European Review for Medical and Pharmacological Sciences 2013; 17: 1716-1721

Computational identification of specific splicing regulatory elements from RNA-seq in lung cancer

R.-L. CHEN, W. GUO[1], Y. SHI[2], H. WU, J. WANG, G. SUN

Department of Respiratory Medicine, Shaanxi Provincial People's Hospital, Xi'an, Shaanxi, China
[1]Department of Emergency, Shaanxi Provincial People's Hospital, Xi'an, Shaanxi, China
[2]Department of West Yard No. 2, Shaanxi Provincial People's Hospital, Xi'an, Shaanxi, China

Co-first Authors: *Ruilin Chen, Wei Guo*

Abstract. – BACKGROUND: Lung cancer is the most common cause of cancer-related death worldwide. Recently, deep transcriptional sequencing has been used as an effective genomic assay to get an insight into this disease.

AIM: This study is carried out to identify specific regulatory elements (SREs) in lung cancer.

MATERIALS AND METHODS: The RNA-sequencing data on lung cancer sample and normal sample were downloaded from NCBI. TopHat and Cufflinks were used to analyze differential alternative splicing in lung cancer by using RNA-sequenceing data. Further, we searched specific SREs in lung cancer through finding over-represented hexamers around high expression exons.

RESULTS: According to the Jensen-Shannon divergence between two samples and the p-value of t-test, we found 53 genes with differential alternative splicing in lung cancer. In the analysis of SREs, we found 763 specific SREs between lung cancer sample and normal sample.

CONCLUSIONS: These results may give an insight into how alternative splicing causes differential expression in lung cancer.

Key Words:

Lung cancer, Splicing regulatory elements, RNA-sequencing data, High expression exons.

Introduction

Lung cancer is the most common cause of cancer-related death in men and women, which is responsible for more than 1.3 million deaths annually worldwide[1]. Epidemiology studies showed that smoking is the mean contributor to lung cancer, which causes 80-90% of lung cancers[2]. For nonsmokers, which accounting for 10-15% of lung cancer cases[3], the causes of disease are attributed to a combination of genetic factors, asbestos, radon gas, and air pollution[4-6].

Alternative splicing is considered as a process by which the exons of the RNA produced by transcription of a gene are joined in different ways during RNA splicing[7]. The resulting different mRNAs usually code for multiple proteins. Alternative splicing is a normal phenomenon in eukaryotes, which greatly increases the biodiversity of species. In addition, alternative splicing is expected to play a major role in gene expression regulation because of its capacity to generate protein diversity[8]. Genome-wide approaches have revealed that tumorigenesis often involves large-scale alterations in alternative splicing[9]. Many cancer-related genes are suggested to be regulated by splicing, involved in all major aspects of cancer cell biology, such as cell proliferation, differentiation, cell cycle control, and cell death[10]. RNA-Sequencing (RNA-seq) refers to the use of next-generation sequencing technologies to sequence cDNA in order to understand a sample's RNA content[11]. The technique has been rapidly adopted in studies of diseases like cancer. Through RNA-Seq analysis, we could get information about differential expression of genes, non-coding RNAs, gene fusions, and mRNA mutations or editing[12,13].

Splicing regulatory elements (SREs) are short nucleotide sequences which play a crucial role in regulating alternative splicing[14]. In this article, we used RNA-Seq sequences to identify candidate enhancers and silencers in lung cancer. According to the Jensen-Shannon (JS) divergence[15] between two samples and the *p*-value of *t*-test, we found 53 genes with differential alternative splicing in lung cancer. In the analysis of SREs, we found 498 specific SREs between lung cancer sample and normal sample. These results give an insight into how alternative splicing causes differential expression in lung cancer.

Corresponding Author: Gang Sun, MD; e-mail: xjtu1228@live.cn

Materials and Methods

Data Source

The RNA sequencing data on lung cancer sample and normal sample were downloaded from National Center for Biotechnology Information. The urls of downloading data are shown in Table I.

Alignment and Assembly of RNA-Seq Reads

In order to find differential alternative splicing, we first mapped RNA-seq reads for each library independently using TopHat version 2.0.6 against the human genome build hg19. Tophat is a fast splice junction mapper for RNA-Seq reads[16]. It aligns RNA-Seq reads to mammalian-sized genomes using the ultrahigh-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons[17].

Further, we used Cufflinks software for transcript assembly. Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples[18]. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Then Cufflinks was used to estimates the relative abundances of these transcripts with default parameters. After that, Cuffmerge was used to merge these transcripts of normal and tumor samples produced by the Cufflinks [19].

Differential alternative splicing in lung cancer

Cuffdiff[20] was used to identify differential alternative splicing related to lung cancer in our study. Cuffdiff not only learns the level of fragment count over-dispersion across replicates, it also uses the replicates to capture the fragment assignment uncertainty between alternative isoforms of each gene. We found the differential splicing by computing Jensen-Shannon (JS) divergence between two samples and the *p*-value of *t*-test[18]. Cuffdiff uses the square root of this divergence to measure the "distance" between the relative abundances of a group of transcripts in two difference conditions. Cuffdiff assigns *p*-values to the observed changes based on the JS divergence. The alternative isoforms were visualized by cummeRbund R package[21].

Identification of specific SREs in lung cancer

In order to analyze the reason why differential alternative splicing is occurred between two samples, we identified their specific SREs by calculating the frequencies of hexamers. The crucial theory used in all bioinformatics methods for identifying SREs is to find hexamers that are over-represented in a positive data set relative to a control data set[22,23]. We want to find hexamers that are over-represented in neighboring region of alternative splicing sites. Firstly, we divided alternatively spliced exons of each sample into two sets: inclusion set and exclusion set. If a majority ($\geq 90\%$) of the isoforms of a gene include one exon, then the exon is added to inclusion set. On the contrary, if a minority ($\leq 10\%$) of the isoforms of a gene include an exon, then the exon is added to exclusion set[24]. Furthermore, 400 intronic nucleotides upstream or downstream of inclusion/exclusion exons were added to inclusion/exclusion set.

In order to find candidate enhancers and silencers in lung cancer, we calculated the frequencies of each of 4^6 (4096) possible six-base combinations in both inclusion set and exclusion set. If frequencies of one hexamer possess significant differences between two sets ($p < 0.01$, two-tail test), the hexamer is considered as over-represented. The hexamers that are over-represented in one tissue but not over-represented in the other tissue are identified as putative specific SREs.

In statistics, z-score represents the distance between the raw score and the population mean in units of the standard deviation. Z-score is negative when the raw score is below the mean, positive when above. We set exclusion set as population when calculated z-score of hexamer from inclusion set. In this article, the z-score of the hexamer was calculated by mathematical formula which is shown in a previous paper[24]. We considered hexamers with z-score > 2.5758 ($p < 0.01$) as being enhancers, and hexamers with z-score < –2.5758 ($p < 0.01$) as being silencers.

Results

Differential alternative splicing

By computing the JS divergence between lung cancer sample and normal sample, We can find the significant differential alternative splicing with computing JS divergence between two samples. In this article, we found 53 genes with differential alternative splicing related to lung cancer with *p*-value < 0.001 (Table II). Here we show the alternative splicing pattern of one of these genes PIK3R5 in Figure 1. We could see that exon positions are usually conservative.

Table I. Detail information of samples.

Type	Data size	ID	Download address
Normal	7.4G base	ERR164475	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/ERR/ERR164/ERR164475/ERR164475.sra
Lung tumor	8.4G base	ERR164552	ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/ERR/ERR164/ERR164552/ERR164552.sra

Table II. The differential alternative splicing genes

Gene Symbol	Location	Sqrt(JS)	p-value
KIAA0430	chr16:15688225-15737023	0.385833	1.00E-05
KAT7	chr17:47865980-47906458	0.547462	1.00E-05
BZRAP1-AS1	chr17:56378587-56494931	0.832555	1.00E-05
PIK3R5	chr17:8782232-8869029	0.832467	1.00E-05
SYNRG	chr17:35874899-35969486	0.534531	1.00E-05
NCR1	chr19:55417507-55424439	0.832555	1.00E-05
E2F6	chr2:11584500-11606297	0.341957	0.000415
LMAN2L	chr2:97371666-97405813	0.471369	0.00012
PLA2G2A	chr1:20301923-20306932	0.832555	1.00E-05
ZNF619	chr3:40518603-40531728	0.701661	1.00E-05
VIPR1	chr3:42530790-42579065	0.715117	1.00E-05
RASSF6	chr4:74437266-74486348	0.675029	1.50E-0
LRRC39	chr1:100598705-100643829	0.832555	1.00E-05
PPP3CA	chr4:101944586-102268628	0.198444	1.00E-05
UBE2D3	chr4:103717132-103790032	0.832555	1.00E-05
TBCK	chr4:106967232-107270381	0.832555	1.00E-05
NEK1	chr4:170314420-170533778	0.39908	0.0005
TARS	chr5:33440801-33468196	0.601329	1.00E-05
PCDHGB6	chr5:140710251-140892548	0.832555	1.00E-05
PCDHGA10	chr5:140710251-140892548	0.832555	1.00E-05
GRIA1	chr5:152870083-153193429	0.832555	1.00E-05
ENC1	chr5:73923230-73937249	0.649359	1.00E-05
CLIC5	chr6:45866189-46048085	0.832555	1.00E-05
PARK2	chr6:161768589-163745505	0.832555	1.00E-05
IKZF1	chr7:50344377-50472798	0.788136	1.00E-05
DMTF1	chr7:86781676-86849031	0.644103	0.000565
CASP2	chr7:142985307-143004789	0.832555	1.00E-05
KIAA0895	chr7:36363758-36493400	0.591206	0.000525
ABHD11	chr7:73150424-73153190	0.447515	0.00042
RBBP5	chr1:205055269-205091150	0.736588	1.00E-05
ZNF484	chr9:95607312-95640320	0.832486	1.00E-05
PPP1R8	chr1:28157251-28178183	0.155521	8.00E-05
SLC9A6	chrX:135067582-135129428	0.626884	1.00E-05
CHRDL1	chrX:109917083-110039286	0.35926	1.00E-05
FXYD4	chr10:43867091-43871783	0.832555	1.00E-05
TCF7L2	chr10:114710008-114927436	0.657125	1.00E-05
NRP1	chr10:33466418-33623833	0.394176	7.50E-05
EGR2	chr10:64571755-64578927	0.394194	2.00E-05
LOC283050	chr10:80703082-80827205	0.832555	1.00E-05
TMEM25	chr11:118401802-118436791	0.80563	3.00E-05
MRVI1	chr11:10562782-10715535	0.832555	1.00E-05
PODN	chr1:53527723-53551174	0.832555	1.00E-05
HRASLS5	chr11:63014620-63330855	0.832555	1.00E-05
ERC1	chr12:1100403-1605099	0.60176	1.00E-05
LARP4	chr12:50794591-50873788	0.451246	1.00E-05
APAF1	chr12:99039077-100378432	0.832424	0.00078
RNF34	chr12:121837885-121862155	0.397679	1.00E-05
P2RX2	chr12:133195402-133198972	0.832555	1.00E-05
MTF2	chr1:93544791-93604638	0.761616	1.00E-05
SLC12A6	chr15:34517244-34630265	0.832555	1.00E-05
MEIS2	chr15:37183221-37393500	0.832555	1.00E-05
RUSC1	chr1:155278538-155300909	0.832555	1.00E-05
GPR56	chr16:57653909-57698944	0.832555	1.00E-05

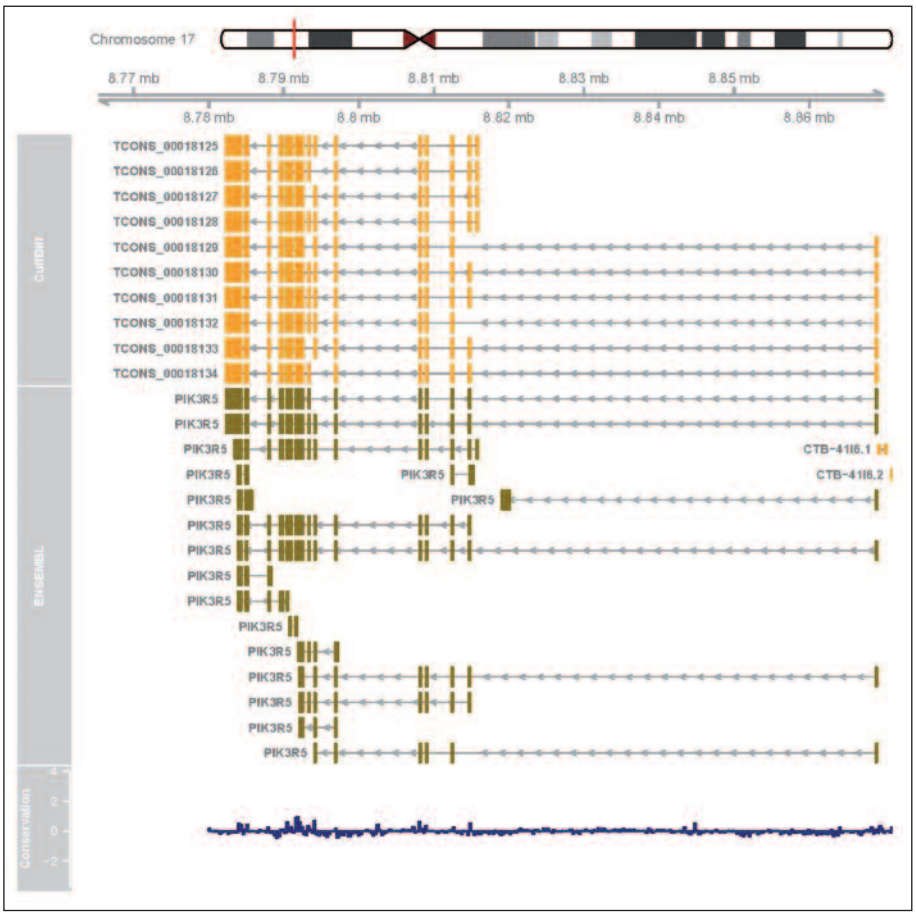

Figure 1. Alternative splicing of PIK3R5 in Cuffdiff results. Alternative splicing results in Cuffdiff are colored in yellow and their Cuffdiff IDs are labeled on the left. The brown part stands for all transcripts of PIK3R5 from ENSEMBL database. The blue histogram on the bottom is conservations in the chromosome region.

Specific SREs in lung cancer

In order to analyze the reason why differential alternative splicing is occurred between two samples, we further analyzed their specific SREs by calculating the frequencies of hexamers. We considered hexamers with z-score > 2.5758 ($p < 0.01$) as being enhancers, and hexamers with z-score < –2.5758 ($p < 0.01$) as being silencers. Finally, we identified 1346 candidate SREs in lung cancer sample and normal sample. Among them, 763 SREs are over-represented in only one of two samples (specific SREs) and 583 SREs are over-represented in both two samples. The distribution of z-scores of all hexamers from the two samples are shown in Figure 2. The SREs with most significant differences are listed in Table III.

Discussion

Abnormal splicing variants are usually implicated in disease, and cause a large proportion of human genetic disorders. Specifically alternative splicing is also thought to contribute to the development of cancer[25,26]. In this article, we found 53 genes with differential alternative splicing between lung cancer sample and normal sample.

Most of these genes were reported to be related to lung carcinogenesis previously. For example, Allen et al[27] presented that suppression of RASSF6 (Ras association domain family member 6) enhanced the tumorigenic phenotype of a human lung tumor cell line. Wen et al[28] reviewed the advancement of RASSF6 and suggested that RASSF6 expression was reduced in lung cancer, cervical cancer and might relate to tumorigenesis. Wu et al[29] demonstrated that NRP-1 (neuropilin 1) protein in lung cancer is related to angiogenesis. NRP1 plays a critical role in tumorigenesis, cancer invasion, and angiogenesis through VEGF, PI3K, and Akt pathways. Besides, Hong et al[30] suggested that NRP1 may have potential as a new therapeutic target in non-small cell lung cancer. These genes with differential alternative splicing in our result are potential markers for lung cancer diagnosis.

We also found out the reason why differential alternative splicing is occurred between two samples. Alternative splicing is thought to be regula-

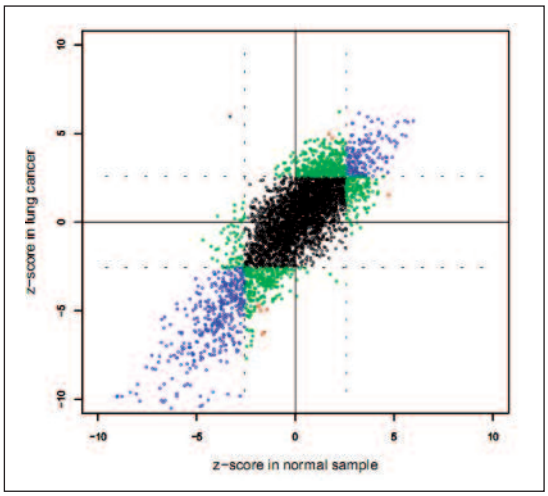

Figure 2. The z-scores distribution of 4096 hexamers in lung cancer sample and normal sample. The coordinates of dashed lines are z-score = –2.5758 and z-score = 2.5758 in the figure. The black points stand for hexamers which are not over-represented in both cancer and normal samples. The hexamers corlored in green are over-represented in only one of two samples. If the hexamers are over-represented in two samples, the points are clored in blue. What's more, the 9 SREs with most significant differences between two samples are colored in red. These SREs locate far from both blue and black area.

ted by tissue-specific enhancers and silencers[31]. In this article, we identified 1346 candidate SREs in two samples, but 763 SREs (56.69%) possess individual specificity. These specific-SREs present the exon preference when enhance or silence pre-mRNA splicing. As a result, leading to the differential alternative splicing between lung tumor cell and normal cell.

Transcripts with high hexamer ratio may play crucial role in the development of lung cancer. We further filtered the most significant transcripts in the 763 SREs. Nine transcripts with high significance were identified. These transcripts were reported to play important roles in the progression of lung cancer. For example, CXCL12 (chemokine ligand 12) encodes a stromal cell-derived alpha chemokine member of the intercrine family. This gene product and its receptor CXCR4 (chemokine receptor 4) can activate lymphocytes and have been implicated in the metastasis of some cancers including lung cancer. Imai et al[32] have indicated that CXCL12 is required for tumor growth and provide a rationale for the anti-CXCL12 treatment strategy in lung cancer. The results of a recent study have confirmed that CXCL12 play significant potential role in lung tumor cell migration[33].

Conclusions

We characterized the alternative splicing patterns of genes in lung cancer. A total of 53 differential alternative splicing genes which may served as the potential markers of lung carcinogenesis were identified. Besides, 763 SREs which may explain the reason of alternative splicing were identified by computational analysis. Further investigation of these alternative splicing genes and SREs may provide new thoughts for lung cancer study.

Conflict of Interest

None.

Table III. The SREs with most significant differences in two samples

SREs	Z-scores in Normal sample	Z-scores in Tumor sample	Transcript with highest hexamer ratio	Gene symbol
CTTATG	4.717323468	1.516202999	NM_003419	ZNF345
ATCAAG	1.695632415	4.972302415	NM_001207043	LOC554223
TCGCTC	-1.878906571	-4.750995592	NM_001171182	CENPL
AGTGCG	-1.852382375	-4.703257356	NM_001161625	NXNL2
GGTCGC	-1.452929527	-4.946891892	NM_001080950	MYO1C
CGGGTG	-1.572565134	-6.227095057	NR_034072	SARS
AAGCTT	1.871684165	4.735567851	NM_006724	MAP3K4
CGGCAG	-1.671979835	-6.341190217	NM_001077700	MIER1
CGCTCT	-1.756774747	-4.99952925	NM_000609	CXCL12

References

1) Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: Globocan 2008. Int J Cancer. 2010;127:2893-2917
2) Biesalski HK, Bueno de Mesquita B, Chesson A, Chytil F, Grimble R, Hermus RJ, Kohrle J, Lotan R, Norpoth K, Pastorino U, Thurnham D. European consensus statement on lung cancer: Risk factors and prevention. Lung cancer panel. CA Cancer J Clin. 1998;48:167-176; discussion 164-166
3) Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee IM, Marugame T, Palmer JR, Riboli E, Sobue T, Avila-Tang E, Wilkens LR, Samet JM. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. PLoS Med 2008; 5: e185.
4) O'Reilly KM, McLaughlin AM, Beckett WS, Sime PJ. Asbestos-related lung disease. Am Fam Physician 2007; 75: 683-688.
5) Chen H, Goldberg MS, Villeneuve PJ. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. Rev Environ Health 2008; 23: 243-297.
6) Alberg AJ, Ford JG, Samet JM; American College of Chest Physicians. Epidemiology of lung cancer: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132(3 Suppl): 29S-55S.
7) Black DL. Mechanisms of alternative pre-messenger rna splicing. Ann Rev Biochem 2003; 72: 291-336.
8) Ghigna C, Valacca C, Biamonti G. Alternative splicing and tumor progression. Curr Genomics 2008; 9: 556-570.
9) 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: 670-676.
10) Brinkman BM. Splice variants as cancer biomarkers. Clin Biochem 2004; 37: 584-594.
11) Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol 2010; Chapter 4: Unit 4.11.1-13.
12) Oshlack A, Robinson MD, Young MD. From RNA-Seq reads to differential expression results. Genome Biol 2010; 11: 220.
13) Cheng P, Cheng Y, Li Y, Zhao Z, Gao H, Li D, Li H, Zhang T. Comparison of the gene expression profiles between smokers with and without lung cancer using rna-seq. Asian Pac J Cancer Prev 2012; 13: 3605-3609.
14) Wang Z, Burge CB. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA 2008; 14: 802-813.
15) Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys Rev E 2002; 65: 041905.
16) Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with rna-seq. Bioinformatics 2009; 25: 1105-1111.
17) Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10: R25.
18) 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 Biotechnol 2010; 28: 511-515.
19) Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. Nat Protoc 2012; 7: 562-578.
20) Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-Seq. Nat Biotechnol 2012; 31: 46-53.
21) Goff LA. Repfpkm-methods 55. Package 'cummeRbund'. 2012: 55.
22) Zhang XHF, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. Genes Develop 2004; 18: 1241-1250.
23) Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science 2002; 297: 1007-1013.
24) Wen J, Chiba A, Cai X. Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. Nucleic Acids Res 2010; 38: 7895-7907.
25) Skotheim RI, Nees M. Alternative splicing in cancer: noise, functional, or systematic? Int J Biochem Cell Biol 2007; 39: 1432-1449.
26) Fackenthal JD, Godley LA. Aberrant RNA splicing and its functional consequences in cancer cells. Dis Model Mech 2008; 1: 37-42.
27) Allen NP, Donninger H, Vos MD, Eckfeld K, Hesson L, Gordon L, Birrer MJ, Latif F, Clark GJ. RASSF6 is a novel member of the RASSF family of tumor suppressors. Oncogene 2007; 26: 6203-6211.
28) Wen Y, Wang Q, Peng Z. Advancement of RASSF6 and malignant tumors. Progr Mod Biomed 2011; 13: 061.
29) Wu G, Guo S, Li G. Expression and significance of nrp-1 in xenografts of lung cancer cell in nude mouse. Chinese-German J Clin Oncol 2007; 6: 254-257.
30) Hong TM, Chen YL, Wu YY, Yuan A, Chao YC, Chung YC, Wu MH, Yang SC, Pan SH, Shih JY. Targeting neuropilin 1 as an antitumor strategy in lung cancer. Clin Cancer Res 2007; 13: 4759-4768.
31) Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nature Genet 2008; 40: 1416-1425.
32) Imai H, Sunaga N, Shimizu Y, Kakegawa S, Shimizu K, Sano T, Ishizuka T, Oyama T, Saito R, Minna J, Mori M. Clinicopathological and therapeutic significance of CXCL12 expression in lung cancer. Int J Immunopathol Pharmacol 2010; 23: 153.
33) Franco R, Pirozzi G, Scala S, Cantile M, Scognamiglio G, Camerlingo R, Botti G, Rocco G. CXCL12-binding receptors expression in non-small cell lung cancer relates to tumoral microvascular density and CXCR4 positive circulating tumoral cells in lung draining venous blood. Eur J Cardiothorac Surg 2012; 41: 368-375.