Alterations of RNA Splicing Patterns in Esophagus Squamous Cell Carcinoma

Jiyu Ding
Shantou University Medical College
Chunquan Li
Harbin Medical University
Yinwei Cheng
Shantou University Medical College
Zepeng Du
Shantou Central Hospital
Qiuyu Wang
Harbin Medical University
Zhidong Tang
Harbin Medical University
Chao Song
Harbin Medical University
Qiaoxi Xia
Shantou University Medical College
Wenjing Bai
Shantou University Medical College
Ling Lin
Shantou University Medical College
Wei Liu
Shantou University Medical College
Liyan Xu
Shantou University Medical College
Enmin Li
Shantou University Medical College
Bingli Wu (blwu@stu.edu.cn)
Shantou University https://orcid.org/0000-0002-0614-4721

Research

Keywords: alternative splicing, MISO, esophagus squamous cell carcinoma

Posted Date: July 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-39232/v1

Version of Record: A version of this preprint was published on February 9th, 2021. See the published version at https://doi.org/10.1186/s13578-021-00546-z.

Abstract

Alternative splicing (AS) is an important biological process for regulating the expression of various isoforms from a single gene and thus to promote proteome diversity. In this study, RNA-seq data from 15 pairs of esophageal squamous cell carcinoma (ESCC) tissue samples and two cell lines were analyzed. AS events with significant differences between ESCC and matched normal tissues were re-annotated to find protein coding genes or non-coding RNAs. A total of 45,439 AS events was found. Of these, 6,019 (13.25%) significant differentially AS events were identified. Exon skipping (SE) events occupied the largest proportion of abnormal splicing events. Fifteen differential splicing events with the same trends of ΔΨ values in ESCC tissues and cell lines were found. Four pathways and 20 biological processes related to cell junction and migration were significantly enriched. The upregulated splicing factor SF3B4, which regulates 92 gene splicing events, could be a potential prognostic factor of ESCC. Sashimi plotting was applied to show the differentially-spliced genes, including HNRNPC, VCL, ZNF207, KIAA1217, TPM1 and CALD1. These results suggest that cell junction- and migration-related biological processes and KEGG pathways are affected by the AS abnormalities, and aberrant splicing events can be affected by changed expression of splicing factors. In summary, we identified significant differentially AS events which might be related to the development of ESCC.

Introduction

Esophageal cancer is the sixth leading cause of cancer fatalities in the world [1], with a five-year survival of less than 20% [2]. Esophageal cancer is mainly comprised two types that differ in epidemiology and pathology: esophageal squamous cell carcinoma and esophageal adenocarcinoma. About 90% of esophageal cancer cases worldwide are esophageal squamous cell carcinoma [1]. Studies committed to the molecular mechanism of esophageal squamous cell cancer development need to be carried out further to help improve diagnosis, treatment and prognosis of ESCC.

Pre-mRNA splicing is a biological process that releases intron loops and puts exons together to produce mature mRNAs [3], and is carried out by a spliceosome consisting of five snRNP complexes and more than 200 auxiliary proteins [4]. In the process of alternative splicing, alternative splice site selection within pre-mRNAs results in the production of various mature mRNAs from a single gene [5].

It has been reported normal organ development and cell differentiation are regulated by alternative splicing [6]. However, splicing patterns abnormality caused by mutations of cis-acting sequences and spliceosomal proteins can also trigger human diseases, including cancer [7, 8]. Altered splicing patterns influence multiple aspects of cancer development, including cell proliferation regulation, programmed cell death, cancer cell metabolism, angiogenesis, and tumor metastasis [5, 9]. Genome-wide methods indicate that tumor development often involves large-scale variations in alternative splicing [10].

The MISO model was originally developed to assess the expression of alternatively spliced exons and transcripts, and to estimate confidence [11]. MISO uses confidence intervals to assess exon and transcript abundance, but also detects differentially-spliced exons or transcripts. This is accomplished by calculating the posterior distribution of unobserved random variables for the RNA-Seq data [11]. In this study, MISO was utilized to characterize the splicing patterns and try to identify the regulatory roles of splicing factors on splicing abnormalities in ESCC.

Materials And Methods

Source of sequencing data and routine analysis

RNA-seq data of 15 pairs of ESCC tissues and the SHEE and SHEEC cell lines came from our laboratory, and all data has been published [12]. FastQC was used to measure the quality of the sequences [13]. The paired-end sequences were then aligned against the UCSC (http://genome.ucsc.edu/) human genome (hg19) [14] using TopHat2 v2.0.13 [15] and Bowtie2 [16] using default parameters. Alignments were further analyzed with Cufflinks [17] to calculate the abundance (FPKM) of each gene and isoform using the hg19 gene annotations, and with DESeq [18] to identify differentially-expressed splicing factors. To validate the differentially-expressed splicing factors, two microarray datasets, named GSE53624 [18–19] and GSE23400-GPL96 [20–21], were utilized to perform differential expression analysis with the siggenes package in R2.15.3.

Identification of AS events in ESCC

The splicing patterns of ESCC were analyzed with MISO [11] using exon-centric annotations V.2.0, and the results were filtered with a cut-off of $\Delta\Psi > 0.2$, Bayes-factor > 5, and a sum of inclusion and exclusion reads number > 10. Sashimi plots were used to visualize AS events with significant differences in ≧ 3 pairs of ESCC tissues, and between the SHEE and SHEEC cell lines.

Re-annotation of Differential AS Events

AS events used in the previous analysis were annotated with human genome (hg19) alternative events v2.0, and their annotations were derived from Ensembl genes, knownGenes (UCSC) and RefSeq genes. In order to further verify the AS event annotations and find out their represented genes and transcripts, AS events with significant differences in at least 3 pairs of ESCC tissue samples were re-annotated. Homo_sapiens.GRCh37.75.gtf from the Ensembl database [22–23] and GRCh37_latest_genomic.gff from the RefSeq database [24] were chosen for re-annotation. Since the mRNA sequences and gene-related annotations of the UCSC database were derived from GenBank and RefSeq databases, those annotations were not adopted for re-annotation of AS events with several criterion described as follows: (i) if three exons in an SE event appear in the same transcript, the SE event was judged to be corrected, (ii) if two exons in an RI event are present in the same transcript and there are no other exons between these two exons, then the RI event annotation was determined to be corrected, (iii) if exons 124 and 134 occurred in the same transcript in an MXE event, and exons 1–4 did not appear in any one transcript, the MXE event was judged as correct, (iv) if the variable exon (the longer or the shorter exon) and constitutive exons in an A3SS event occurred in the same transcript, and between these two exons there were no other exons, the A3SS event annotation was determined to be corrected, (v) if the variable exon and the constitutive exon in an A5SS appeared in the same transcript, and between the two exons there were no other exons, the A5SS event annotation was judged to be correct. Since there are no chromosome numbers in the RefSeq exon annotations, the chromosomal locations of exons using RefSeq annotations should be checked later.

Functional Enrichment Analysis

DAVID online tools [25–26] were utilized to perform functional enrichment analysis towards GO biological process and KEGG pathways. The significant cut-off was set to $P < 0.05$. Dot plots from R package named ggplot2 [27] were then applied to visualize the significantly enriched GO terms and KEGG pathways.

Differential AS events from ESCC tissues and SHEE/SHEEC cell lines

After the results of differential AS events from the SHEE and SHEEC cell lines were filtered, the following information, including AS event names, Ψ values from the SHEE cell line, Ψ values from the SHEEC cell line and ΔΨ values from SHEE/SHEEC cells were selected. The re-annotation results of the AS events in ESCC tissues using gene annotations from the Ensembl and RefSeq databases were overlapped with SHEE/SHEEC AS events, then whether the trend of ΔΨ values of each AS event in tissues and in cell lines was similar would be judged.

Splicing regulatory network establishment

Gene expression of each sample was combined and merged with the splicing factor list mainly from the SpliceAid2 database [28]. Spearman correlation tests were performed between splicing factor expression and the splicing event psi values, with a threshold value set at $P < 0.05$. The significant pairs of splicing factors and genes of the related splicing events were chosen to establish a splicing regulatory network based on the results of Spearman tests. Cytoscape software was used to visualize the splicing regulatory network [29].

Survival analysis

The expression and clinical data of 87 ESCC patients were obtained from the Xena browser database (https://xenabrowser.net/datapages/). The log-rank (Mantel-Cox) test was used to perform survival analysis and a K-M curve was plotted using GraphPad software.

Results

Overall mapping rate after genome alignments

Reads from each clinical case were mapped to human genome sequences. The results are shown in Table 1, where the total alignment number of the sequenced fragments ranged from 15 to 39 million, and the comparison rate ranged between 83%-92%.

Table 1
Summary of TopHat2 genomic alignments

Number of Reads	Total reads		Mapped reads		Overall reads	
	Left	Right	Left	Right	Mapping_rate	Aligned_pairs
B782N	51906891	51906891	45391832	42285862	84.50%	39419636
B782T	20187921	20187921	18022068	16864424	86.40%	15851609
B783N	22772025	22772025	20876811	20292461	90.40%	19220668
B783T	22026613	22026613	20199656	19665894	90.50%	18669500
B785N	23881681	23881681	21879991	20946206	89.70%	19806614
B785T	27545958	27545958	25310820	24617731	90.60%	23314980
B786N	19526756	19526756	18043943	17643909	91.40%	16752687
B786T	20424033	20424033	18841925	18319670	91.00%	17453969
B788N	22166170	22166170	20319140	19346438	89.50%	18276699
B788T	19329401	19329401	17949865	17468161	91.60%	16651551
B791N	25487339	25487339	22803788	22091751	88.10%	20629163
B791T	19730469	19730469	17587161	16688030	86.90%	15713877
B794N	21773037	21773037	20089962	18134197	87.80%	17197455
B794T	27630784	27630784	24796070	22364260	85.30%	20946236
B797N	24393756	24393756	22517747	21882580	91.00%	20713525
B797T	20082035	20082035	18745753	18269290	92.20%	17451339
B798N	32153939	32153939	29786198	26930352	88.20%	25572896
B798T	32468547	32468547	30246172	27414170	88.80%	26154479
B799N	25194910	25194910	23448172	21151887	88.50%	20095525
B799T	20897751	20897751	19481307	17526890	88.50%	16701290
B800N	26328835	26328835	24474467	22048959	88.40%	20927261
B800T	24186510	24186510	22481444	20362666	88.60%	19317914
B801N	22799054	22799054	21390618	19363520	89.40%	18518368
B801T	26946581	26946581	25259555	22713737	89.00%	21717932
B804N	27121722	27121722	25391238	22801422	88.80%	21774130
B804T	24985226	24985226	23383899	21149641	89.10%	20212413
C199N	29776640	29776640	26261213	25254804	86.50%	23366546
C199T	35651673	35651673	30502787	29013594	83.50%	26708316
C200N	21494047	21494047	19784129	19242503	90.80%	18270154

Number of Reads	Total reads		Mapped reads		Overall reads	
	Left	Right	Left	Right	Mapping_rate	Aligned_pairs
C200T	19588861	19588861	17953109	17510470	90.50%	16566068
B782N	51906891	51906891	45391832	42285862	84.50%	39419636
B782T	20187921	20187921	18022068	16864424	86.40%	15851609
SHEE	55085029	55085029	47868214	44812830	84.1%	41691361
SHEEC	54582526	54582526	47769458	44796267	84.8%	42325127

Changes in AS patterns between ESCC tissue, matched normal tissue and cell lines

From the results of splicing analysis, a total of 45,439 AS events were detected using MISO. Of these, 6019 AS events were differently spliced with significance in paired ESCC tissues and matched normal tissues, with a cut-off of BF > 5, |ΔΨ|> 0.2, and number of inclusion and exclusion reads > 10, accounting for 13.25% of the total number of AS events (Fig. 1A). After filtering out the AS events with inconsistent ΔΨ trends in the paired samples, there were 5150 differential splicing events, including 2324 exon skipping (SE) events, 1014 intron retention (RI) events, and 693 mutually exclusive exon (MXE) events, 592 alternative 3’ splice site (A3SS) events, and 527 alternative 5’ splice site (A5SS) events. The AS results of ESCC tissue samples showed that the proportion of SE events that could be detected was the largest, being about 33.6%, RI events comprised 6.9%, MXE events comprised 6.8%, A3SS comprised 8.5%, and A5SS events comprised 6.4%. The percentage of SE events in differential splicing events was still the largest, comprising 4.8%, RI events comprising 2.5%, MXE events comprising 1.5%, A3SS events comprising 1.2%, and A5SS events comprising approximately 1.1%. Individual differences were detected during MISO analysis. These individualized differences were more obvious in differential splicing events (Fig. 1B). From the volcano plot, it could be seen that the ΔΨ values and log2(BF) values in detectable AS events and the differentially variably spliced events were symmetrically distributed (Fig. 1C). Also, it could be seen from the bar plot that the ΔΨ trend did not have a significant difference in each type of AS event (Fig. 1D).

A total of 32,891 AS events were detected, of which a total of 928 AS events were differently spliced with significance in the SHEE/SHEEC cell lines, accounting for approximately 2.8% of the total detectable AS events (Fig. 1E). The threshold for judging the differences in AS events in the SHEE/SHEEC cell lines was the same as used for ESCC AS event filtration. In detectable AS events, the proportion of SE was the largest, being about 55.9%, RI comprising 9.2%, MXE comprising 10.9%, A3SS accounting for 13.6%, and A5SS accounting for 10.3%. In differential splicing events, the number of SE, RI, MXE, A3SS, and A5SS events were 455, 131, 152, 96, and 94, accounting for 49%, 14.1%, 16.4%, 10.3%, and 10.1% of the total number of differential splicing events separately (Fig. 1F).

Re-annotation of AS events in paired ESCC tissues

To identify the genes and transcripts representing the differential AS events and filter out incorrect AS events, as described in the methods, we re-annotated all differentially splicing events found in at least 3 paired samples. For SE events, there were a total of 222 and 151 significant alternatively spliced events before and after re-annotation, reducing the number of AS events to approximately 68% of all events. For RI events, there were a total of 134 significant AS events before re-annotation, 115 after re-annotation, and a reduction of 14.2% of the total number of

all events. For MXE events, there were a total of 62 significant alternative splicing events before re-annotation and 6 after re-annotation, reducing the number of AS events to 9.7% of the total. For the A3SS events, there were 57 significant AS events before re-annotation and 42 after re-annotation, reducing the number of AS events to approximately 73.7% of all events. For the A5SS events, there were a total of 52 significant AS events before re-annotation and 35 after re-annotation, reducing the number of AS events to approximately 67.3% of all events (Fig. 2A). The MXE events were reduced the most because the two exons that should have been mutually exclusive in the MISO official MXE event annotation were observed to exist in the same transcript. A total of 349 AS events were identified after re-annotation in 15 pairs of ESCC tissue samples, of which transcripts and genes representing 169 AS events could be found in the Ensembl and RefSeq database gene annotations, whereas 114 could only be identified having representing transcripts and genes in the RefSeq database gene annotations, and 66 AS events could only be identified having representing transcripts and genes in the Ensembl database gene annotations (Fig. 2B). We have also calculated the fractions of AS events representing mRNA and ncRNA (Fig. 2C). The top 10 splicing events with the most paired sample counts are shown in the following table (Table 2). Splicing modes of CALD1 (Fig. 3B), KIAA1217 (Fig. 3C), TPM1 (Fig. 3D), HNRNPC (Fig. 3A), VCL (Fig. 3E), and ZNF207 (Fig. 3F) were visualized with Sashimi plotting.

Table 2
Summary of alternative splicing events with top 10 sample count (paired) in ESCC tissues

Event	Gene_Name	Count (paired)	Sample	Type
chr1:150483352:150483674:+ @chr1:150483933:150484307: +@chr1:150484828:150485048:+	ECM1	14	B783, B798, C199, B788, B782, B799, B797, B791, B800, B785, B786, B804, C200, B801	SE
chr10:24833910–24834032:+ @chr10:24834756–24836772:+	KIAA1217	13	C199, B783, B785, B782, C200, B786, B799, B788, B797, B794, B804, B791, B800	RI
chr9:117835882:117836145:- @chr9:117808689:117808961:- @chr9:117804498:117804620:-	TNC	12	B804, B785, B783, B794, B786, B798, B800, B797, C199, B782, C200, B788	SE
chr15:63334838:63335142:+@chr15:63335905:63336030:+ @chr15:63336226:63336351:+@chr15:63349184:63349317:+	TPM1	11	C200, B788, B785, B800, B794, B801, C199, B782, B797, B804, B783	MXE

Event	Gene_Name	Count (paired)	Sample	Type
chr2:238289558:238290142:-@chr2:238287279:238287878:-@chr2:238285415:238285987:-	COL6A3	11	B797, B783, B786, C199, B782, B800, B788, C200, B785, B798, B794	SE
chr22:31495731:31495882:+@chr22:31496871:31497035:+@chr22:31500302:31500610:+	SMTN	10	B785, B797, B782, C199, B788, C200, B804, B800, B794, B801	SE
chr16:15808766:15808938:-@chr16:15802660:15802698:-@chr16:15796992:15797980:-	MYH11	9	C200, B799, B794, B788, B800, B801, B785, C199,B782	SE
chr17:30692349:30692506:+@chr17:30693684:30693776:+@chr17:30694791:30695033:+	ZNF207	9	B782, C200, B788, B794, B786, B785, C199, B797, B800	SE
chr7:134617739:134618141|134618828:+@chr7:134620439:134620516:+	CALD1	9	B788, C200, B800, B785, B804, B782, C199, B797, B786	A5SS
chr9:117835882:117836145:-@chr9:117819432:117819704:-@chr9:117808689:117808961:-	TNC	9	B798, B782, B800, B794, C199, C200, B788, B786, B785	SE

Overlapping AS events in ESCC tissues and SHEE/SHEEC cell lines

A total of 37 AS events were identified to be differentially spliced in both ESCC tissues and cell lines, based on statistical significance, 15 of which had the same trend in ΔΨ values in ESCC tissues and cell lines. Among these 15 AS events, there were 4 SE events, 7 RI events, 2 MXE events, and 2 A5SS events (Supplementary Table 1).

Classification of AS events according to representative transcript types

Of the 349 AS events after re-annotation, 312 events represented at least one protein-coding transcript, which accounted for 89.4% of the total AS events. The remaining 37 AS events did not represent any protein-coding transcripts, including some processed protein coding transcripts, lncRNAs, and microRNAs (Supplementary Table 2). Among the 235 AS events re-annotated with Ensembl transcript annotations, 200 events represented at least one protein-coding transcript, which accounted for 85.1% of the total AS events, and the remaining 35 AS events represented non-protein-coding transcripts. Of the 283 variable splicing events re-annotated with RefSeq transcript annotations, 266 events represented at least one protein-coding transcript, which accounted for 94% of the total AS events, and the remaining 17 AS events represented non-protein-coding transcripts.

Functional enrichment revealed the potential role of AS transcripts

To identify the biological processes and pathways affected by splicing abnormalities in ESCC, we performed enrichment analysis. A total of 270 protein-coding genes were selected for functional and pathway enrichment analysis from the results of alternative splicing events after re-annotation of differential AS events with gene annotations from the Ensembl and RefSeq databases. A total of 234 genes were enriched in 543 biological processes. There were 382 biological processes satisfying the threshold of $P < 0.05$, and 39 genes were enriched in 12 pathways, of which 10 pathways satisfied $P < 0.05$. Furthermore, there were 4 KEGG pathways and 20 biological processes that related to invasion and metastasis (Fig. 4).

Splicing regulatory networks in ESCC

To find the regulators of the splicing alterations in ESCC, we constructed a splicing regulatory network with FPKM values of splicing factors and PSI values of AS events. Filtered with a cut-off of $P < 0.05$ after Spearman correlation analysis, 5,999 splice factor-splicing event pairs and 5,511 splice factor-target gene pairs were chosen to build a splicing regulatory network. This network involved 81 splicing factors and 223 differential AS events (Fig. 5). Splice factors that were differentially expressed in ESCC and matched normal tissues were measured by two public microarray datasets and the RNA-seq dataset generated from our laboratory. It turned out that SF3B4 was significantly differentially expressed in ESCC with a fold-change of about 1.5 (Table 3).

Table 3
Summary of SF3B4 expression using three ESCC datasets

Dataset	Num of Samples	FoldChange	Pvalue	Qvalue	R packages
RNA-seq	15 pairs	1.422555267	0.196605694	0.458244843	DESeq
GSE53624	119 pairs	1.642700989	0	0	siggenes
GSE23400	53 pairs	1.49618277	0	0	siggenes

The splicing regulatory network with SF3B4 and its target genes were also provided and contained 92 target genes and 102 abnormal AS events (Fig. 6A). SF3B4 was found to be a survival-related gene in ESCC. ESCC patients with a lower expression level of SF3B4 had a longer survival time (Fig. 6B). The PSI value of SRSF5's RI event was found to have the smallest *p*-value with Spearman correlation analysis with the FPKM of SF3B4 (Fig. 6C). Functional enrichment analysis was also performed to identify the biological processes for SF3B4's target genes, resulting in a dot plot of 9 biological processes with $P < 0.05$ (Fig. 6D).

Discussion

We identify several AS events in ESCC. It has been reported that a natural endogenous soluble form of VEGFR-2, a product of alternative splicing, is present in humans and mice [30]. Our previous research found that endogenous soluble VEGFR-2 is down-regulated in ESCC, but its high expression is an independent prognostic factor for poor survival [31]. Previous studies have identified at least three isoforms of the human RASSF5 gene, of which RASSF5A is silenced in esophageal tumor cell lines [32]. However, alternative splicing studies in ESCC remain at a low-throughput level, and analyses of AS events in ESCC on a genome-wide scale are rare. Although aberrant splicing events in ESCA samples have been found and survival-related, differentially-spliced genes have been identified, detailed splicing patterns of ESCC are still unknown, and splicing-mode changes between ESCC and matched normal tissues have not been characterized [33]. In this paper, previously published next-generation sequencing data was used to study AS in ESCC. Because our RNA-seq data was generated from ESCC tissue and matched normal tissue of the same patient, our data was very suitable for MISO analysis and splicing-mode change characterization. There was also a pair of samples, in the RNA-seq data, that was from a normal esophageal epithelial (SHEE) and an esophageal squamous cell carcinoma (SHEEC) cell line, which could be used as a validating dataset.

In total, 6,019 AS events were identified with significant differences in 15 pairs of ESCC tissues. It is also noteworthy that a large proportion of AS events was found to only have significant differences in a small number of paired samples, which suggests that AS differences in many genes might have a similar impact on the development of ESCC. The proportion of AS events with $\Delta\Psi > 0$ and $\Delta\Psi < 0$ in ESCC tends to be consistent, which indicates that the existing AS classifications cannot generalize all the true AS types, and the real AS modes may be more complicated. We mentioned that MISO's AS annotations were based on existing transcripts, so alternative splicing analysis could not find new transcripts and genes, but unknown AS regulatory mechanisms based on known genes and transcripts could be revealed by the MISO analysis.

Even so, there are many problems with official MISO AS annotations, in which several exons in mutually exclusive exon events are found present in the same transcript, making the accuracy of the annotation confusing. In this study, transcriptional annotations from Ensembl and RefSeq database were used to re-annotate the previously analyzed AS events and to remove the wrong splicing events to make the subsequent analysis more accurate.

Other laboratories have reported on the mechanism of alternative splicing affecting invasiveness of cancer. For example, the normal transcript for CD44 was expressed in normal gastric tissue, and the upregulation of its variants is associated with invasion and metastasis of gastric tumors [34]. Higher expression of CD44 variant 6 has been reported in gastric tumor with lymph node metastasis, suggesting that CD44 variant 6 plays a role in the metastasis of gastric cancer [35]. It has also been reported that the up-regulation of OPN-b is associated with deeper local invasion and late metastasis of tumor lymph nodes [36]. For the first time, from the perspective of systems biology, we show the biological processes and KEGG pathways affected by alternative splicing in ESCC development. There

are many GO terms and pathways related to cell junction and cell migration, indicating that altered splicing may govern invasion and metastasis in ESCC.

HNRNPC is a typical splicing factor, and its regulation by alternative splicing has not been reported. Alternative splicing of ZNF207 has been reported to be regulated by SRSF11 [37], but its AS pattern has not been reported. VCL encodes an F-actin-binding cytoskeletal protein, compared to VCL-001, VCL-201 contains an extra exon named exon 19. The skipped exon 19 is reported to be increasingly preferred in tumor cells compared to normal colon mucosa [38], which is consistent with the Sashimi plot here. Calmodulin is an actin-binding protein with regulatory functions, and has been reported to play an important role in cell movement. In colon cancer, the longest Calmodulin was reported to be down-regulated [38]. However, Sashimi plot results show that there is an alternative 5' splice site event in CALD1, which also involves a relatively down-regulated expression of the longest isoform. Intron retention of KIAA1217 has been found in non-small cell lung cancer, but RT-PCR results do not support this result [39]. Sashimi plot results here indicate that there is an RI event for KIAA1217, but the retained intron tends to be included in matched normal tissues. These results need further verification. Regarding TPM1, it has been reported that exon 6 of TPM1 has two types of AB with the same size of 76 bp. In bladder cancer and prostate cancer, exon 6A and 6B are reported to have opposite splicing trends [40]. We also show there is an MXE event in TPM1s, and both the mutually exclusive exons were 76 bp.

SF3B4 is reported to be a diagnostic marker and overrepresented in HCC [41], but the relationship between SF3B4 and ESCC has not been identified. In this paper, we have found SF3B4 up-regulated in ESCC with a fold-change of about 1.5, which is similar to the results of an HCC study [41]. Moreover, according to the log-rank (Mantel-Cox) test, SF3B4 can be identified as a prognostic marker of ESCC. Spearman correlation analysis was utilized to calculate the correlation between the degree of splicing events and the expressions of splicing factors in ESCC, according to a previous method [42]. The candidate splicing factor SF3B4 was found to play an important role in regulating 92 protein coding gene aberrant splicing events in ESCC. Functional enrichment analysis indicates that SF3B4's downstream targets may play a role in cell-cell junction and nuclear factor kB (NF-kB) processes. However, it is still necessary to verify their regulatory roles through validation experiments.

Conclusion

In this study, AS events were analyzed from RNA-seq data of pairs ESCC tissue samples and two cell lines and 13.25% significant differentially AS events were identified which are involved in cell junction and migration. The splicing factor SF3B4 is a potential prognostic factor of ESCC. In summary, we identified thousands of significant differentially AS events which might be related to the development of ESCC.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no conflicts of interest.

Funding

This work was supported in part by the National Science Foundation of China (Grant No. 81672473, 8150213861602292, 81772532 and 81872372), National Cohort of Oesophageal Cancer of China (Grant No. 2016YFC09014000), and the Science and Technology Program of Guangdong (No.2014A030310390, No.2017A030313181).

Authors' contributions

Conceptualization, L.Y.X., EM.L. and B.L.W.; Funding acquisition, L.Y.X., EM.L. and B.L.W.; Data curation, J.Y.D. and C.Q.L.; Software and Methodology, J.Y.D., C.Q.L., Y.W.C., Z.P.D., Q.Y.W., Z.D.T. and C.S.; Project administration, Q.X.X., W.J.B., L.L. and W.L.; Writing—original draft, J.Y.D.; Writing—review and editing, L.Y.X., EM.L. and B.L.W. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We thank Dr. Stanley Lin for proof-read the manuscript.

References

1. Smyth EC, Lagergren J, Fitzgerald RC, Lordick F, Shah MA, Lagergren P, Cunningham D. Oesophageal cancer. Nat Rev Dis Primers. 2017; 3:17048.
2. Siegel RL, Miller KD. Cancer statistics, 2019. 2019; 69(1):7-34.
3. Bonnal S, Vigevani L, Valcarcel J. The spliceosome as a target of novel antitumour drugs. Nat Rev Drug Discov. 2012; 11(11):847-859.
4. Wahl MC, Will CL, Luhrmann R. The spliceosome: design principles of a dynamic RNP machine. Cell. 2009; 136(4):701-718.
5. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & development. 2010; 24(21):2343-2364.
6. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221):470-476.
7. Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell. 2009; 136(4):777-793.
8. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nature reviews Genetics. 2016; 17(1):19-32.

9. Kaida D, Schneider-Poetsch T, Yoshida M. Splicing in oncogenesis and tumor suppression. Cancer science. 2012; 103(9):1611-1616.
10. Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E, Lucier JF, Thibault P, Rancourt C, Tremblay K, Prinos P, Chabot B, Elela SA. Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol. 2009;16(6):670-6..
11. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods. 2010; 7(12):1009-1015.
12. Li CQ, Huang GW, Wu ZY, Xu YJ, Li XC, Xue YJ, Zhu Y, Zhao JM, Li M, Zhang J, Wu JY, Lei F, Wang QY, Li S, Zheng CP, Ai B, Tang ZD, Feng CC, Liao LD, Wang SH, Shen JH, Liu YJ, Bai XF, He JZ, Cao HH, Wu BL, Wang MR, Lin DC, Koeffler HP, Wang LD, Li X, Li EM, Xu LY. Integrative analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma. Oncogenesis. 2017;6(2):e297..
13. Andrews S. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
14. Pennisi E. The human genome. Science. 2001 Feb 16;291(5507):1177-80.
15. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013; 14(4):R36.
16. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012; 9(4):357-359.
17. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010; 28(5):511-515.
18. Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010; 11(10):R106.
19. Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, Wang S, Zhou F, Shi S, Feng X, Sun N, Liu Z, Skogerboe G, Dong J, Yao R, Zhao Y, Sun J, Zhang B, Yu Y, Shi X, Luo M, Shao K, Li N, Qiu B, Tan F, Chen R, He J. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. Gut. 2014;63(11):1700-10.
20. Li W, Zhang L, Guo B, Deng J, Wu S, Li F, Wang Y, Lu J, Zhou Y. Exosomal FMR1-AS1 facilitates maintaining cancer stem-like cell dynamic equilibrium via TLR7/NFkappaB/c-Myc signaling in female esophageal carcinoma. Molecular cancer. 2019; 18(1):22.
21. Su H, Hu N, Yang HH, Wang C, Takikita M, Wang QH, Giffen C, Clifford R, Hewitt SM, Shou JZ, Goldstein AM, Lee MP, Taylor PR. Global gene expression profiling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes. Clin Cancer Res. 2011;17(9):2955-66.
22. Li WQ, Hu N, Burton VH, Yang HH, Su H, Conway CM, Wang L, Wang C, Ding T, Xu Y, Giffen C, Abnet CC, Goldstein AM, Hewitt SM, Taylor PR. PLCE1 mRNA and protein expression and survival of patients with esophageal squamous cell carcinoma and gastric adenocarcinoma. Cancer Epidemiol Biomarkers Prev. 2014;23(8):1579-1588..
23. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P, Searle SM. The Ensembl gene annotation system. Database (Oxford). 2016;2016:baw093..

24. Pruitt, KD, Tatusova T, Brown GR, Maglott, DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012; 40, D130-5.
25. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009; 4(1):44-57.
26. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009; 37(1):1-13.
27. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. 2016.
28. Piva F, Giulietti M, Burini AB, Principato G. SpliceAid 2: a database of human splicing factors expression data and RNA target motifs. Human mutation. 2012; 33(1):81-85.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13(11):2498-2504.
30. Albuquerque RJ, Hayashi T, Cho WG, Kleinman ME, Dridi S, Takeda A, Baffi JZ, Yamada K, Kaneko H, Green MG, Chappell J, Wilting J, Weich HA, Yamagami S, Amano S, Mizuki N, Alexander JS, Peterson ML, Brekken RA, Hirashima M, Capoor S, Usui T, Ambati BK, Ambati J. Alternatively spliced vascular endothelial growth factor receptor-2 is an essential endogenous inhibitor of lymphatic vessel growth. Nat Med. 2009;15(9):1023-30.
31. Wu ZY, Chen T, Zhao Q, Huang JH, Chen JX, Zheng CP, Xu XE, Wu JY, Xu LY, Li EM. Altered expression of endogenous soluble vascular endothelial growth factor receptor-2 is involved in the progression of esophageal squamous cell carcinoma. J Histochem Cytochem. 2013 May;61(5):340-7.
32. Guo W, Wang C, Guo Y, Shen S, Guo X, Kuang G, Dong Z. RASSF5A, a candidate tumor suppressor, is epigenetically inactivated in esophageal squamous cell carcinoma. Clin Exp Metastasis. 2015;32(1):83-98.
33. Mao S, Li Y, Lu Z, Che Y, Sun S, Huang J, Lei Y, Wang X, Liu C, Zheng S, Zang R, Li N, Li J, Sun N, He J. Survival-associated alternative splicing signatures in esophageal carcinoma. Carcinogenesis. 2019;40(1):121-130..
34. Li Y, Yuan Y. Alternative RNA splicing and gastric cancer. Mutation research. 2017; 773:263-273.
35. Chen XY, Wang ZC, Li H, Cheng XX, Sun Y, Wang XW, Wu ML, Liu J. Nuclear translocations of beta-catenin and TCF4 in gastric cancers correlate with lymph node metastasis but probably not with CD44 expression. Human pathology. 2005; 36(12):1294-1301.
36. Tang X, Li J, Yu B, Su L, Yu Y, Yan M, Liu B, Zhu Z. Osteopontin splice variants differentially exert clinicopathological features and biological functions in gastric cancer. Int J Biol Sci. 2013;9(1):55-66.
37. Toh CX, Chan JW, Chong ZS, Wang HF, Guo HC, Satapathy S, Ma D, Goh GY, Khattar E, Yang L, Tergaonkar V, Chang YT, Collins JJ, Daley GQ, Wee KB, Farran CA, Li H, Lim YP, Bard FA, Loh YH. RNAi Reveals Phase-Specific Global Regulators of Human Somatic Cell Reprogramming. Cell Rep. 2016;15(12):2597-607.
38. Bisognin A, Pizzini S, Perilli L, Esposito G, Mocellin S, Nitti D, Zanovello P, Bortoluzzi S, Mandruzzato S. An integrative framework identifies alternative splicing events in colorectal cancer development. Molecular oncology. 2014; 8(1):129-141.
39. Langer W, Sohler F, Leder G, Beckmann G, Seidel H, Grone J, Hummel M, Sommer A. Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. BMC genomics. 2010; 11:676.
40. Thorsen K, Sørensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhøffer M, Laurberg S, Borre M, Wang K, Brunak S, Krainer AR, Tørring N, Dyrskjøt L, Andersen CL, Orntoft TF. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. Mol Cell Proteomics. 2008;7(7):1214-24.

41. Shen Q, Eun JW, Lee K, Kim HS, Yang HD, Kim SY, Lee EK, Kim T, Kang K, Kim S, Min DH, Oh SN, Lee YJ, Moon H, Ro SW, Park WS, Lee JY, Nam SW. Barrier to autointegration factor 1, procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3, and splicing factor 3b subunit 4 as early-stage cancer decision markers and drivers of hepatocellular carcinoma. Hepatology. 2018;67(4):1360-1377.
42. Richards AL, Watza D, Findley A, Alazizi A, Wen X, Pai AA, Pique-Regi R, Luca F. Environmental perturbations lead to extensive directional shifts in RNA processing. PLoS Genet. 2017;13(10):e1006995.

Figures

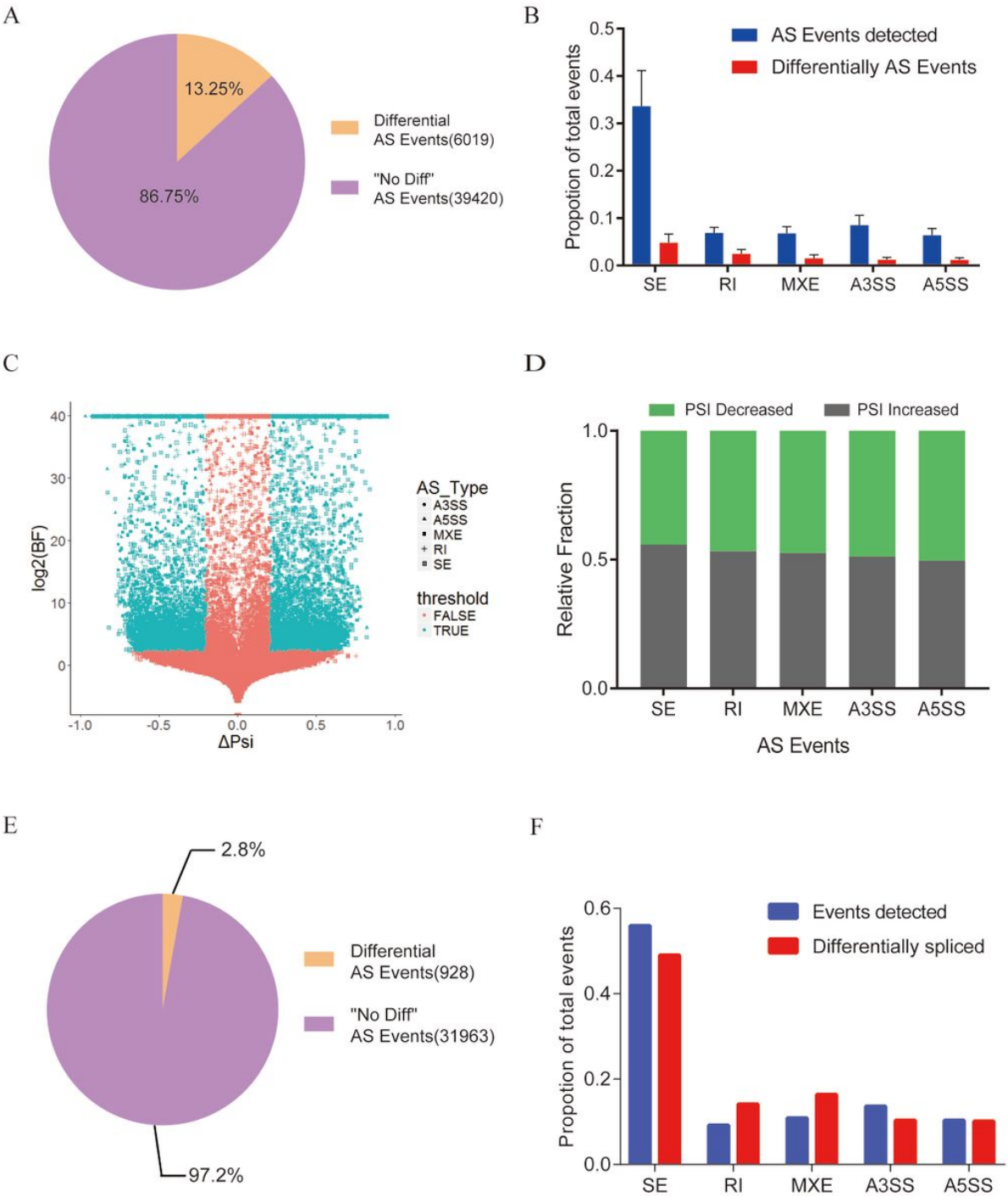

Figure 1

Distribution of RNA splicing events in ESCC after MISO analysis. Percentages of differentially-spliced AS events with significance in total AS events found in ESCC clinical samples compared to the matched normal samples. (B) Plot depicts the fraction of events detected for each AS type (blue) compared to events with significantly different expression in ESCC tissues compared to matched normal tissues (red). (C) Scatter plot of all the AS events identified using MISO. The X-axis represents ΔΨ values, and the Y-axis represents log2 (BF) values. The shape of the dots indicates the type of AS event. Specifically, a hollow square indicates an SE event; cross indicates an RI event; solid square indicates an MXE event; triangle indicates an A5SS event; and circle indicates an A3SS event. Alternative splicing events with BF > 5 and ΔΨ > 0.2 are colored in blue. (D) Relative fraction of each AS event affected positively or negatively by ESCC. (E) Percentage of alternative splicing events showing significant changes in SHEE/SHEEC cell line association compared to detectable alternative splicing events showing no difference. (F) Plot depicts the fraction of events detected for each AS type (blue) compared to events with significantly different expressions in SHEE/SHEEC cell lines to matched normal tissues (red).

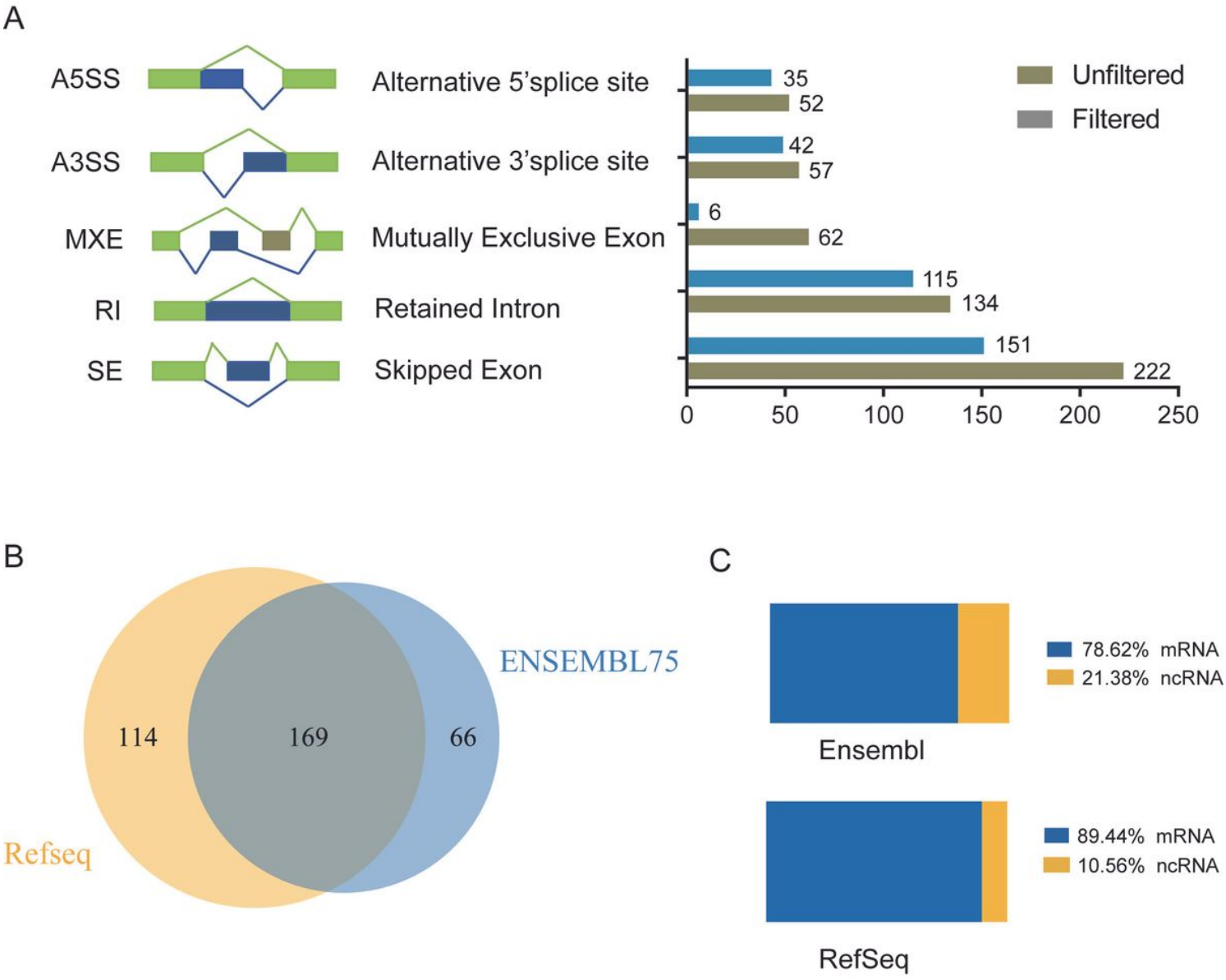

Figure 2

Distribution of RNA splicing events with significant differences in ESCC after re-annotation with Ensembl and RefSeq gene annotations. (A) Plot in the left panel indicates exon models of the types of AS assessed by MISO

analysis. In the right panel, the different types of AS events before and after re-annotation were quantified with transcript annotations from Ensembl and RefSeq database, respectively. (B) A total of 283 and 235 splicing events with significantly altered psi values were identified using transcript annotations from the RefSeq and Ensembl databases, respectively. In total, 169 splicing events were identified in both database transcript annotations. (C) Percentage of protein coding transcripts (blue) and non-coding transcripts (yellow) representing AS events after re-annotation analysis.

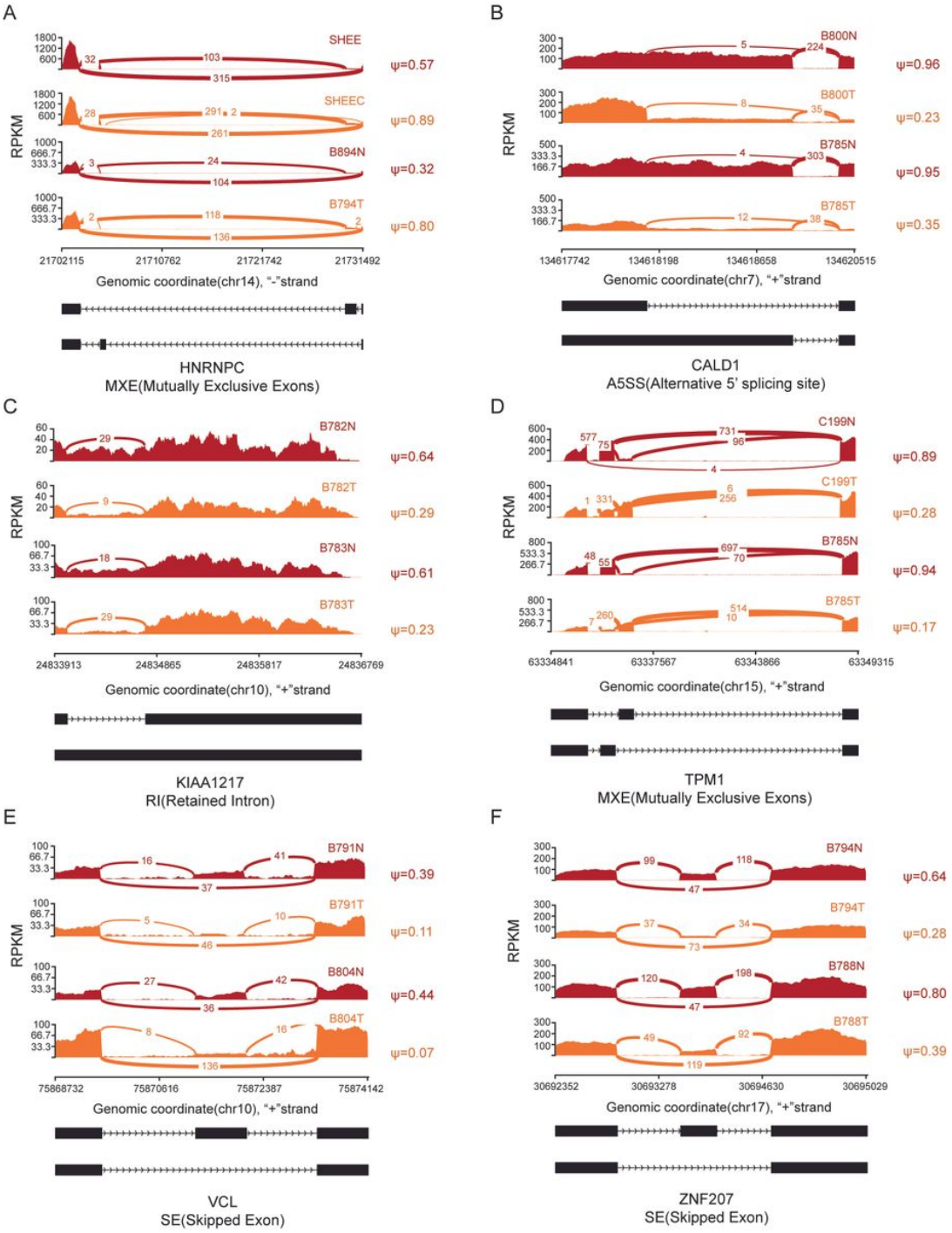

Figure 3

Sashimi plot of splicing changes in ESCC. Diagrams on the left show the read coverage of exons. Plots on the right show the full posterior distribution of the Ψ values, along with their confidence intervals. The AS model of this region is represented below. (A)-(F) AS models of HNRNPC, CALD1, KIAA1217, TPM1, VCL and ZNF207, respectively.

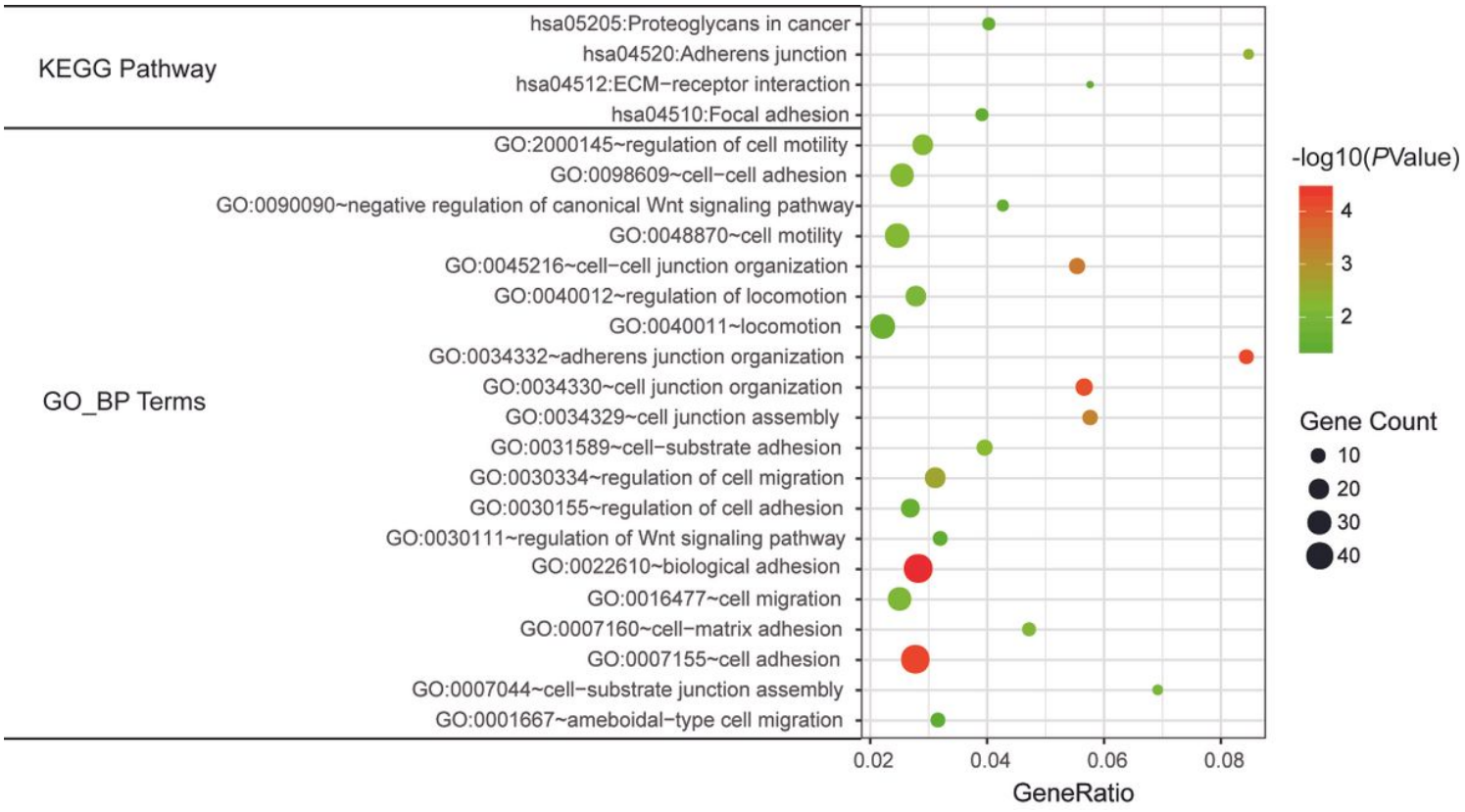

Figure 4

Dot plot of significant GO_BP terms and KEGG pathways using functional enrichment analyses with aberrant splicing event-related genes.

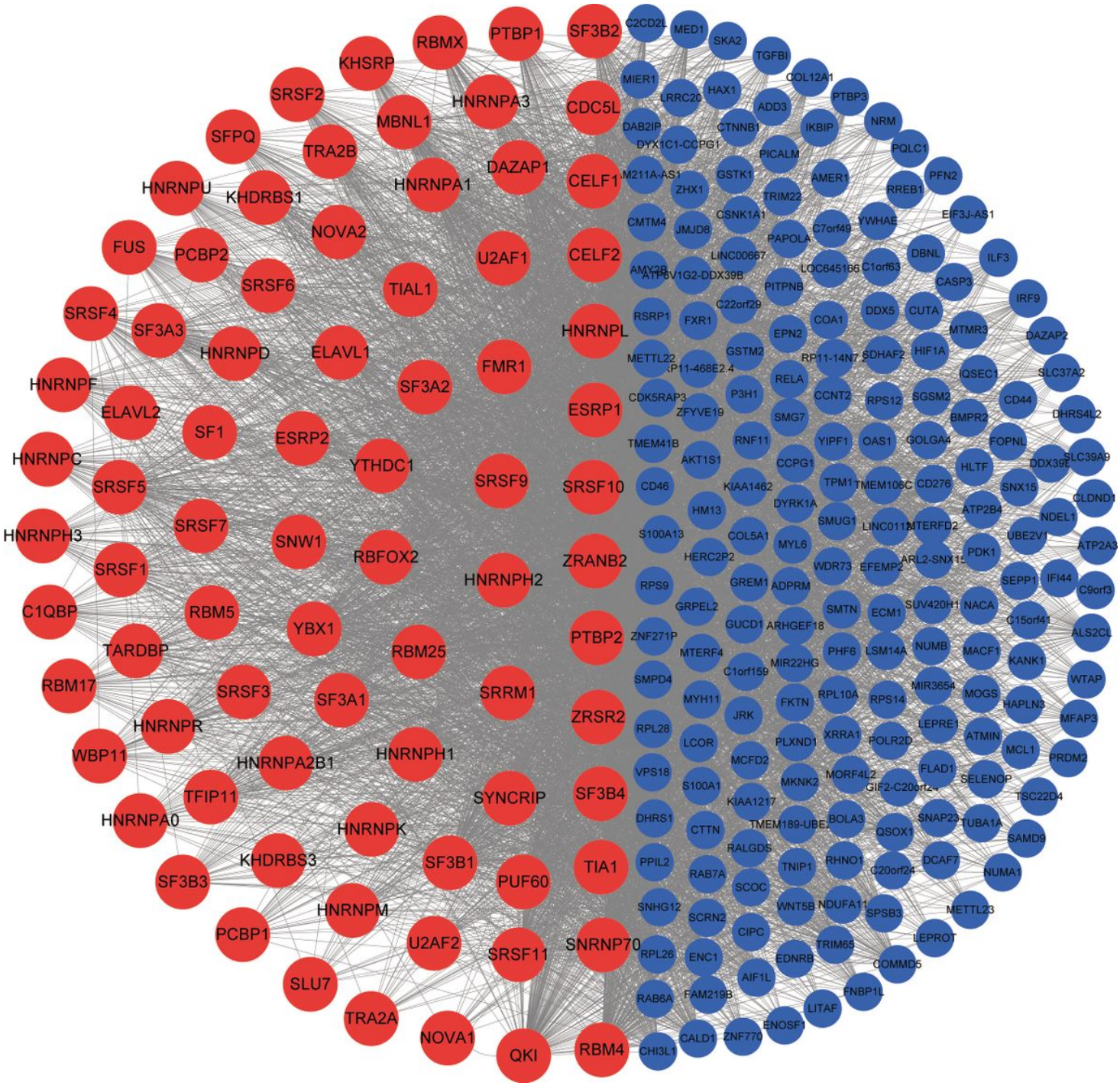

Figure 5

Regulatory network of splicing factor-AS events representing genes.s

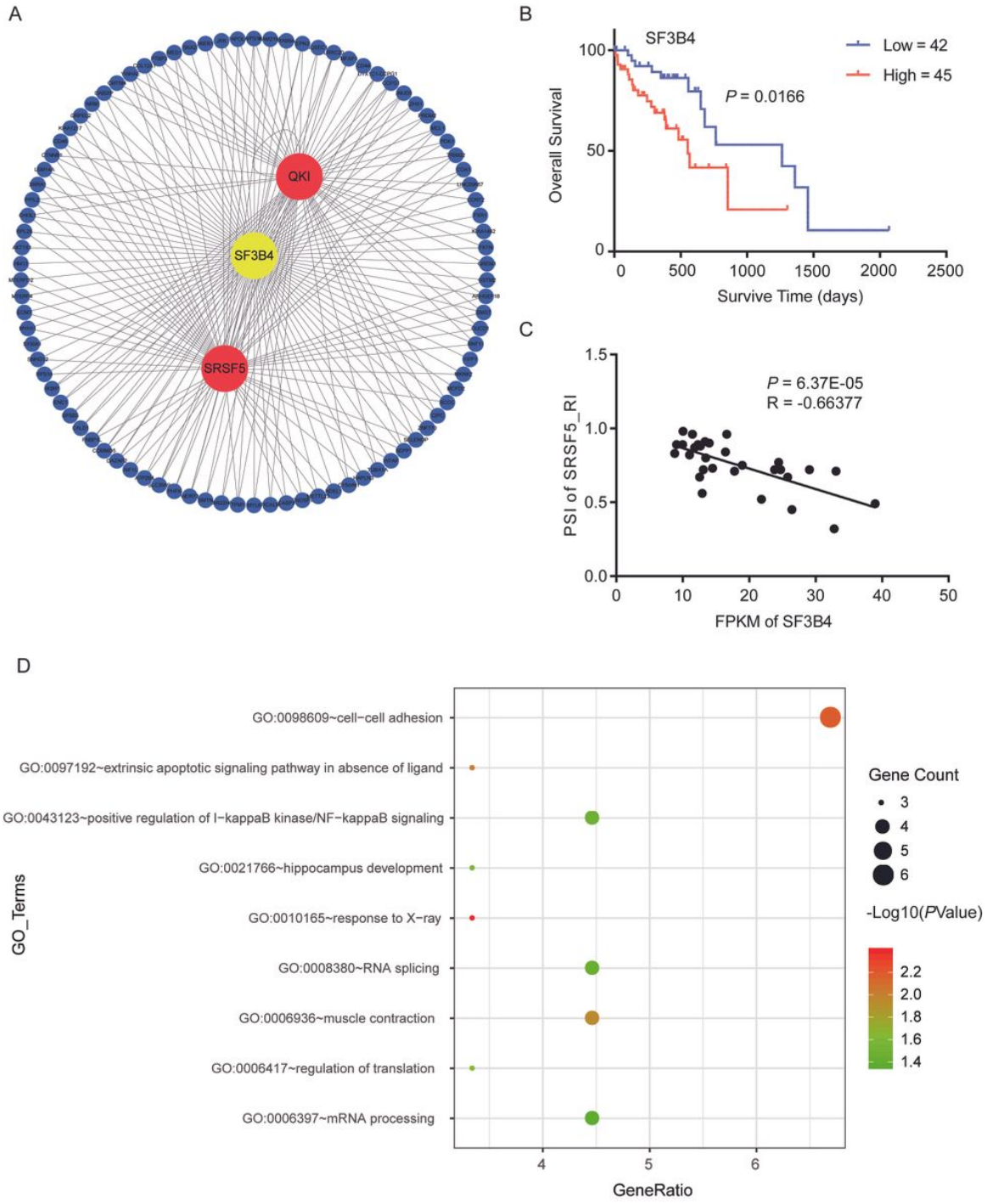

Figure 6

SF3B4-related regulatory network in ESCC. (A) SF3B4 splicing regulatory network. (B) K-M curve for SF3B4 in ESCC. (C) Scatter plot indicating correlation between SF3B4 expression and psi value of the RI event of SRSF5. (D) Dot plot of GO_BP terms using alternatively-spliced SF3B4-targeted genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable2.docx

- SupplementaryTable1.docx