Front. Electr. Electron. Eng. China 2008, 3(3): 260–266
DOI 10.1007/s11460-008-0067-6

RESEARCH ARTICLE

Jianning BI, Tao PENG, Yanda LI

Evidence for association of multi-exon skipping events with tumors

Abstract Alternative splicing (AS) has been shown to be frequently present in human tumors. Specifically, it has been observed in some experimental studies that multi-exon skipping (MES) events often appear in tumorous tissues. Prompted by this observation, we conducted a genome-wide analysis of MES events to investigate their association with tumors. The results show that MES events are more likely associated with tumors than single-exon skipping (SES) and the degree of association increases with the number of skipped exons. Furthermore, MES events are found to be less conserved than their SES counterparts, which provides additional evidence for our results because disease-associated AS events should be eliminated during evolution. Interestingly, these differences still existed even after comparison of MES and SES events with similar-length skipped regions. These results demonstrate that MES events may be associated with tumors and suggest that MES isoforms might be useful in cancer diagnosis.

Keywords multi-exon skipping event, single-exon skipping event, tumor

1 Introduction

Alternative splicing (AS) is a wide-spread phenomenon in the expression of eukaryotic genes [1]. It plays such a crucial role in expanding proteome diversity that by using AS, a single gene can produce multiple distinct protein products [1].

The regulation of AS is a complex process involving many regulatory elements [1]. Defects in any regulatory element can possibly induce splicing alterations, which are often observed in human disease. It has once been estimated that about 15% of disease-causing mutations affect splicing [2]. In recent reports, the proportion has reached a striking 50% [3]. Specifically, some computational studies have identified many tumor-associated AS events [4–6]. Therefore, AS might be prevalent in human tumors.

The relationship between AS and human tumors has crucial clinical meanings. Although numerous efforts have been made in the diagnosis of cancers, there is still an urgent need to identify good biomarkers for clinical diagnostic purposes [7–9]. Microarrays have been shown to be a powerful tool to discover biomarkers for cancers [7]. However, most of these only measure the expression of overall mRNA transcripts of each gene and ignore the complexity of transcripts originating from AS [7–9]. Nevertheless, alternative splicing is frequently observed in tumors, and distinct splice variants may be associated with specific tumor subclasses [3,7,8]. Initial attempts to distinguish different splice variants have yielded some promising results. By exploiting a splicing array, Li et al. reported that it is more effective to use signature mRNA isoforms than overall transcripts in distinguishing between normal prostatic epithelia and prostate cancer [8]. In another microarray study, it was found that for as many as 30% of the cancer-relevant genes on the array, the ratios of distinct splice variants were different, while the overall gene expression levels remained unchanged [9]. The better performance of splicing microarrays than conventional ones in these studies demonstrates the usefulness of specific splice variants in diagnosis. Therefore, it will be helpful in cancer diagnosis if we could obtain more information about which splice variants are more likely associated with tumors.

Interestingly, several experimental studies have reported some differences regarding multi-exon skipping (MES) events and single-exon skipping (SES) events in terms of association with tumors. Here, MES events refer to exon skipping events in which at least two consecutive exons are removed, while SES events are those with only a single skipped exon. Human estrogen receptor α (hER) is a hormone-activated nuclear transcription factor that regulates the transcription of estrogen-responsive genes [10,11]. In normal breast tissue, it has been found that hER splice

Received April 3, 2008; accepted April 15, 2008

Jianning BI, Tao PENG, Yanda LI (✉)
MOE Key Laboratory of Bioinformatics and Bioinformatics Div, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: daulyd@tsinghua.edu.cn

variants are mainly composed of SES isoforms, while MES isoforms are rare. Nevertheless, in breast tumors, SES isoforms account for only half of all variants, and splice variants lacking multiple exons are abundant [12]. Another example is the RON receptor tyrosine kinase, which belongs to the MET proto-oncogene family. As a crucial factor for cancer development, RON has the ability to induce cell dissociation, migration and matrix invasion [13,14]. It has been found that in primary human colorectal adenocarcinomas, RON splice variants undergoing an MES event have much stronger oncogenic potential than that only having an SES event [14]. These phenomena have attracted our attention and encouraged an investigation of MES events.

As the most prevalent type of AS [15], exon skipping has been extensively studied. However, the majority of studies focus on SES events [16–18], while MES events receive little attention. Despite the aforementioned observations reported in some experimental studies, large-scale analysis of MES events is still lacking, while research in this field may provide some insights into the relationship between AS and tumors.

In this paper, we conducted a genome-wide analysis of MES events with the emphasis on their association with tumors. Exon skipping events, including MES and SES, were first identified through expressed sequence tag (EST) alignment. Next, we investigated the association of MES/SES events with tumors by evaluating their expression in normal and tumorous tissues. An example was then analyzed in detail to illustrate the difference between MES and SES. We also made a conservation analysis, which provided further evidence for our results. Finally, we investigated the origin of differences between MES and SES events by comparing MES and SES subsets whose skipped regions are of similar length distributions. The results show that MES events may be associated with tumors, thereby rendering them as potential diagnostic indicators.

2 Materials and methods

2.1 Data set generation

Figure 1 shows the pipeline of data set generation (Fig. 1). Human-mouse orthologous gene pairs were downloaded from HomoloGene (ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene). The structures and sequences of these orthologous genes were subsequently obtained based on human and mouse genomes. If the sequence of one gene could not be determined, the orthologous pair was eliminated. Those orthologous genes with equal numbers of exons were then selected. This requirement was set to facilitate conservation analysis. It should be noted that this criterion will not cause bias in our results, because most human and mouse orthologous genes have equal numbers of exons [19].

We then sought to identify exon skipping events occurring in human genes in two steps. First, we constructed exon-exon junctions for all possible exon skipping events, and second, we aligned these junctions with ESTs. In the first step, we constructed exon-exon junction sequences by concatenating the last 50 nt of the upstream exon with the first 50 nt of the downstream exon. If either exon was shorter than 50 nt, the sequence of whole exon was adopted.

In the second step, these exon-exon junction sequences were aligned with ESTs downloaded from NCBI (ftp://ftp.ncbi.nih.gov/blast/db). ESTs are sequence fragments of entire genes that are widely used in fields such as AS discovery, expression studies, and polymorphism analysis. Here, we employed ESTs to identify MES and SES events. The alignment of junctions with ESTs were made using BLAST with an E-value threshold of 1e–10. Furthermore, it was required that each alignment spanned at least the middle 50 nt of a junction ($\geqslant 25$ nt for each exon) and showed nucleotide identity of $\geqslant 95\%$, for which the threshold was set according to the estimation that ESTs have an error rate of $\sim 3\%$ [20]. Using these criteria, MES and SES events were identified. For convenience, we will refer to splice variants produced by MES/SES events as MES/SES isoforms. The number of ESTs corresponding to an MES/SES isoform was used as an estimate of expression level of this isoform. This feasible method was also employed in other works (e.g., Ref. [21]).

2.2 Calculation of proportion of MES/SES isoforms expressed only in tumors

The tissue origins of ESTs were obtained from UniLib (ftp://ftp.ncbi.nih.gov/repository/UniLib) annotations of libraries where ESTs were derived. We then determined whether an MES/SES isoform expressed in normal tissues or tumors based on the expression status of ESTs supporting this isoform.

When calculating proportions of MES/SES isoforms expressing only in tumors, we considered only isoforms supported by multiple ESTs, because it was unreliable to judge the expression status of those events supported only by a single EST.

2.3 Calculation of fold change

The fold change of an MES/SES isoform is defined as the isoform expression level in tumors divided by the expression level in normal tissues. In the case that normal expression level was zero, a policy was adopted which increased both normal and tumor expression levels by a small number. Here we added them to one, which is the smallest quantity. Just as mentioned

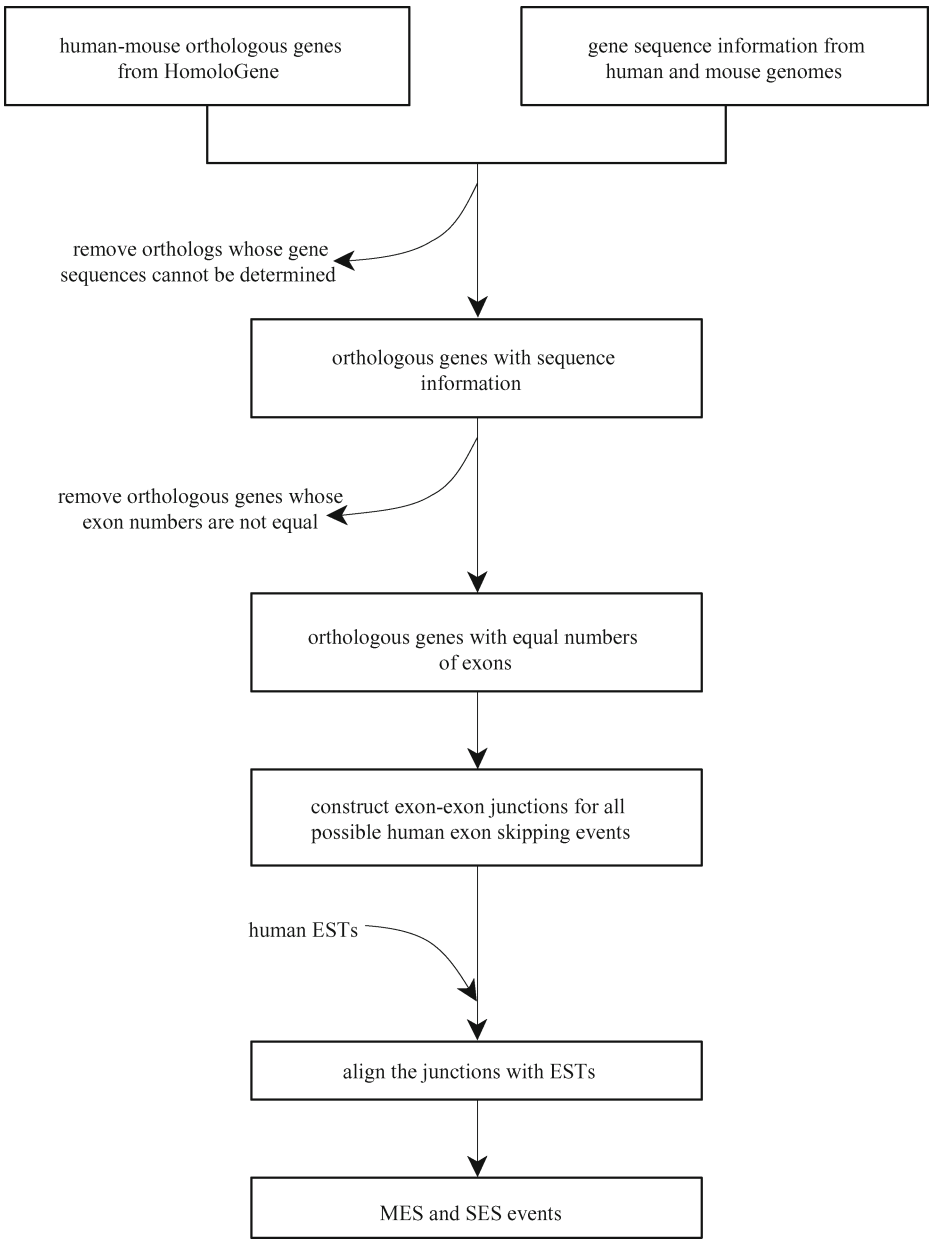

Fig. 1 Pipeline of data set generation

above, we considered only isoforms supported by multiple ESTs to improve the precision of calculated fold changes in the analysis.

2.4 Determination of conservation status of MES and SES events

To judge which events in human genes are conserved, we identified exon skipping events in mouse orthologous genes using the same method mentioned above. A normal MES/SES event in human was considered as conserved if it was observed in the same location in the mouse orthologous gene and also expressed in normal tissues.

2.5 Selection of MES and SES subsets with skipped regions of similar length distributions

Three steps were taken to generate subsets of events in which MES and SES removed regions of similar length distributions. First, we set a skipped region length range of 80–400 nt, which covers most MES and SES events. Second, this range was divided into many intervals, each with a length of 10 nt. Finally, for each interval, equal numbers of MES and SES events with skipped region lengths lying in this interval were selected. According to this procedure, MES and SES events in the subsets must have skipped regions of similar length distributions, and there must be equal numbers of MES and SES events in

the subsets. In this way, we obtained 660 MES and 660 SES events in human.

3 Results

3.1 Data set

Using EST alignment, we identified MES and SES events in the human genes (see Sect. 2), and MES events were further grouped by their skipped exon numbers. Figure 2(a) shows the counts of events skipping different numbers of exons. It can be observed that the count of events basically decreases as the skipped exon number increases. Moreover, there are more SES events than in any other groups, and exon skipping events removing large numbers of exons are rarely seen (Fig. 2(a)). These results are consistent with the observation that skipped region lengths of AS events follow the power law distribution and that small-sized events are predominant [22]. Based on these observations, we considered only exon skipping events which remove $\leqslant 12$ exons in this paper, because events skipping higher numbers of exons can be expected to be rare. Eventually, we obtained 1419 MES and 4539 SES events in the human genes, which constituted our data set.

3.2 Analysis of association of MES and SES events with tumors

To investigate the possible association of MES and SES events with tumors, we first examined MES and SES isoforms (i.e., splice variants produced by MES/SES events) that express only in tumors (see Sect. 2). These isoforms are probably associated with tumors, and the proportion of these tumor-specific isoforms in MES is 27.27%, higher than that in SES, which is 22.28%. A Fisher's exact test showed that this difference is significant (p-value = 0.017). Furthermore, MES events were classified according to their skipped exon numbers, and the proportion of isoforms expressing only in tumors was calculated for each group. Figure 2(b) shows the results for events with one to five skipped exons. It can be seen that the proportion of isoforms only in tumors generally becomes larger as the skipped exon number increases (Fig. 2(b)). Results of other groups with higher numbers of skipped exons are not shown, because events in these groups are quite few.

We also analyzed the fold change, which measures the increase of expression level in tumors relative to that in normal tissues (see Sect. 2). A fold change greater than one shows that the corresponding isoform expresses at a higher level in tumors than in normal tissues. The greater a fold change, the larger the increase of expression level in tumors. Figure 2(c) shows medians of fold changes of

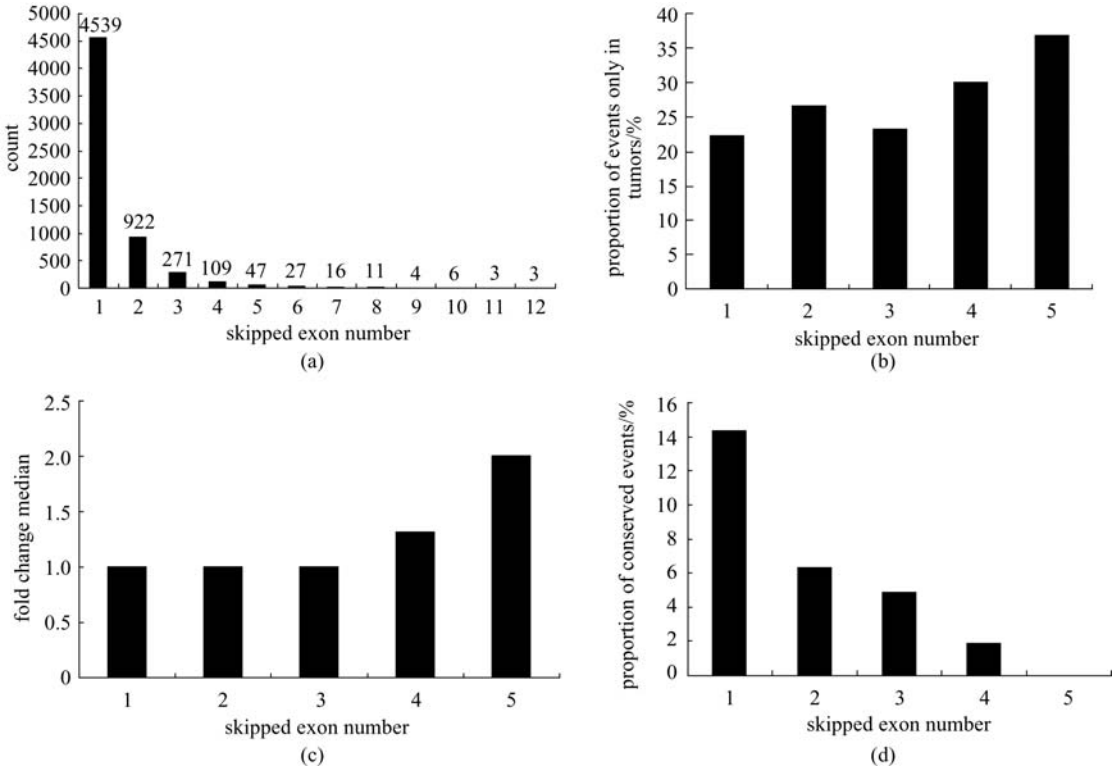

Fig. 2 Relations of some measures with skipped exon number. (a) Count of events; (b) proportion of events expressing only in tumors; (c) median of fold changes; (d) proportion of conserved events

exon skipping events removing different numbers of exons. Events skipping >5 exons were not included for their small quantities. There is an ascending trend of fold change as the number of skipped exons increases, indicating that the more skipped exons, the greater the increase of expression level in tumors.

3.3 ALG-2: a case study

An example was selected for detailed analysis to achieve an in-depth understanding of the association of MES events with tumors. Among genes involved in our data set, the human ALG-2 gene is a well-studied one. It has been found that ALG-2 has the ability of Ca^{2+} binding and is involved in multiple cellular processes such as apoptosis, proliferation and protein trafficking [23–25]. In our analysis, an MES event and an SES event have been observed in ALG-2 (Fig. 3). The SES event skips exon 4, while the MES event skips exon 3 and exon 4. The common skipped region, exon 4, is 159 nt in length, and the MES event removes an additional part, the 45 nt-long exon 3, whose length is only about one third of the common skipped region. Interestingly, it was observed that the MES isoform only expresses in tumors, while the SES isoform is present only in normal tissues.

To understand the association of the MES event with tumors, we investigated the impacts of the MES and SES events on the protein product. As both these events are located in the annotated CDS region and preserve the reading frame, the effect of either event on the annotated protein product comes in the form of simply removing a fragment of protein sequence. Based on the Conserved Domain Database (CDD v2.13 http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), three domains were predicted on the annotated protein sequence (Fig. 3): two EFh domains and one FRQ1 domain, all of which are related to Ca^{2+} binding. In Fig. 3, the thick black line represents the protein sequence of ALG-2, while the gray boxes represent the domains predicted by CDD. MES and SES events are also indicated. It was observed that for one EFh domain and the FRQ1 domain, the MES event removed longer parts from them than the SES event (Fig. 3). Particularly, for one EFh domain, the part deleted by the MES event was almost two-fold as long as that removed by the SES event (Fig. 3). Interestingly, EFh domains usually appear in pairs or high copy numbers, indicating that multiple EFh domains might cooperate to function properly and that destruction of any of these domains may abolish such function. In this situation, the MES event may influence the protein function more drastically than its SES counterpart, and this drastic effect might contribute to its association with tumors.

3.4 Analysis of conservation level

Gene products with functional defects are generally thought to be eliminated during evolution. Therefore, conservation analysis may provide additional evidence for the association of MES events with tumors.

Human normal MES and SES events conserved in mouse were first identified, and the proportion of these conserved events in the MES category is 5.51%, much lower than that in SES, which is 14.30%. A Fisher's exact test showed that the difference is significant (p-value $= 4.64e-13$). This result indicates that MES events are less conserved than SES events.

Following the same procedure above, we next subdivided MES events into different groups according to their skipped exon numbers and calculated conservation level in each group. The results, together with the conservation level of SES, are shown in Fig. 2(d) (results of groups skipping >5 exons are not shown), which shows that the conservation level decreases monotonously with the increasing number of skipped exons. Strikingly, the conservation level drops sharply ($>50\%$) from one-skipped-exon group to two-skipped-exon group (Fig. 2(d)).

It should be noted that our results could be affected by exon numbers of annotated genes. It is to be expected that the skipping of many exons in a gene with a small number of exons is rare. For example, MES events cannot occur in genes with just three annotated exons. Therefore, it is possible that the above results are simply caused by different annotated exon numbers between conserved and non-conserved events. To address this possibility, exon skipping events in MES and SES categories were further

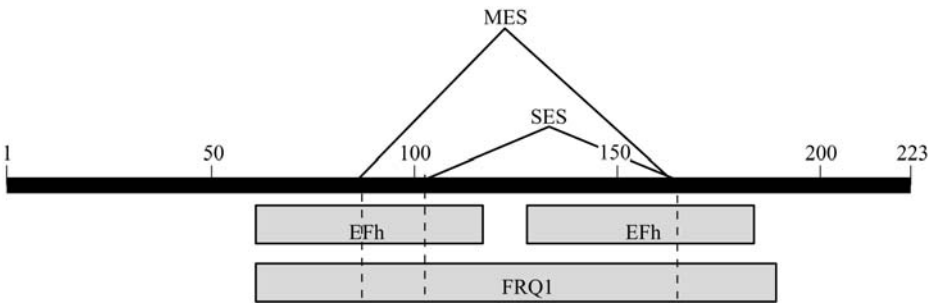

Fig. 3 MES and SES events of ALG-2

classified into several groups based on exon numbers of corresponding annotated genes. In each group, proportions of conserved events were calculated for MES and SES (Fig. 4). It was found that in each group, the conservation level of MES events was still much lower than that in the SES category (Fig. 4). Therefore, the annotated exon number bears little impact on our results.

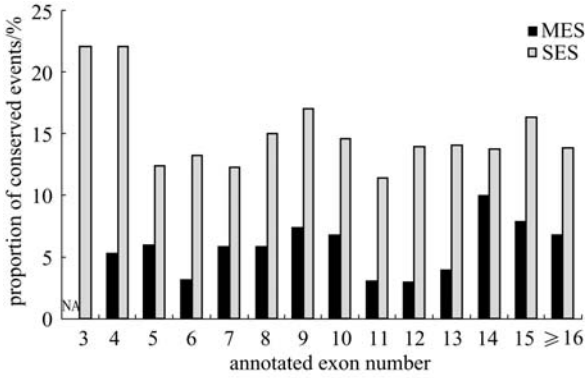

Fig. 4 Proportions of conserved events among MES and SES events grouped by annotated exon numbers

3.5 Comparison of MES and SES subsets with skipped regions of similar length distributions

After observing the above differences between MES and SES events, we sought to investigate their origin. Do these distinctions stem solely from the removal of longer regions by MES than by SES events? Or do they originate from some intrinsic differences between MES and SES events? To answer these questions, we selected subsets of MES and SES events with skipped regions of similar length distributions (see Sect. 2) and compared them in the aspects mentioned above.

It was observed that the MES subset still had a significantly larger fraction of events which expressed only in tumors than the SES subset (p-value = 0.029, Fisher's exact test). Furthermore, the analysis of another indicator of tumor association, fold change, showed that the median of MES was greater than that of SES, although the difference was not significant (MES median: 2, SES median: 1). We then examined conservation levels in the MES and SES subsets and found that the proportion of conserved events in the MES subset was 6.32%, significantly lower than the proportion in the SES subset, which was 11.63% (p-value = 0.012, Fisher's exact test). This result indicated that in the subsets, MES events were still less conserved than SES ones.

Overall, these findings show that the distinctions between MES and SES events still exist even if their difference in skipped region length vanishes.

4 Discussions

In this paper, we presented a genome-wide analysis of MES events and showed that MES events are different from SES events in terms of association with tumors and conservation level. It is to be noted that these differences remained even if we compared subsets of MES and SES events with skipped regions of similar length distributions.

Our results showed that MES events are more likely associated with tumors than SES events. Detailed analysis further revealed an increasing possibility of association with tumors as more exons are skipped. Consistent with our results, Hui et al. suggested that different splice variants might each possess distinct potential in carcinogenesis [4]. The association of MES events with human tumors obtained further evidence from the observation that MES events are much less conserved than their SES counterparts because tumor-associated AS events can be expected to be eliminated during evolution.

There might be two reasons explaining why MES events are more likely associated with tumors. First, compared with SES events, MES events generally delete much longer sequences, which may drastically influence the final protein products and might generate non-functional or even deleterious products. This explanation is supported by the aforementioned ALG-2 example as well as by a recent report which showed that AS events with mild changes are favored in evolution [26]. In fact, our data set is dominated by SES events, which is also consistent with a study showing that as much as 60% of AS events are less than 50 aa [22]. Second, compared with SES events, MES events may have a more complex regulatory mechanism, thereby increasing the likelihood of splicing mistakes. This can explain our finding that differences could still be observed even when MES and SES subsets with skipped regions of similar length distributions were compared. The complexity of the regulation of MES events may not only originate from the fact that they need to remove more exons, but may also stem from possible linkage of these skipped exons because they must all be removed in a single event. For example, in a hereditary non-polyposis colorectal cancer (HNPCC) family, affected family members were found to express hMLH1 mRNAs lacking exons six and seven, and further analysis suggested that the skipping of exon seven might be the prerequisite of skipping exon six [27].

Our results may also have some clinical meanings. We have shown that MES isoforms are possibly associated with human tumors. Meanwhile, studies based on splicing microarrays have demonstrated that individual mRNA isoform profiling can provide critical diagnostic information which cannot be captured by total transcript profiling [8,9]. Therefore, MES isoforms may possess great diagnostic potential and should receive more attention in bio-

marker identification. It should be noted that the development of diagnostic strategies employing specific MES isoforms does not require knowledge about whether these isoforms are causative factors for tumors or products of abnormal cellular activities; only the association between specific MES isoforms and tumors is sufficient to make these isoforms potential biomarkers [3,8].

Alternative splicing is often observed in human tumors, and research work in this field has both theoretical and clinical relevance. In this paper, we studied MES events that previously attracted little attention, and provided evidence for the possible association of these events with tumors. Our results shed light on the relationship between splicing and tumors and may be helpful in developing novel diagnostic approaches.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 60775002, 60572086).

References

1. Black D L. Mechanisms of alternative pre-messenger RNA splicing. Annual Review of Biochemistry, 2003, 72: 291–336
2. Krawczak M, Reiss J, Cooper D N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Human Genetics, 1992, 90(1–2): 41–54
3. Wang G S, Cooper T A. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature Review Genetics, 2007, 8(10): 749–761
4. Hui L, Zhang X, Wu X, et al. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. Oncogene, 2004, 23(17): 3013–3023
5. Xu Q, Lee C, Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. Nucleic Acids Research, 2003, 31(19): 5635–5643
6. Wang Z, Lo H S, Yang H, et al. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Research, 2003, 63(3): 655–657
7. Brinkman B M. Splice variants as cancer biomarkers. Clinical Biochemistry, 2004, 37(7): 584–594
8. Li H R, Wang-Rodriguez J, Nair T M, et al. Two-dimensional transcriptome profiling: identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. Cancer Research, 2006, 66(8): 4079–4088
9. Zhang C, Li H R, Fan J B, et al. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. BMC Bioinformatics, 2006, 7: 202
10. Wang Z, Zhang X, Shen P, et al. Identification, cloning, and expression of human estrogen receptor-alpha36, a novel variant of human estrogen receptor-alpha66. Biochemical and Biophysical Research Communications, 2005, 336(4): 1023–1027
11. Matthews J, Almlof T, Kietz S, et al. Estrogen receptor-alpha regulates SOCS-3 expression in human breast cancer cells. Biochemical and Biophysical Research Communications, 2005, 335(1): 168–174
12. Van Dijk M A, Hart A A, Van Veer L J. Differences in estrogen receptor alpha variant messenger RNAs between normal human breast tissue and primary breast carcinomas. Cancer Research, 2000, 60(3): 530–533
13. Cote M, Miller A D, Liu S L. Human RON receptor tyrosine kinase induces complete epithelial-to-mesenchymal transition but causes cellular senescence. Biochemical and Biophysical Research Communications, 2007, 360(1): 219–225
14. Zhou Y Q, He C, Chen Y Q, et al. Altered expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: generation of different splicing RON variants and their oncogenic potential. Oncogene, 2003, 22(2): 186–197
15. Ast G. How did alternative splicing evolve? Nature Review Genetics, 2004, 5(10): 773–782
16. Pan Q, Shai O, Misquitta C, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular Cell, 2004, 16(6): 929–941
17. Sorek R, Shamir R, Ast G. How prevalent is functional alternative splicing in the human genome? Trends in Genetics, 2004, 20(2): 68–71
18. Xing Y, Lee C J. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. PLoS Genetics, 2005, 1(3): e34
19. Thanaraj T A, Clark F, Muilu J, Conservation of human alternative splice events in mouse. Nucleic Acids Research, 2003, 31(10): 2544–2552
20. Boguski M S, Lowe T M, Tolstoshev C M. dbEST–database for ''expressed sequence tags''. Nature Genetics, 1993, 4(4): 332–333
21. Castillo-Davis C I, Mekhedov S L, Hartl D L, et al. Selection for short introns in highly expressed genes. Nature Genetics, 2002, 31(4): 415–418
22. Wang P, Yan B, Guo J T, et al. Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(52): 18920–18925
23. la Cour J M, Mollerup J, Berchtold M W. ALG-2 oscillates in sub-cellular localization, uni-temporally with calcium oscillations. Biochemical and Biophysical Research Communications, 2007, 353(4): 1063–1067
24. Shibata H, Suzuki H, Yoshida H, et al. ALG-2 directly binds Sec31A and localizes at endoplasmic reticulum exit sites in a Ca^{2+}-dependent manner. Biochemical and Biophysical Research Communications, 2007, 353(3): 756–763
25. Tarabykina S, Mollerup J, Winding P, et al. ALG-2, a multifunctional calcium binding protein. Frontiers in Bioscience, 2004, 9: 1817–1832
26. Zhang C, Krainer A R, Zhang M Q. Evolutionary impact of limited splicing fidelity in mammalian genes. Trends in Genetics, 2007, 23(10): 484–488
27. Tanko Q, Franklin B, Lynch H, et al. A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutation Research, 2002, 503(1–2): 37–42