**. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). pp. 226-231, doi: 10.1109/ICCMC48092.2020.ICCMC-00043.

33. Tomczak, A., Mortensen, J. M., Winnenburg, R., Liu, C., Alessi, D. T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N. H., Musen, M. A., & Khatri, P. (2018). **Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations.** Scientific reports, 8(1), 5115. https://doi.org/10.1038/s41598-018-23395-2

34. **KEGG Mapping**. (n.d.). Retrieved June 22, 2023, from https://www.genome.jp/kegg/kegg1b.html

35. **ShinyGO 0.76**. (n.d.). Retrieved June 24, 2023, from http://bioinformatics.sdstate.edu/go76/

36. **STRING: functional protein association networks**. (n.d.-a). Retrieved June 22, 2023, from https://string-db.org/

37. **What are genome editing and CRISPR-Cas9?** MedlinePlus Genetics. (n.d.). Retrieved June 22, 2023, from https://medlineplus.gov/genetics/understanding/genomicresearch/genomeediting/

38. **CRISPR Cas 9 Nuclease RNA-guided Genome Editing**. (n.d.). Retrieved June 22, 2023, from https://www.sigmaaldrich.com/IE/en/technical-documents/protocol/genomics/advanced-gene-editing/crispr-cas9-genome-editing

39. **CRISPOR.** (n.d.). Retrieved June 22, 2023, from http://crispor.org/

40. **Addgene: CRISPR Guide**. (n.d.). Retrieved June 22, 2023, from https://www.addgene.org/guides/crispr/

41. Song M. (2017). **The CRISPR/Cas9 system: Their delivery, in vivo and ex vivo applications and clinical development by startups**. Biotechnology progress, 33(4), 1035–1045. https://doi.org/10.1002/btpr.2484

42. **Bile Duct Cancer (Cholangiocarcinoma): Statistics** | Cancer.Net. (n.d.). Retrieved June 22, 2023, from https://www.cancer.net/cancer-types/bile-duct-cancer-cholangiocarcinoma/statistics

43. Turati F, Bertuccio P, Negri E &Vecchia CL. (2022) **Epidemiology of cholangiocarcinoma.** Hepatoma Res 2022; 8:19. http://dx.doi.org/10.20517/2394-5079.2021.130

44. Mosconi, S., Beretta, G. D., Labianca, R., Zampino, M. G., Gatta, G., & Heinemann, V. (2009). **Cholangiocarcinoma. Critical Reviews in Oncology/Hematology**, 69(3), 259–270. https://doi.org/10.1016/j.critrevonc.2008.09.008

45. **Cholangiocarcinoma Foundation—Helping those with bile duct cancer**. (n.d.). Retrieved June 22, 2023, from https://cholangiocarcinoma.org/

46. Xue, C., Gu, X., Li, G., Bao, Z., & Li, L. (2021). **Expression and Functional Roles of Eukaryotic Initiation Factor 4A Family Proteins in Human Cancers**. Frontiers in Cell and Developmental Biology, 9, 711965. https://doi.org/10.3389/fcell.2021.711965

47. Li, J., Cheng, D., Zhu, M., Yu, H., Pan, Z., Liu, L., Geng, Q., Pan, H., Yan, M., & Yao, M. (2019). **OTUB2 stabilizes U2AF2 to promote the Warburg effect and tumorigenesis via the AKT/mTOR signaling pathway in non-small cell lung cancer**. Theranostics, 9(1), 179–195. https://doi.org/10.7150/thno.29545

48. Mu, Q., Lv, Y., Luo, C., Liu, X., Huang, C., Xiu, Y., & Tang, L. (2021**). Research Progress on the Functions and Mechanism of circRNA in Cisplatin Resistance in Tumors**. Frontiers in Pharmacology, 12, 709324. https://doi.org/10.3389/fphar.2021.709324

49. Chen, B.-L., Wang, H.-M., Lin, X.-S., & Zeng, Y.-M. (2021). **UPF1: A potential biomarker in human cancers.** Frontiers in Bioscience (Landmark Edition), 26(5), 76–84. https://doi.org/10.52586/4925

50. Zhou, Y., Li, Y., Wang, N., Li, X., Zheng, J., & Ge, L. (2019). **UPF1 inhibits the hepatocellular carcinoma progression by targeting long non-coding RNA UCA1**. Scientific Reports, 9(1), 6652. https://doi.org/10.1038/s41598-019-43148-z

51. Bordonaro, M., & Lazarova, D. (2019). **Amlexanox and UPF1 Modulate Wnt Signaling and Apoptosis in HCT-116 Colorectal Cancer Cells**. Journal of Cancer, 10(2), 287–292. https://doi.org/10.7150/jca.28331

52. Li, L., Geng, Y., Feng, R., Zhu, Q., Miao, B., Cao, J., & Fei, S. (2017). **The Human RNA Surveillance Factor UPF1 Modulates Gastric Cancer Progression by Targeting Long Non-Coding RNA MALAT1**. Cellular Physiology and Biochemistry:

International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology, 42(6), 2194–2206. https://doi.org/10.1159/000479994

53. Liu, C., Karam, R., Zhou, Y., Su, F., Ji, Y., Li, G., Xu, G., Lu, L., Wang, C., Song, M., Zhu, J., Wang, Y., Zhao, Y., Foo, W. C., Zuo, M., Valasek, M. A., Javle, M., Wilkinson, M. F., & Lu, Y. (2014). **The UPF1 RNA surveillance gene is commonly mutated in pancreatic adenosquamous carcinoma**. Nature Medicine, 20(6), 596–598. https://doi.org/10.1038/nm.3548

54. Zhong, Z.-B., Wu, Y.-J., Luo, J.-N., Hu, X.-N., Yuan, Z.-N., Li, G., Wang, Y.-W., Yao, G.-D., & Ge, X.-F. (2020). **Knockdown of long noncoding RNA DLX6-AS1 inhibits migration and invasion of thyroid cancer cells by upregulating UPF1**. European Review for Medical and Pharmacological Sciences, 24(16), 8246. https://doi.org/10.26355/eurrev_202008_22587

55. Pei, C.-L., Fei, K.-L., Yuan, X.-Y., & Gong, X.-J. (2019). **LncRNA DANCR aggravates the progression of ovarian cancer by downregulating UPF1**. European Review for Medical and Pharmacological Sciences, 23(24), 10657–10663. https://doi.org/10.26355/eurrev_201912_19763

56. Yang, C., Ströbel, P., Marx, A., & Hofmann, I. (2013). **Plakophilin-associated RNA-binding proteins in prostate cancer and their implications in tumor progression and metastasis**. Virchows Archiv: An International Journal of Pathology, 463(3), 379–390. https://doi.org/10.1007/s00428-013-1452-y

57. Yang, C., Ströbel, P., Marx, A., & Hofmann, I. (2013). **Plakophilin-associated RNA-binding proteins in prostate cancer and their implications in tumor progression and metastasis**. Virchows Archiv: An International Journal of Pathology, 463(3), 379–390. https://doi.org/10.1007/s00428-013-1452-y

58. Shkreta, L., Delannoy, A., Salvetti, A., & Chabot, B. (2021). **SRSF10: an atypical splicing regulator with critical roles in stress response, organ development, and viral replication**. RNA (New York, N.Y.), 27(11), 1302–1317. https://doi.org/10.1261/rna.078879.121

59. Sun, Q., Li, F., Yu, S., Zhang, X., Shi, F., & She, J. (2018). **Pontin Acts as a Potential Biomarker for Poor Clinical Outcome and Promotes Tumor Invasion in**

Hilar Cholangiocarcinoma. BioMed research international, 2018, 6135016. https://doi.org/10.1155/2018/6135016

60. Ian, A., Pu, K., Li, B., Li, M., Liu, X., Gao, L., & Mao, X. (2019). **Weighted gene coexpression network analysis reveals hub genes involved in cholangiocarcinoma progression and prognosis.** Hepatology research: the official journal of the Japan Society of Hepatology, 49(10), 1195–1206. https://doi.org/10.1111/hepr.13386

61. Gapinske, M., Luu, A., Winter, J., Woods, W. S., Kostan, K. A., Shiva, N., Song, J. S., & Perez-Pinera, P. (2018). **CRISPR-SKIP: Programmable gene splicing with single base editors**. Genome Biology, 19(1), 107. https://doi.org/10.1186/s13059-018-1482-5

62. Dm, M., Ao, M.-P., Jij, O., & Dsb, H. (2018). **Alternative splicing and cancer metastasis: Prognostic and therapeutic applications. Clinical & Experimental Metastasis**, 35(5–6). https://doi.org/10.1007/s10585-018-9905-y