Cancer Genetics 230 (2019) 58–65

Cancer Genetics

SHORT COMMUNICATION

Genome-wide isoform-level analysis reveals tumor-specific isoforms for lung adenocarcinoma diagnosis and prognosis

Hu Zhuhong[a], Bai Zhenyu[b], Chen Xiangyuan[c], Xu Tingzhen[c], Song Libin[d,*]

[a] *The First Affiliated Hospital of Zhejiang Chinese Medical University, Zhejiang, China;* [b] *Department of Laboratory, General Hospital of Pingmei Shenma Medical Group, Henan, China;* [c] *Department of Integrative Oncology, Fudan University Shanghai Cancer Center, Shanghai, China;* [d] *Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, China*

Abstract

Last decades have witnessed the great progress in exploration of tumor transcriptome. However, most researches were restricted in gene-level expression. mRNA isoforms, especially tumor-specific isoforms have not been fully explored in tumor. Here, by analyzing RNA-seq data derived from hundreds of samples in TCGA projects, we comprehensively characterized the expression variations of mRNA isoforms in adenocarcinoma of lung (LUAD), which is one of leading causes of cancer-related death. Our analysis found that a variety of mRNA isoforms showed differential expression in LUAD tumor samples. Some of them even showed distinct variations compared to their host genes. Further analysis of functional enrichment revealed that up- and down-regulated mRNA isoforms took part in different types of biological process. In addition, we also identified hundreds of isoforms that expressed exclusively in LUAD tumor samples. Furthermore, the expression level of several isoforms, such as uc001kuk.3 and uc003yls.2, could separate tumor patients by overall survival periods. Our study provided new candidates for the diagnosis and prognosis of lung cancer.

Keywords Lung cancer, mRNA isoform, Functional analysis, Tumor-specific, Diagnosis.

Introduction

The primary goal across diverse areas of oncology is identification of tumor-specific molecules for applications in early diagnosis, prognosis and the therapeutic strategy design. Among various molecules, RNA has been proven to be one of the most attractive ones mainly because of its prevalence in cells which enables the highly sensitive early detection and diagnostic assays [1,2]. The main themes of tumor investigations during the last few decades have been focused on gene-level aberrance. Actually, “gene” is a collective term for all RNA isoforms transcribed from a genomic locus [2]. Experiments based on “gene” concept hindered the efforts to discover molecules expressing in a tumor-specific way. Recent studies have proven that isoform-level expression could represent the molecule alterations more clearly in tumor [3–6].

High-throughput sequencing technologies have provided researchers powerful tools to investigate RNA level expression [7]. Transcriptome sequencing (RNA-seq) could enable genome-wide quantification of mRNA isoforms expression. Subhash et al. employed high throughput RNA-seq analysis of buccal mucosal cancer (BMC) to identify aberrant transcriptional events [8]. Their analysis revealed hundreds of aberrant isoforms and splice variants that might be helpful in deep understanding of pathological changes. A recent study in ovarian cancer developed custom bioinformatics algorithms to identify mRNA isoforms with tumor-specific expression from hundreds of RNA-seq datasets [2]. Their study found that tumor-specific isoforms played important roles in the development of ovarian cancer and might be useful in the clinical diagnosis and molecular therapy. Wei et al. systematically analyzed the isoform-level expression patterns and

Received June 25, 2018; received in revised form October 9, 2018; accepted November 7, 2018

* Corresponding author.
E-mail address: huzh1020@163.com

2210-7762/$ - see front matter © 2018 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.cancergen.2018.11.004

isoform switching events by using 819 mRNA-seq datasets of breast tumor and normal samples [5]. Wei et al. found that most isoforms were associated with breast cancer subtypes and showed biological significance. In addition, a number of large databases such as The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) have released mountains of RNA-seq data derived from malignant and normal tissues. Public resources provide opportunities to promote tumor research, especially in diagnosis and therapy. Nevertheless, to date these data have not been rigorously evaluated for the discovery of tumor-specific molecules in many types of tumors, including lung adenocarcinoma.

Lung cancer is one of the leading causes of cancer-related mortality worldwide, the most common histological type of which is adenocarcinoma [9]. Smoking is the major risk for lung adenocarcinoma, however, last decades have witnessed increasing proportion cases in non-smoking individuals [10]. Recently, molecule-target therapies proved dramatic improvement on treatment for patients with specific genomic features. Such genes as EGFR [11], TP53 [12], ALK [13] and KEAP1 [14] were investigational targets for significant genomic alterations. However, the genomic abnormalities of these genes are not clinically actionable. Recent study focusing in tumor transcriptome revealed various RNA molecules as potential diagnosis and therapy biomarkers. Hence, it is a necessity to investigate the expression alterations of isoforms and identify tumor-specific isoforms genome-wide in lung adenocarcinoma.

Our study is the first to comprehensively characterize isoform-level expression in lung adenocarcinoma (LUAD). A variety of isoforms were found to express differentially in LUAD tumor samples compared to normal samples. Differentially expressed isoforms showed biological significance in LUAD. We also identified a number of tumor-specific isoforms that could be potential biomarkers for clinical diagnosis. Several tumor-specific isoforms showed prognostic potential to predict the overall survival period of LUAD patients.

Materials and methods

TCGA RNA-seq data

RNA-seqV2 read counts and patient clinical parameters of 573 samples from the TCGA lung adenocarcinoma (LUAD) cohort were downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). The raw sequencing reads were aligned to the human genome (hg19) by the TCGA consortium using MapSplice [15]. Additionally, the quantification was performed by RSEM [16] against the reference transcriptome of UCSC hg19 GAF2.1(17) for known genes and isoforms [18]. Furthermore, the raw read counts for each gene and isoform were normalized as RPKM (Reads Per Kilobase per Million mapped reads). RPKM value is a method for quantifying gene expression from RNA sequencing data by normalizing for feature (such as genes) length and the number of total sequencing reads [19]. We calculated the RPKM value for each gene or isoform as follows:

$$\mathrm{RPKM}_i = 1000,000,000 \times \frac{R_{ij}}{N_j \times L_i}$$

Where R_{ij} is the count of reads mapped to gene (or isoform) i in sample j. N_j represents the total number of reads mapped to all genes in sample j and L_i for the length of exonic regions in gene (or isoform) i.

Identification of differentially expressed genes and isoforms

Included in the LUAD samples were 58 paired tumor/normal samples, on which differential expression analysis was performed using paired Student's t test. All genes and isoforms that were differentially expressed (Benjamini-Hochberg corrected p-value ≤ 0.01 and fold change ≥ 2 or ≤ 0.5) between paired tumor and normal samples were selected for further analysis.

Calculate the PSI values for isoforms

For isoforms, the Percent Spliced In (PSI) value represents relative abundance of the included isoforms among the same genes. The PSI values were calculated as follows:

$$PSI_{gi} = \frac{RPKM_{gi}}{\sum_{j=1}^{n} RPKM_{gj}}$$

Where PSI_{gi} denotes the PSI value of isoform i that belongs to gene g, and gene g has a total of n isoforms. RPKM_{gi} represents the RPKM value of isoform i.

Functional enrichment analysis

Gene Ontology (GO) and KEGG pathway gene sets were retrieved from Molecular Signatures Database (MSigDB v6.0) [20]. Functional enrichment were performed using Hypergeometric test to identify significantly relevant biological process and pathways of the differentially expressed isoforms (related to 2432 up-regulated genes and 1842 down-regulated genes).

Identification of genes with switch isoforms

The expression contribution of distinct isoforms from the same genes might alter in different biological conditions. For each gene, the dominant isoforms were first identified in all paired samples. Genes that expressed different dominant isoforms in normal and paired tumor samples were considered to have switch isoforms in tumor. We then computed the occurrence frequency of switch isoforms in tumor samples. The PSI difference between dominant and second rank isoforms in tumor samples was also calculated. A switch isoform with occurrence frequency higher than 30% and average PSI difference no less than 0.3 was defined as valid switch isoforms.

Identification of tumor-specific isoforms

First of all, isoforms with RPKM value larger than 0.1 were considered as expressed isoforms. Isoforms that were not

expressed in all benign adjacent samples ($n=58$) were identified as tumor-expressed isoforms. To filter out isoforms that were low-expressed across all samples, isoforms that were detected expressed in less than ten percent of all tumor samples ($n=515$) were further discarded. All remained isoforms were considered as tumor-specific isoforms.

Survival analysis

Expression levels across all tumor samples of each tumor-specific isoforms were used to explore whether they were associated with prognosis of tumor patients. For each tumor-specific isoform, all tumor samples were dichotomized into low- and high-expression group by using the median expression level as cutoff. The Kaplan-Meier method was further used to evaluate the overall survival time for the two groups. The differences between overall survival times of two groups were compared by log-rank test. In addition, isoforms that divided patients with longer survival time into high-expression group were considered as positive tumor-specific isoforms. Conversely, negative tumor-specific isoforms divided patients with shorter survival time into high-expression group.

Statistical analysis and figures

All statistical analyses used in our study were performed in R software [21]. Figures were also built using R software. Without specific statement, statistical tests with a *P*-value less than 0.05 were considered significant.

Results

Exploring the expression of genes and isoforms in TCGA LUAD cohort

To identify genes and isoforms that exhibit potential roles in the carcinogenesis process of LUAD, expression profiles of 573 LUAD samples were obtained from TCGA, including 58 paired tumor/normal samples (Fig. 1(A)). In addition, the UCSC annotation (GAF2.1) [17] was used as the gene model, which contains 20,531 genes and 73,599 previously determined isoform definitions (Fig. 1(A)). In our study, only these known isoform definitions or gene model were analyzed and no *de novo* isoform discovery was performed. Most genes can be transcribed into diverse isoforms, which

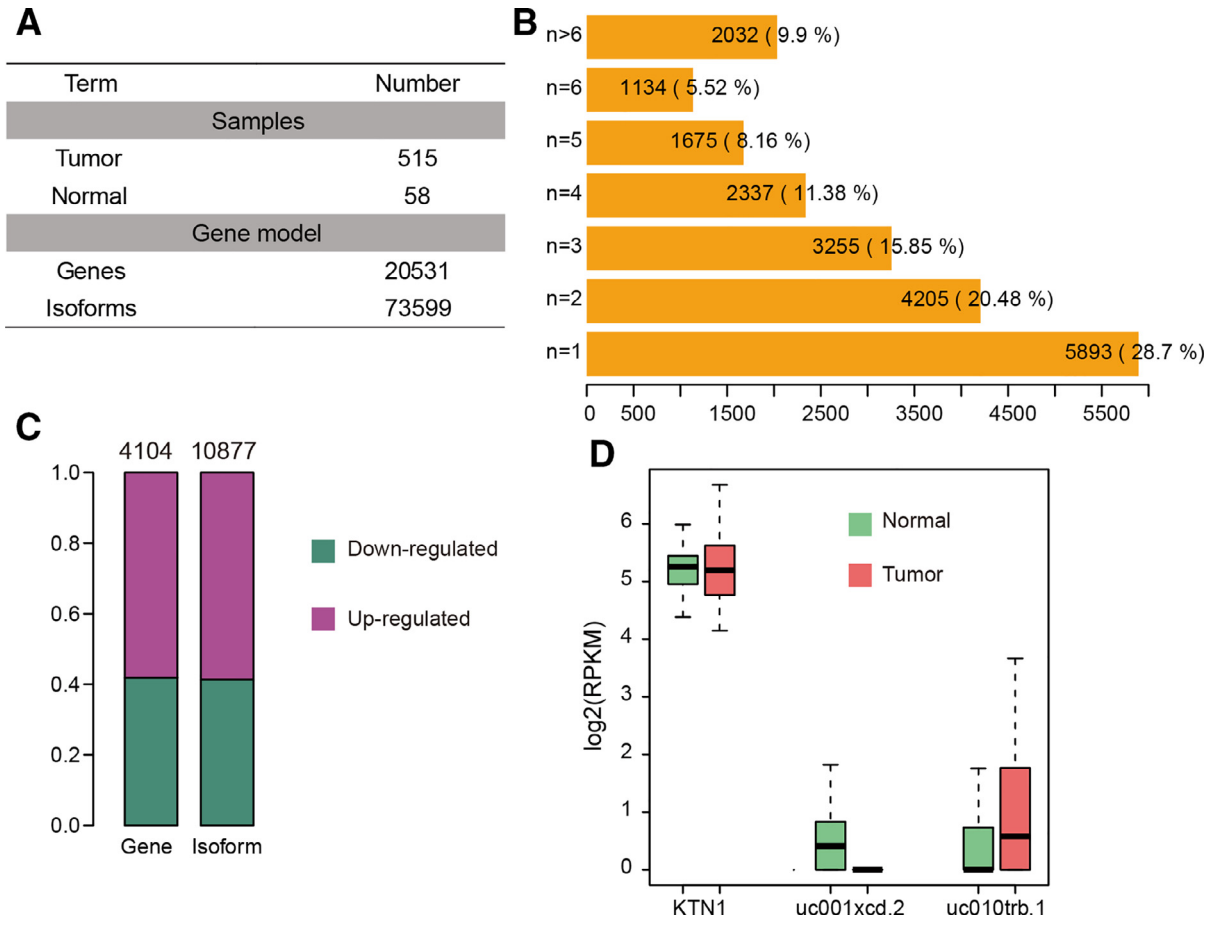

Fig. 1 The status of genes and isoforms in TCGA annotation. (A) Numbers of normal and tumor samples, numbers of genes and isoforms in TCGA LUAD cohort. (B) Distribution of isoform numbers in genes. (C) Numbers of up-regulated and down-regulated genes and isoforms in tumor samples compared to normal samples. (D) Expression level of *KTN1* gene and two of its isoforms *uc001xcd.2* and *uc010trb.1* in normal and tumor samples.

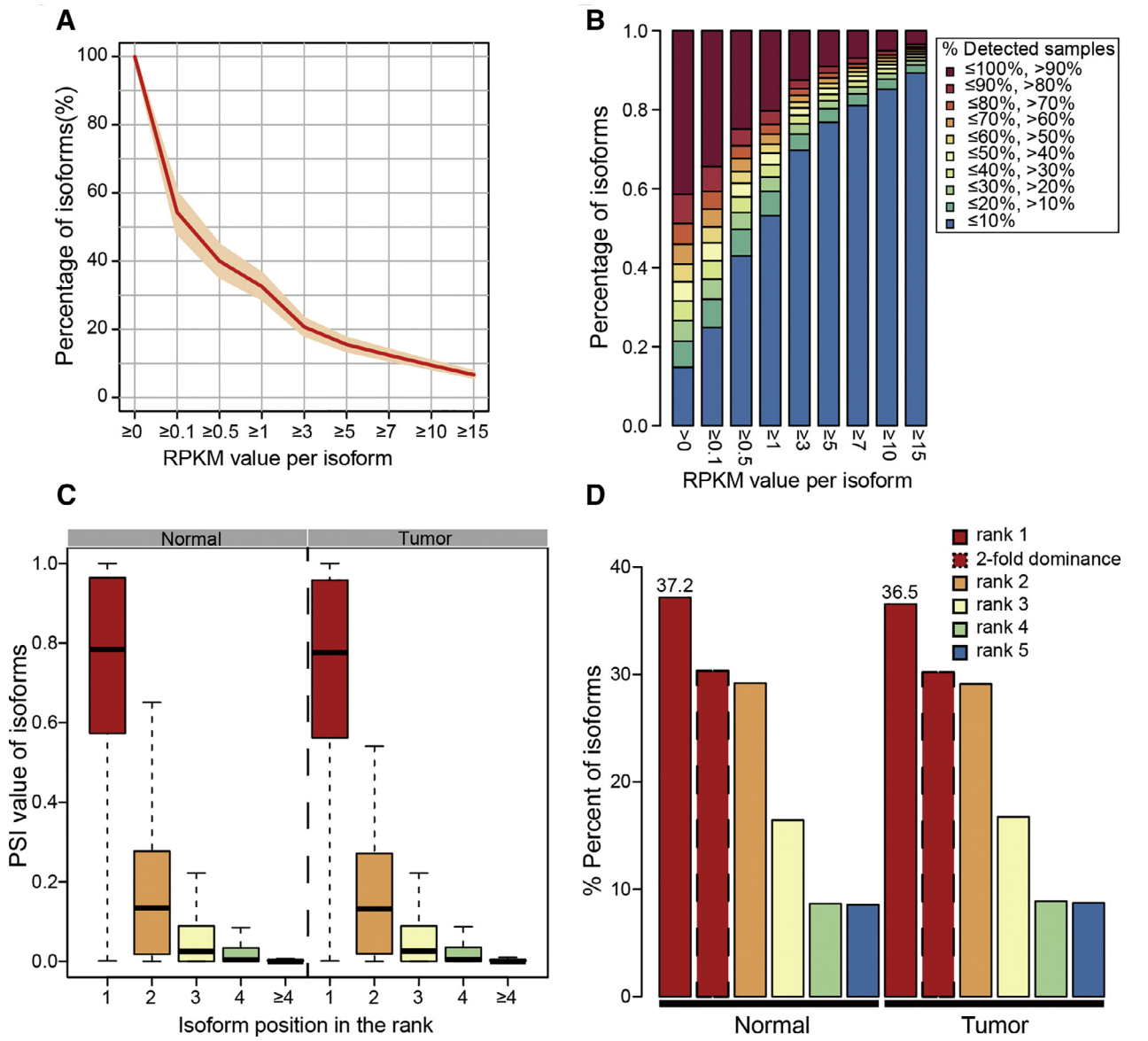

Fig. 2 The expression overview of isoforms in TCGA LUAD cohort. (A) Overall distribution of expression level of isoforms. (B) Sample distribution of isoforms with different expression levels. (C) The expression contributions of distinct isoforms from the same genes in normal or tumor samples. (D) The percentage of isoforms with different expression ranks.

extends the transcriptional diversity. In our gene model, over two third (71.3%) of genes have multiple isoforms (Fig. 1(B)), and 34.96% of those have more than three. The expression levels of genes were quantified by counting the reads mapped to the whole exonic regions, whereas the junction coverage was an extra consideration in the estimation of isoform expression. Obviously, reads that were counted into the isoform expression also contributed to the expression of corresponding genes. For genes having diverse isoforms, not all isoforms were supposed to positively correlate with genes in expression level. In order to testify this, correlations of expression levels between genes and their isoforms were calculated. Consequently, about one third (35.2%) of the multi-isoform genes showed this discrepancy.

The prevalent discrepancy between different isoforms belonging to the same genes raised the question that genes that were not identified differential expression might harbor differentially expressed isoforms. To further the investigation, differential expression analysis for both genes and isoforms was performed by comparing paired tumor and normal samples (see Methods). Consequently, 4104 differentially expressed genes (DEGs) and 10,877 differentially expressed isoforms (DEIs) were identified in LUAD (Fig. 1(C)), and most of these DEGs or DEIs were up-regulated. Although the host genes of most DEIs (7434, 68.3%) were DEGs, a non-negligible part of DEIs were transcribed from genes that were not differentially expressed. Among the non-differentially expressed genes, 2545 genes harbored isoforms that exhibited differential expression levels. Several cancer-related genes were amid these genes, such as KTN1 (Fig. 1(D)), which was reported in liver cancer [22] and colorectal cancer [23], but not in lung cancer.

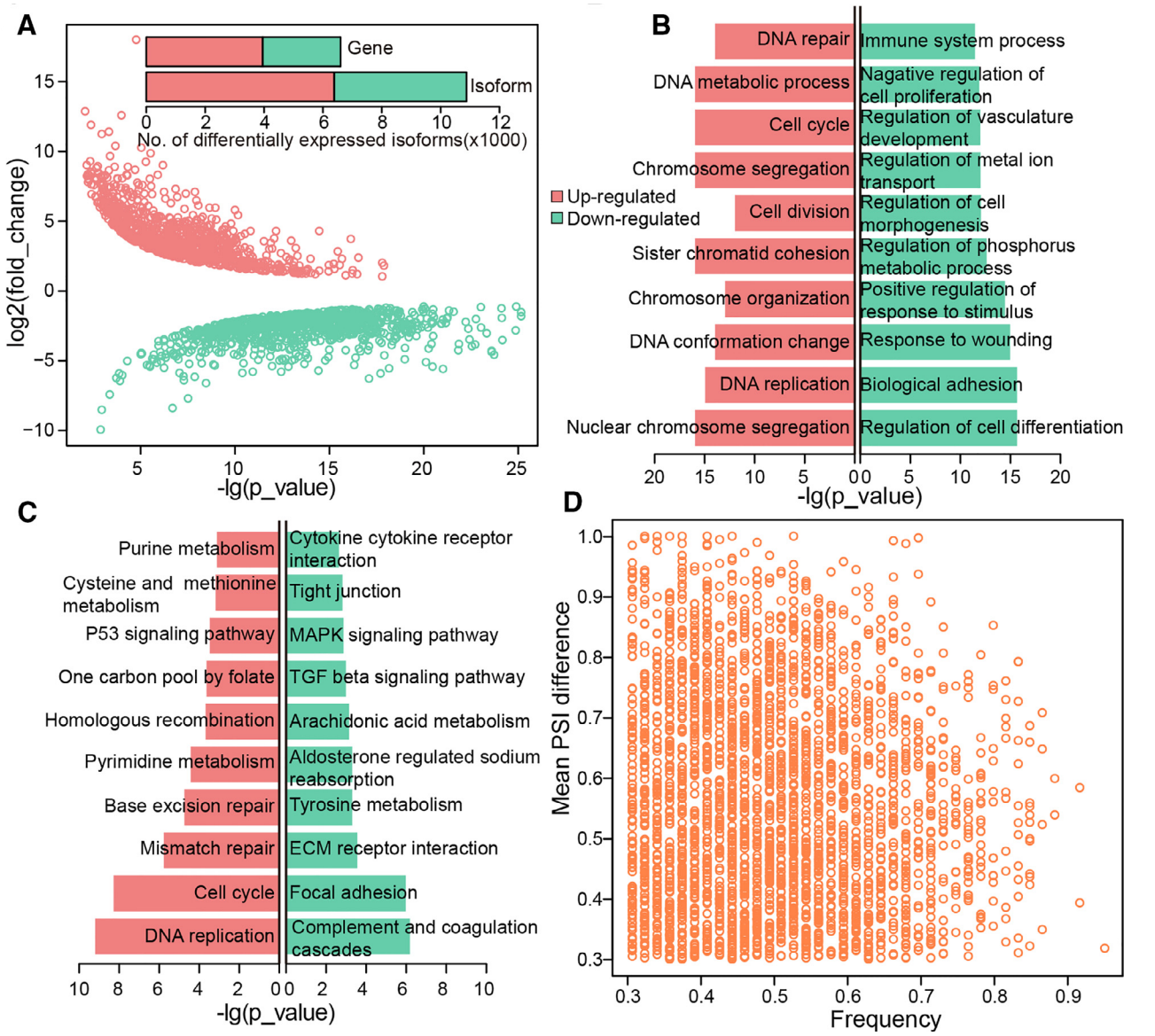

Fig. 3 Functional investigation of differential expression isoforms. (A) Scatter plot shows the difference (fold change and p values) of 1000 up-regulated and 1000 down-regulated isoforms with top difference scores. (B) Significant GO biological process of functional enrichment in up- and down-regulated isoforms. (C) Significant KEGG pathways of functional enrichment in up- and down-regulated isoforms. (D) Scatter plot indicates the occurrence frequency and PSI values of genes with switch isoforms.

These observations showed that some non-differential genes can be identified to play potential roles in tumor through the analysis of mRNA isoforms. The "gene-centered" research in tumor might make some intriguing things still unexplored.

Most genes expressed dominant isoforms

As the discrepancy observed above implicated that mRNA isoforms might play crucial roles in tumor biology, we further investigated the variations of isoforms in LUAD. In order to study the potential roles of mRNA isoform in LUAD, we first analyzed the isoform-level expression across the lung tumors and paired normal lung tissues. Different cutoff for isoform detection was used to inquire the distribution of isoform expression. Across all these samples, an average of 45.6% isoforms were expressed under 0.1 RPKM value and 6.7% isoforms showed expression level that were higher than 15 RPKM (Fig. 2(A)). On average, 59.7% of all isoforms were detected in less than 10% samples and approximately 17.6% isoforms were expressed in more than 90% samples (Fig. 2(B)).

Among the various isoforms expressed from the same gene, there exists a predominant isoform that shows the most abundant expression. The PSI (Percent Spliced In) values of each isoform from the same genes were compared to investigate the dominant isoforms. When quantifying all the annotated isoforms within a gene based on their relative abundance, we observe the existence of a predominant isoform for most genes in both normal and tumor samples as expected (Fig. 2(C)). In order to limit the degree of sparsity while

including the majority of expressed isoforms, only isoforms that have at least 0.1 RPKM value in more than 50% samples in either normal or tumor condition were considered as expressed isoforms. As a result, 36,473 and 36,605 isoforms were identified as expressed in normal and tumor samples, respectively. Among the pool of expressed isoforms, over one third (37.2% and 36.5% in normal and tumor, respectively, Fig. 2(D)) were the dominant isoforms in both normal and tumor conditions. In addition, isoform dominance of each gene was determined as described in a previous study [6]. Specifically, isoform dominance was quantified by computing the ratio of the expression levels between the dominant isoform and the second most abundant one. Overall, in both normal and tumor conditions, we observed that about one third of the genes shave a two-fold dominant major isoform (that is, expressed at least twice as much as the second most abundant one).

These results showed that a large slice of isoforms were expressed in a relatively small portion of all samples (including both normal and tumor conditions), which indicated their specific roles in either condition. Additionally, over one third of the genes expressed a dominant isoform that expressed more than twice as much as the second most abundant isoform.

Differentially expressed isoforms play roles in the process of LUAD

To further investigate the variation and importance of isoforms in lung cancer, we performed differential expression analysis on all isoforms. Consequently, 10,877 isoforms (16.15% of all isoforms) were shown to differentially express in tumor samples. In particular, 6385 isoforms were up-regulated and 4492 isoforms were down-regulated in LUAD (Fig. 3(A)). Functional enrichment was performed separately in up- and down-regulated isoforms to inspect the abnormality of biological process and pathways in lung cancer. Interestingly, we found that up-regulated isoforms were enriched in process related to genome stability, whereas down-regulated isoforms were more likely to be enriched in cell function process (Fig. 3(B)). Particularly, up-regulated isoforms were inclined to be related to such biological process as "Nuclear chromosome segregation", "DNA replication", "Chromosome organization" and "DNA repair". Nevertheless, isoforms that showed down-regulation in tumor samples were enriched in process like "Regulation of cell differentiation", "Responding wounds", "Positive regulation of response to stimulus" and "Negative regulation of cell proliferation". In addition, most up-regulated isoforms took part in pathways related to nucleotide metabolism and genome stability, such as "Pyrimidine metabolism", "Purine metabolism", "Mismatch repair" and "Base excision repair" (Fig. 3(C)). However, down-regulated isoforms were enriched in pathways like "Complement coagulation cascades", "Focal adhesion", "TGF beta signaling pathway" and "MAPK signaling pathways" (Fig. 3(C)). Genes may use distinct dominant isoforms in different biological conditions to perform various functions. In order to explore the switch of dominant isoforms in tumor, we identified the dominant isoform for each gene in normal and paired tumor samples (see Methods). The switch of dominant isoforms were revealed to occur in thousands of genes between normal and paired tumor samples (Fig. 3(D)). About fifty percent of these isoform switch occurred in over half of the samples. Additionally, the PSI difference of switch isoform was larger than 0.5.

Our analysis revealed that differentially expressed isoforms play important roles in progress of LUAD. Furthermore, distinct biological process and pathways were impacted through up- or down-regulated isoforms. Intriguingly, the alterations of functions in many genes were performed by switching the dominant isoforms.

Table 1 Potential isoform targets for pathway-inhibiting therapies.

Isoform	Gene	Pathway
uc010vtw.1	FGF11	MAPK signaling pathway; Pathways in cancer
uc003iac.2	PITX2	TGF beta signaling pathway
uc009xco.2	LAMB3	Focal adhesion;Pathways in cancer;Small cell lung cancer
uc002tsu.3	TUBA3D	Gap junction
uc010irs.2	NEIL3	Base excision repair
uc010bkn.2	NRG4	ERBB signaling pathway

Tumor-specific isoforms proved clinical relevance

Our analyses have revealed that the altered isoforms play important roles in the biological process of LUAD. Among these altered isoforms, there exists a group of isoforms that only expressed in tumor samples (see Methods). In particular, we identified 882 tumor-specific isoforms (Fig. 4(A)). Most of these tumor-specific isoforms were detected in 10–20% of all LUAD tumor samples ($n = 515$). Tumor-specific isoforms might only express in certain types of LUAD patient, which made them as potential prognosis marker. To further examine the clinical relevance of these isoforms, we evaluated the relations between overall survival time of patients and isoform expression in tumor samples. These analysis found that 96 tumor-specific isoforms be associated with patient overall survival (Log-rank test p value < 0.05, Fig. 4(B)). Expression level of some isoforms showed negative prognosis, whereas some showed positive prognosis. For example, higher expression level of uc001kuk.3 indicated shorter overall survival (Fig. 4(C)), whereas higher expression of uc003yls.2 indicated longer overall survival time (Fig. 4(D)). In addition, uc001kuk.3 (LDB1) was involved in pathways like transcriptional misregulation in cancer. And uc003yls.2 (RIMS2) played roles in ion channel binding. They both showed biological significance in cancer. Furthermore, we identified 6 tumor-specific isoforms were involved in tumor-related pathways (Table 1). These isoforms could be useful targets for pathway-inhibiting therapies. These observations showed that tumor-specific isoforms might be potential diagnosis or even prognosis markers.

Discussion

We have comprehensively characterized the expression of isoforms in LUAD samples. Our analysis found that a group

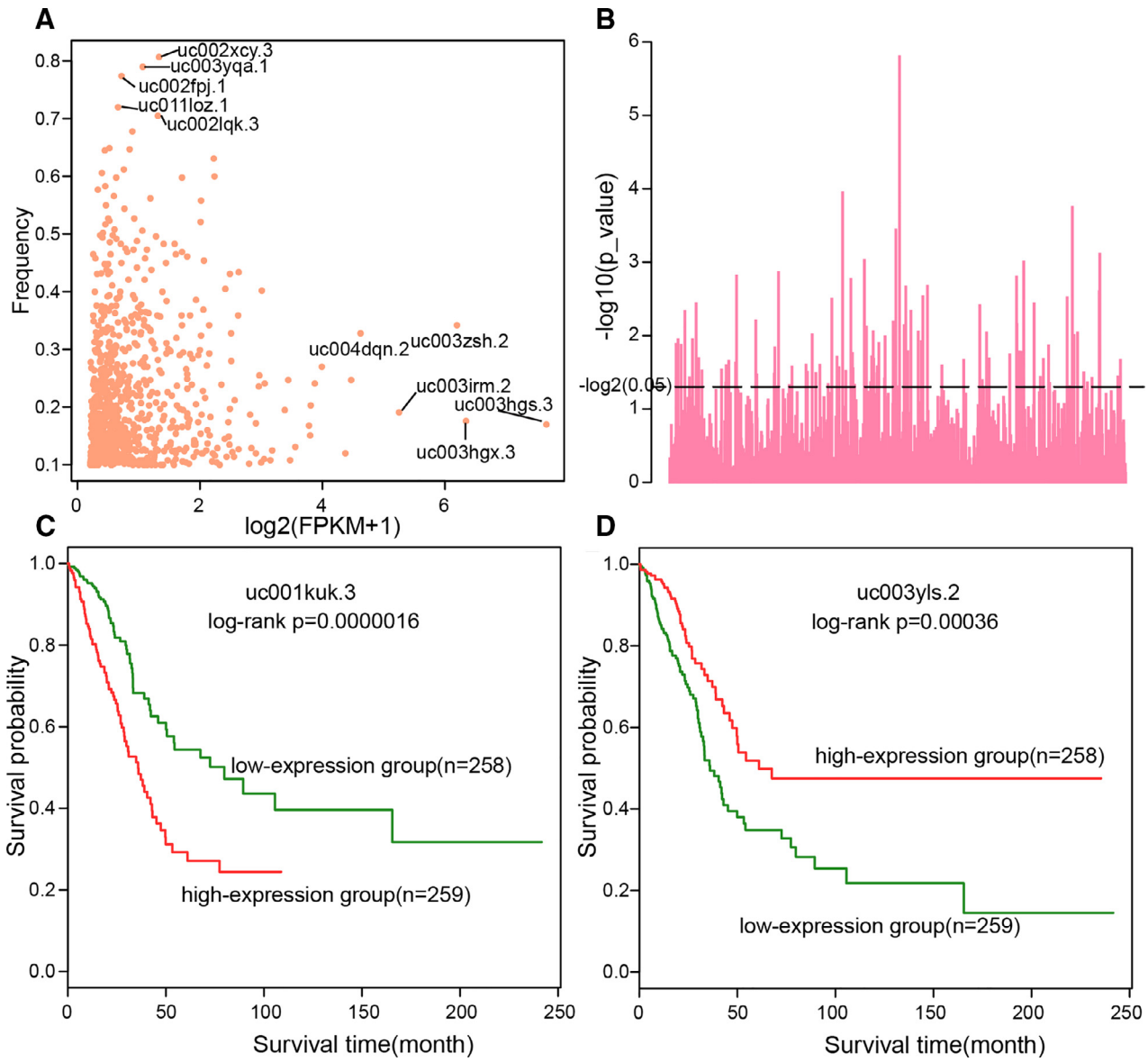

Fig. 4 Survival analysis of tumor-specific isoforms. (A) Scatter plot presents the expression level and frequency of tumor-specific isoforms. (B) The significance (p values) of relations between expression level of tumor-specific isoforms and patients overall survival time. (C) Survival curve of a negative prognosis isoform *uc001kuk.3*. (D) Survival curve of a positive prognosis isoform *uc003yls.2*.

of genes showed no expression variations but some of their isoforms expressed differentially. We next identified all differentially expressed isoforms in LUAD. Up-regulated and down-regulated isoforms showed distinct biological significance in tumor process. These observations indicated that expression analysis from isoform level could reveal the molecular alterations in LUAD more clearly. Isoform-level analysis in other types of tumor also revealed amazing observations [2,5]. Quantifying isoforms from raw sequencing reads with latest gene structure annotation would enrich these results. However, RNA-seq level 3 data of TCGA LUAD cohort was the largest dataset with clinical follow-up information we could access so far.

In eukaryotes, alternative splicing process leads to distinct mRNA isoforms transcribed from the primary transcription products [24]. Thus, the dis-regulation of alternative splicing in tumor might bring about differential expression levels of isoforms. The alternative splicing process could be affected by many factors, such as DNA methylation [25]. Further investigation on the mechanism of dis-regulated isoforms in tumor might lead to the discovery of drug targets.

Our further investigation found some hundreds of isoforms expressed exclusively in tumor. Even more, the expression level of several of these isoforms showed significant association with patient overall survival time. Therefore, our investigation provides some potential biomarkers for clinical diagnosis and therapy.

Funding

This work was supported by Supported by Foundation of Zhejiang Chinese Medical University (No.2014zz02) and the China Postdoctoral Science Foundation (No. 2015M581532).

Conflicts of interest

None.

References

[1] Adamia S, Haibe-Kains B, Pilarski PM, Bar-Natan M, Pevzner S, Avet-Loiseau H, Lode L, Verselis S, Fox EA, Burke J, Galinsky I, Dagogo-Jack I, Wadleigh M, Steensma DP, Motyckova G, Deangelo DJ, Quackenbush J, Stone R, Griffin JD. A genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets. Clin Cancer Res 2014;20(5):1135–45.

[2] Barrett CL, DeBoever C, Jepsen K, Saenz CC, Carson DA, Frazer KA. Systematic transcriptome analysis reveals tumor-specific isoforms for ovarian cancer diagnosis and therapy. Proc Natl Acad Sci U S A 2015;112(23):E3050–7.

[3] David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev 2010;24(21):2343–64.

[4] Sebestyen E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucl Acids Res 2015;43(3):1345–56.

[5] Zhao W, Hoadley KA, Parker JS, Perou CM. Identification of mRNA isoform switching in breast cancer. BMC Genom 2016;17:181.

[6] Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol 2013;14(7):R70.

[7] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10(1):57–63.

[8] Jakhesara SJ, Koringa PG, Bhatt VD, Shah TM, Vangipuram S, Shah S, Joshi CG. RNA-Seq reveals differentially expressed isoforms and novel splice variants in buccal mucosal cancer. Gene 2013;516(1):24–32.

[9] Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511(7511):543–550.

[10] Swanton C, Govindan R. Clinical implications of genomic discoveries in lung cancer. N Engl J Med 2016;374(19):1864–73.

[11] Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304(5676):1497–500.

[12] Takahashi T, Nau MM, Chiba I, Birrer MJ, Rosenberg RK, Vinocour M, Levitt M, Pass H, Gazdar AF, Minna JD. p53: a frequent target for genetic abnormalities in lung cancer. Science 1989;246(4929):491–4.

[13] Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou SH, Dezube BJ, Janne PA, Costa DB, Varella-Garcia M, Kim WH, Lynch TJ, Fidias P, Stubbs H, Engelman JA, Sequist LV, Tan W, Gandhi L, Mino-Kenudson M, Wei GC, Shreeve SM, Ratain MJ, Settleman J, Christensen JG, Haber DA, Wilner K, Salgia R, Shapiro GI, Clark JW, Iafrate AJ. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010;363(18):1693–703.

[14] Singh A, Misra V, Thimmulappa RK, Lee H, Ames S, Hoque MO, Herman JG, Baylin SB, Sidransky D, Gabrielson E, Brock MV, Biswal S. Dysfunctional KEAP1-NRF2 interaction in non-small–cell lung cancer. PLoS Med 2006;3(10):e420.

[15] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucl Acids Res 2010;38(18):e178.

[16] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf 2011;12:323.

[17] Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. Database issue. The UCSC genome browser database: 2014 update. Nucl Acids Res 2014;42:D764–70.

[18] Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genom 2014;15:419.

[19] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5(7):621–8.

[20] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–50.

[21] Dorner BG, Dorner MB, Zhou X, Opitz C, Mora A, Guttler S, Hutloff A, Mages HW, Ranke K, Schaefer M, Jack RS, Henn V, Kroczek RA. Selective expression of the chemokine receptor XCR1 on cross-presenting dendritic cells determines cooperation with CD8+ T cells. Immunity 2009;31(5):823–33.

[22] Bard-Chapeau EA, Nguyen AT, Rust AG, Sayadi A, Lee P, Chua BQ, New LS, de Jong J, Ward JM, Chin CK, Chew V, Toh HC, Abastado JP, Benoukraf T, Soong R, Bard FA, Dupuy AJ, Johnson RL, Radda GK, Chan EC, Wessels LF, Adams DJ, Jenkins NA, Copeland NG. Transposon mutagenesis identifies genes driving hepatocellular carcinoma in a chronic hepatitis B mouse model. Nat Genet 2014;46(1):24–32.

[23] Takeda H, Wei Z, Koso H, Rust AG, Yew CC, Mann MB, Ward JM, Adams DJ, Copeland NG, Jenkins NA. Transposon mutagenesis identifies genes and evolutionary forces driving gastrointestinal tract tumor progression. Nat Genet 2015;47(2):142–150.

[24] Faustino NA, Cooper TA. Pre-mRNA splicing and human disease.. Genes Dev 2003;17(4):419–37.

[25] Li S, Zhang J, Huang S, He X. Genome-wide analysis reveals that exon methylation facilitates its selective usage in the human transcriptome. Brief Bioinform 2018;19(5):754–64.