

## Sequence analysis

**Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans**Xiang Shao<sup>1,†</sup>, Valery Shepelev<sup>2</sup> and Alexei Fedorov<sup>1,\*</sup><sup>1</sup>Department of Medicine and Program in Bioinformatics and Proteomics/Genomics, Medical University of Ohio, Toledo, OH 43614, USA and <sup>2</sup>Department of Bioinformatics, Institute of Molecular Genetics, RAS, Moscow 123182, Russia

Received on July 20, 2005; revised on November 3, 2005; accepted on November 18, 2005

Advance Access publication November 24, 2005

Associate Editor: Dmitrij Frishman

**ABSTRACT****Motivation:** Using bioinformatic approaches we aimed to characterize poorly understood abnormalities in splicing known as exon scrambling, exon repetition and trans-splicing.**Results:** We developed a software package that allows large-scale comparison of all human expressed sequence tags (EST) sequences to the entire set of human gene sequences. Among 5 992 495 EST sequences, 401 cases of exon repetition and 416 cases of exon scrambling were found. The vast majority of identified ESTs contain fragments rather than full-length repeated or scrambled exons. Their structures suggest that the scrambled or repeated exon fragments may have arisen in the process of cDNA cloning and not from splicing abnormalities. Nevertheless, we found 11 cases of full-length exon repetition showing that this phenomenon is real yet very rare. In searching for examples of trans-splicing, we looked only at reproducible events where at least two independent ESTs represent the same putative trans-splicing event. We found 15 ESTs representing five types of putative trans-splicing. However, all 15 cases were derived from human malignant tissues and could have resulted from genomic rearrangements. Our results provide support for a very rare but physiological occurrence of exon repetition, but suggest that apparent exon scrambling and trans-splicing result, respectively, from *in vitro* artifact and gene-level abnormalities.**Availability:** Exon–Intron Database (EID) is available at <http://www.meduohio.edu/bioinfo/eid>. Programs are available at <http://www.meduohio.edu/bioinfo/software.html>. The Laboratory website is available at <http://www.meduohio.edu/medicine/fedorov>**Contact:** [afedorov@meduohio.edu](mailto:afedorov@meduohio.edu)**Supplementary information:** Supplementary file is available at <http://www.meduohio.edu/bioinfo/software.html>**1 INTRODUCTION**

Excision of introns from pre-mRNA is a complex process in which several types of small nuclear RNAs and several dozens of proteins assemble into a spliceosomal complex (Maniatis and Reed, 2002). The precise recognition of exon–intron junctions by a spliceosome is crucial for the production of functional mRNAs. However, there is often ambiguity in the choice of exon–intron junctions. This

results in a process known as alternative splicing, which occurs in ~50% of mammalian genes and enables production of multiple mRNA isoforms from the same gene, often in a tissue-specific or development-stage-specific manner (Modrek and Lee, 2002; Stamm, 2002).

There are three other types of splicing-associated peculiarities reported to result in abnormal exon arrangement within mRNA molecules. The first of these is called exon repetition (ExRep) and was first characterized by Caudevilla *et al.* (1998). These studies showed that two copies of exon 2, from the rat carnitine octanoyltransferase gene, were positioned adjacent to one another in some mRNAs while the genomic sequence contained only a single copy. ExRep has been further described in other genes of vertebrates (Frantz *et al.*, 1999; Finta and Zaphiropoulos, 2000; Hide *et al.*, 2001; Rigatti *et al.*, 2004). A second form of abnormal exon arrangement is exon scrambling (ExScram) and was first described by Nigro *et al.* (1991), in which some products of the human tumor suppressor gene DCC had an aberrant order of exons. For example, in some DCC mRNA molecules they found that exon 4 preceded exon 2, while in other mRNAs exon 4 preceded exon 3. This exon arrangement could not be explained by alternative splicing, and thus, this process was identified as scrambling (perhaps resulting from circular mRNAs). Other cases of ExScram were described thereafter by others (Cocquerelle *et al.*, 1993; Zaphiropoulos, 1996, 1997; Caldas *et al.*, 1998; Crawford *et al.*, 1999; Takahara *et al.*, 2000; Flouriot *et al.*, 2002). Third is a process known as trans-splicing (TransSpl), in which transcripts from two different genes are joined at a splicing junction (Bonen, 1993). Trans-splicing is widespread in some protozoa (such as trypanosomes) and in nematodes, where a common 5' leader sequence on mRNAs is the result. TransSpl has also been described in *Drosophila* (Dorn *et al.*, 2001), rodents (Caudevilla *et al.*, 2001; Hirano and Noda, 2004) and human (Takahara *et al.*, 2000; Li *et al.*, 1999; Akopian *et al.*, 1999).

While ExRep and ExScram have been studied for a decade, no systematic statistical assessment has been done yet to estimate the prevalence of these phenomena, or to provide clues as to their physiological bases. TransSpl has been computationally analyzed by Romani *et al.* (2003) on a dataset of human mRNAs. Modern sequencing techniques have resulted in the characterization of hundreds of thousands of genes and millions of mRNA products. Such efforts have facilitated the investigation of biological processes *in silico* by computer examination of nucleotide sequence

\*To whom correspondence should be addressed.

†Current address: Zebrafish Information Network, University of Oregon, Eugene, OR 97403-5291, USA

databases. The most important source of data for computational analysis of splicing is the expressed sequence tag (EST) database (Boguski *et al.*, 1993). Recently, analyses of EST data demonstrated that at least 50% of human genes produce alternative isoforms (Modrek and Lee, 2002). The current release of the human EST database contains sequences of over six million mRNA fragments. Therefore, on average, each human gene is represented by more than 200 different ESTs. Here, we present a program package to study ExRep, ExScram and TransSpl *in silico* by comparison of the EST database with the database of their native gene sequences. We have applied our programs to examine six million human EST sequences, and report here 742 ESTs having sequences consistent with ExRep, ExScram or TransSpl. Our results suggest that a small subset of ExRep cases (2–3%) may have a physiological basis, while the other apparent examples of these three splicing variations may be artifactual.

## 2 MATERIALS AND METHODS

### 2.1 Databases

Human genomic sequences were obtained from GenBank, Build 35.1. After processing this database with the Exon-Intron Database (EID) toolkit (Saxonov *et al.*, 2000), a comprehensive collection of human gene, mRNA, and individual exon and intron sequences was obtained based on annotations in GenBank. This secondary database (human genome EID) is publicly available from our EID website <http://www.meduohio.edu/bioinfo/eid/> (file `hs35p1.EID.tar.gz`), and is accompanied with a description of the database (see README file at [http://www.meduohio.edu/bioinfo/eid/word/README\\_Sept05.DOC](http://www.meduohio.edu/bioinfo/eid/word/README_Sept05.DOC)). The advantages to using EID in our project include a convenient representation of exon/intron gene structures and availability of each gene sequence together with the related sequences of its mRNA, exons and introns that share a common EID identifier. The order of genes in EID strictly follows that of GenBank, and therefore corresponds to the physical order of genes in chromosomes. Several programs within the EID toolkit check the original input database and report possible errors and problems with primary sources.

Human EST sequences were obtained from GenBank, release 138 (Benson *et al.*, 1999). This dataset contains nearly 19 million EST sequences from numerous organisms (GenBank files `gbestN.seq`, where  $N$  is 1 to 279). An in-house Perl-script was used to select human ESTs (5987567 sequences) from the original dataset. All of these human ESTs were used in our analysis without additional filtering.

### 2.2 Description of sequence comparison algorithms

For fast comparison of millions of sequences on a workstation (Athlon, 2200+ CPU) with a large Random Access Memory capacity (4 GB RAM), we did not use standard alignment programs, but instead chose an alternative approach to direct matching of short sequence fragments. Each human mRNA sequence was divided into overlapping 28 nt long fragments (step size of 1 nt) that were stored in the very large associative array (Perl hash). The oligonucleotide sequences were the keys in this array and the mRNA identifiers were the values. Having all these oligonucleotides of all human mRNA sequences in the memory allowed us to immediately assign each EST sequence to its native mRNA sequence. For this purpose, each EST sequence was scanned with a 28 nt window using a step size of 1 nt and the associative array showed which gene(s) the current EST fragment belongs to. Often, poor quality of EST 5' and 3' ends make EST to gene correspondence difficult to establish. Since we used all possible 28 nt windows for each EST, the better quality EST regions compensated for this problem and made possible the EST to mRNA correspondence. After initial screening with this approach, the final subsets of EST candidates for ExRep, ExScram and TransSpl were aligned with their genes using BLAST. We

have successfully utilized this computational approach in several large-scale projects and it is our policy to make this software publicly available from our website.

### 2.3 Programs for computations of ExRep, ExScram and TransSpl

SCRAMBLING package represents a pipeline of programs written in PERL that are available from our web page <http://www.meduohio.edu/bioinfo/software.html>. The algorithms underlying these programs are described in Section 3. Among these programs `ESR1.pl` and `ESR2.pl` scripts perform large-scale comparisons of ESTs with all human mRNA sequences. Program `ESR3.pl` maps EST ends on the corresponding mRNA and calculates the lengths of the EST and the mRNA fragment. Programs `ESR4.pl` and `ESR5.pl` compare the EST and exon sequences for the calculation of exon representation score and exon order within ESTs, respectively. Program `ESR6.pl` searches intron sequences for the presence of possible cryptic exons. Finally, `TRANSSPLICING1.pl` and `TRANSSPLICING2.pl` screens ESTs for those in which the two ends correspond to different genes.

It took 5 days for a desktop computer (AMD Athlon, 2200+ processor) to run this program pipeline to get the final results.

## 3 RESULTS AND DISCUSSION

### 3.1 Stage I screening for ExScram and ExRep

In order to detect and to characterize exon scrambling (ExScram) and exon repetition (ExRep) events in ESTs, we developed a program package, SCRAMBLING, for comparison of all currently available human EST sequences (5992495) with the entire set of 20342 human intron-containing genes. SCRAMBLING begins computations from establishing a correspondence between each EST sequence and the gene it was produced from. As a result, 3221193 ESTs were linked unambiguously with their native genes. Next, we mapped the beginning and the end of each EST onto the corresponding mRNA sequence and calculated the length of the mRNA fragment between these two mapped positions. If the length of this mRNA fragment differed from the length of EST sequence by >40 nt, that EST was processed further to test for possible ExScram and ExRep events. The threshold of 40 nt is high enough to discard all small variations in length due to inaccuracy of EST sequencing as well as most of the cases of alternative splicing with alternative usage of nearby 5'- or 3'-splice sites. At the same time, the 40 nt threshold is relatively low in comparison with the average human exon length (125 nt). So the vast majority of known cases of ExScram and ExRep would exceed this threshold. When we obtained several mapping positions of EST termini onto corresponding mRNA (due to exon duplication or repetitive segments within mRNA), we considered all possible combinations between mapped EST beginnings and ends. There were a total of 182100 EST sequences that passed this initial selection (Table 1).

### 3.2 Stage II screening for ExScram and ExRep

The next computational procedure of SCRAMBLING pipeline was designed to distinguish possible ExScram and ExRep events from alternative splicing, which represent the vast majority of the selected 182100 ESTs.

In order to find putative ExRep, the candidate EST sequence was compared with exon sequences of the corresponding gene to generate a series of identity scores showing the proportion of each exon that was present in the EST. (If the entire exon was present the score was 1.0; if a half of exon was present the score was 0.5; etc.) For this

**Table 1.** Large-scale comparison of all human ESTs from GenBank with 20 342 human intron-containing protein-coding genes in order to find exon scrambling, exon repetition, and trans-splicing cases within ESTs

Stage of the project	Number of ESTs
Initial dataset	5 992 495
Stage I: matching EST onto 20 342 human genes	3 221 193
Computer search for exon scrambling and repetition	
Stage II: selection for exon scrambling, and repetition	182 100
Stage III: final set of exon repetition	401
final set of exon scrambling	416
Computer search for trans-splicing	
Stage I: ESTs representing two different gene fragments (putative trans-splicing)	253 838
Stage II: repetitive and convinced cases of trans-splicing ( $\geq 2$ independent ESTs represent same trans-spliced fragments from two genes)	15

purpose, the individual exon sequences were taken from the exon-only form of EID (file `hs35p1.exEID`), which is described in Section 2. During this procedure, putative ExRep cases were selected when the identity score for one of the exons was greater than 1.2.

In order to find putative ExScram events, the candidate EST sequence was compared with exon sequences of the corresponding gene to determine the order of exons along the EST. For example, if the program outputs (2, 3, 4, 5, 6) it means that the beginning of the EST is represented by the second exon followed by the third through sixth exons in order. When the program generates an abnormal order of exons (for instance, the order (1, 2, 4, 5, 6, 3, 4, 5) for EST sequence AU118844 shown in Fig. 3a) such cases were designated as putative ExScram events.

The EST candidates selected so far for ExRep and ExScram could have a special arrangement of their exonic sequences due to alternative splicing. For example, putative repetitive or scrambling exons could be alternatively spliced cryptic exons located within introns. Therefore, to exclude the possibility of alternative splicing among the selected ESTs, we compared them with all intronic sequences from their genes. For these computations, intron sequences were taken for the intron-only form of EID (dataset `hs35p1.intrEID`). When sequences of putative scrambling or repetitive exons were not detected inside introns, the relevant ESTs were collected into the final list of putative ExScram and ExRep occurrences. Finally, we selected 401 putative cases of ExRep and 416 putative cases of ExScram, as shown in Table 1. The number of ESTs containing both ExRep and ExScram events was 90. The accession numbers of ESTs from this list, as well as their tissue sources/libraries, are available from the supplementary file (`ES_ER_TS.doc`).

### 3.3 Stage III screening for ExScram and ExRep

Further characterization of the 401 putative ExRep and 416 putative ExScram candidates was performed by individually aligning the ESTs with their corresponding mRNA and genomic sequences, using BLAST alignment for two sequences (`bl2seq`) with the help of in-house Perl-scripts designed for automation of this examination. The majority of the 401 ExRep candidates represent cases when repetition occurs only with a portion of and not the whole exon. Thus, we refer to such cases as partial ExRep. Over half of the

partial ExRep cases had the repeated exonic sequence bordered by direct 3–12 nt repeats. A typical instance of partial ExRep is demonstrated in Figure 2, where short DNA repeats bordering the duplicated exonic segment are indicated by arrows. We detected repetition of an entire exonic sequence in only 18 cases, as described in Table 2. The first 11 ESTs from Table 2 represent convincing instances of ExRep and their structures are further illustrated in Figure 1. The remaining seven cases from the bottom of Table 2 represent artificial mRNA constructs that were obtained from the RAGE library (Harrington *et al.*, 2001), or resulted due to an exon amplification technique (Church *et al.*, 1994). Therefore, we do not consider these seven instances as real ExRep events. It is evident from Table 2 and Figure 1, that detected ExRep (1) do not have a strong tendency to be at a specific position inside the gene (both 5' and 3' end locations are possible) and (2) individual exons as well as exon pairs can undergo repetition. The described ESTs with whole-exon repetition were thoroughly investigated for alternative explanations of this phenomenon. For this purpose, we examined 500 000 nt long sequences from GenBank upstream and downstream from these eight genes in the presence of their duplicated copies or alternative promoters that could generate mRNA with an ExRep-like structure. All intron sequences inside these genes were carefully examined. We could find no evidence for alternative interpretations of the whole ExRep cases from Figure 1. However, we cannot rule out the possibility that the observed EST abnormalities with whole ExRep are due to errors in the human genome sequence assembly (Istrail *et al.*, 2004; She *et al.*, 2004). If we initially worked with a subset of erroneous human gene sequences, the abnormalities in their products would be an expected consequence. Another alternative explanation of observed ExRep could be allele-specific exon duplications that occurred in the tissues from which the ESTs were obtained. It is known that exon duplication is a very common genomic process that has produced ~10% of human exons (Fedorov *et al.*, 1998; Kondrashov and Koonin, 2001; Letunic *et al.*, 2002). Therefore, it is possible that some of these reported whole ExRep cases are due to unknown exon duplication events.

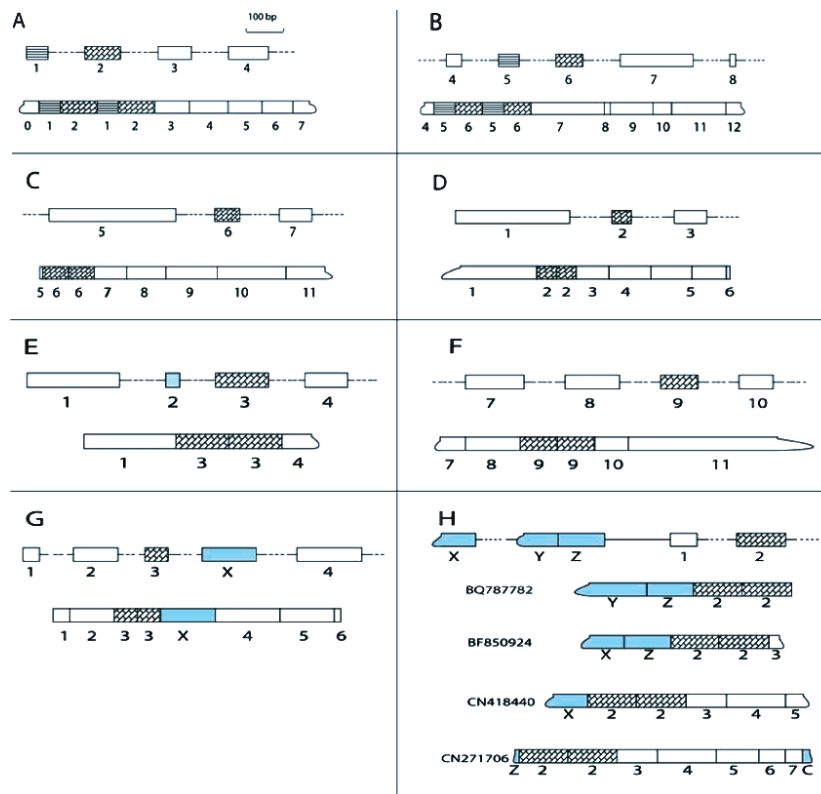
After investigation of the 416 EST candidates with putative ExScram we found no cases of scrambling that occurred with an entire exon but only with exonic pieces. In half of these ExScram cases we consistently saw short direct (3–12 nt) DNA repeats bordering the scrambling sequence (the same structures we detected for ExRep events). Two typical examples of such partial ExScram are demonstrated in Figures 2 and 3. We speculate that the observed partial ExRep and ExScram could occur during experimental procedures of EST cloning such as polymerase slippage during reverse transcription (Zhang *et al.*, 2001) or DNA rearrangements within some intrinsically unstable cDNA fragments. Because we could not detect even one example of entire exon scrambling among 3.2 million ESTs, it seems unlikely that exon scrambling occurs normally in human cells. Several cases of exon scrambling that have been previously reported in the literature may have occurred through other means such as experimental procedures during cDNA generation.

### 3.4 Computer search for trans-splicing

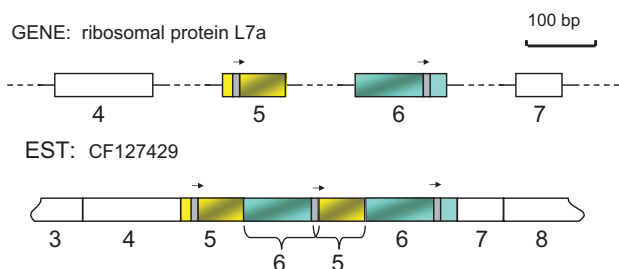
In order to find trans-splicing among 5 992 495 human ESTs, we employed `TRANSSPLICING1.pl` and `TRANSSPLICING1.pl`

**Table 2.** Summary of ESTs with whole-exon repetition

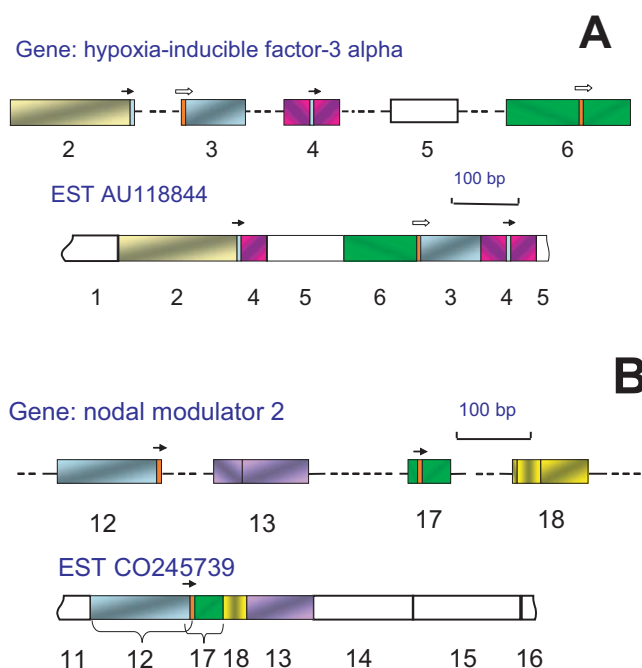
Fig 1	EST accession (GenBank)	EST length (nt)	Gene identifier in EID	Total # of exons	Repeat unit	Length of repetition unit	Note
1A	BG779951	800	14279_NT_010194	12	Exon 1+2	59 + 100	Additional exon 0
1B	AU125297	920	9423_NT_023935	20	Exon 5+6	63 + 75	
1C	CN267294	777	581_NT_004511	15	Exon 6	75	
1D	BI833847	738	6352_NT_034880	11	Exon 2	45	
1E	BP214078	565	18280_NT_011109	5	Exon 3	127	Skipped exon 2
1F	BQ940384	915	4556_NT_005612	11	Exon 9	90	
1G	CN263782	742	8800_NT_008183	12	Exon 3	66	Optional exon X
1H	BQ787782	579	14883_NT_010393	8	Exon 2	129	Altern. Exons X, Y, Z, 1
1H	BF850924	526	14883_NT_010393	8	Exon 2	129	Altern. Exons X, Y, Z, 1
1H	CH418440	660	14883_NT_010393	8	Exon 2	129	Altern. Exons X, Y, Z, 1
1H	CN271706	752	14883_NT_010393	8	Exon 2	129	Altern. Exons X, Y, Z, 1
	BG187500	449	5644_NT_006713	4	Exon 3	41	RAGE library
	BG189196	867	5356_NT_022792	10	Exon 4	224	RAGE library
	BG193343	751	5065_NT_016354	8	Exon 6	208	RAGE library
	H55089	284	19474_NT_011522	9	Exon 8	84	Exon amplification
	H55200	204	19464_NT_011521	13	Exon 11	102	Exon amplification
	T12575	221	9804_NT_035014	18	Exon 2	111	Exon amplification
	T12625	177	9835_NT_019501	4	Exon 3	79	Exon amplification



**Fig. 1.** Scheme of whole-exon repetition found in 11 ESTs. Top row represents structure of the gene that produced the ESTs. Exons are shown as boxes and are numbered; introns are shown as lines. Bottom row represents the structure of the EST. Repetitive exons are shadowed. Alternative exons are shown in gray and named with letters. (A) EST BG779951 representing annexin A2 gene from chr 15. ‘Exon 0’ is an additional exon representing 5’-UTR, which we found in the 5’ end of the gene and which was not described in the GenBank feature-table. (B) EST AU125297 representing transducin-like enhancer protein 1 gene from chr 9. (C) EST CN267294 representing forkhead box J3 gene on chr 1. (D) EST BI833847 representing GDP-mannose 4,6-dehydratase gene from chr 6. (E) EST BP214078 representing zinc finger protein 419 gene from chr 19. Exon 2 is skipped in the EST. (F) EST BQ940384 representing eukaryotic translation initiation factor 4A gene from chr 3. (G) EST representing staufen homolog 2 gene from chr 8. Optional exon X is present in the EST. (H) ESTs BQ787782, BF850924, CN418440 and CN271706 representing nuclear pore complex interacting protein gene from chr 16.



**Fig. 2.** Scheme of partial exon repetition found in EST sequence CF127429. Top row represents structure of the ribosomal protein L7a gene that produced this EST. Exons are shown as boxes and are numbered; introns are shown as lines. Bottom row represents the structure of the EST sequence. Exons 5 and 6, pieces of which are repeated within the EST, are shown in color. Short direct repeated sequence CACCAC inside exons 5 and 6 is shown as a gray bar with an arrow above it. This repeat lies precisely in the junction of partial exon repetition in the EST.

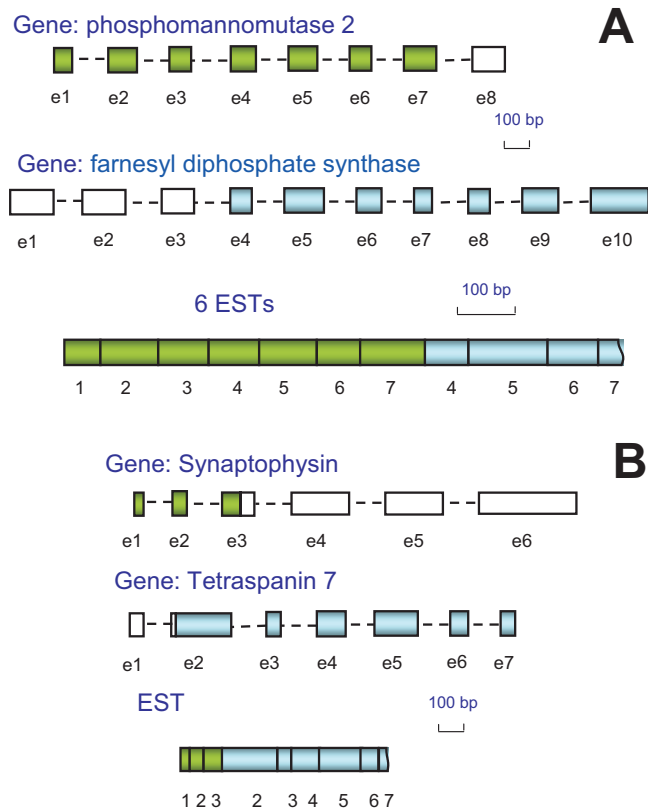


**Fig. 3.** Scheme of partial exon scrambling found in ESTs. Top row represents structure of the gene that produced the ESTs. Exons are shown as boxes and are numbered; introns are shown as lines. Bottom row represents the structure of the EST sequence. Scrambling exon pieces are shown in color. Short direct repeated sequences that border scrambled exon pieces are shown as gray or red bars with an arrow above them. (A) EST AU118844 representing Hypoxia-inducible factor 3 gene. Short 3 bp direct repeat inside exons 2 and 4 showed as black arrow is CAG. 7 bp direct repeat inside exons 3 and 6 showed as white arrow is GGGAGTG. (B) EST CO245739 representing Nodal modulator 2 gene. 12 bp direct repeat inside exons 12 and 17 showed as black arrow is TACAAAGTGCAG.

programs. In these computations we initially selected 253 838 ESTs in which the 5' and 3' ends corresponded to different human genes. During this selection we also required that two genes, having sequences present in the same EST, should not have adjacent

positions within the same chromosomal locus (in our algorithm they had to be separated by at least 30 other genes). This constraint was designed to avoid false-positive results from alternative distal promoters (alternative 5'-exons). Example cases of long transcription across neighboring genes are described by Romani *et al.* (2003) in their computational search for TransSpl in mRNA sequences in GenBank. Then, at the next cycle of selection, we required that at least two independent ESTs should represent the same putative trans-splicing event (ends of ESTs must be represented by the same pair of genes and also have the same junction site for putative TransSpl). The rationale for this filtering is to eliminate false-positive cases due to ligation of two random cDNA sequences or other experimental procedures that could occur in the process of EST cloning. As a result of this selection, we narrowed our sample to 2014 putative cases of TransSpl. Among these selected ESTs, many candidates for TransSpl were represented by gene pairs having very similar sequences. After excluding cases with the program that searches for 20 nt long identical fragments in gene pairs, we had 972 cases left. Further investigation of these 972 putative TransSpl events was done individually with the help of short Perl scripts to automate the screening process. A vast majority of these candidates for TransSpl appeared to be false positive due to DNA-repetitive elements inside mRNAs (Alu-repeats, mini-satellites, etc.). The removal of all EST candidates with repetitive elements resulted in the identification of 15 ESTs that represented five different cases of putative TransSpl (Table 1). Among the final set of 15 candidates, 6 independent ESTs (BM555536, BM809442, BM811153, BM913457, BM914863 and BG491331) were represented by the sequences of the phosphomannomutase-2 gene (NP\_000294) on chromosome 16 and the farnesyl diphosphate synthase gene (NP\_001995) on chromosome 1 (Fig. 4a). The 5' ends of these six ESTs are identical to the first seven exons of the phosphomannomutase-2 gene, while the 3' ends of the ESTs correspond to the 4–7 exons of the farnesyl diphosphate synthase gene. Remarkably, all these six ESTs were obtained in the same laboratory from the same tissue (amelanotic melanoma) according to the feature table of their GenBank records. Hence, there is a possible trivial explanation for this phenomenon other than TransSpl. Since genomic rearrangement frequently occurs in cancer tissues (Rabbitts, 1994), we propose that chromosomal recombination occurred *de novo* in the amelanotic melanoma cells which brought together the loci of the phosphomannomutase-2 and farnesyl diphosphate synthase genes. Therefore, the detected six ESTs likely represent the product of a novel human fusion gene.

The remaining four cases with putative TransSpl occurred between the following genes: (1) synaptophysin and tetraspanin 7 (EST: BI599645 and BI602014; Fig. 4b); (2) MHC DP alpha 1 precursor and myeloid/lymphoid or mixed-lineage leukemia 2 (EST: BF931318 and BF931405); (3) procollagen type II alpha 1 and GABA(A) receptor-associated protein (EST: BE813120 and BF326430); (4) GDP dissociation inhibitor 2 and ribosomal protein L3 (EST: BQ349883, BQ350058 and BQ350484). All four of these cases with putative TransSpl have the following characteristics as shown in Figure 4b: (1) ESTs were obtained by the same laboratory from the same cancer tissue; (2) junctions for the putative trans-splicing occur in the middle of exons. Therefore, none of these cases could be considered as real trans-splicing events beyond reasonable doubt. These data demonstrate that our computational procedures function well, yet TransSpl,



**Fig. 4.** Scheme of putative trans-splicing events found in ESTs. The first two rows represent the structure of the genes for which sequences were found within EST candidates for TransSpl. The exonic sequences presented in EST are shown in color, while exonic sequences that are absent in EST are shown as white boxes. Bottom row represents the structure of EST sequences. (A) ESTs (BM555536, BM809442, BM811153, BM913457, BM914863 and BG491331). (B) ESTs BI599645 and BI602014.

with even minimal reproducibility, is not detectable among 3.2 million ESTs.

Our results for bioinformatics investigation of TransSpl are consistent with the previous smaller-scale study on the same subject by Romani *et al.* (2003). Despite the fact that a percentage of publicly available EST and mRNA sequences have chimeric structures, the vast majority of these chimeres are the artifacts of the cDNA construction process. The occurrences of intriguing reproducible chimeric mRNAs represent either chromosomal translocations identified in cancer cells, or long transcription events across neighboring genes.

#### 4 CONCLUSIONS

We found that whole-exon repetition is an extremely rare process that was detected eleven times among 3.2 million human EST molecules from different tissues and cellular conditions. The whole-exon scrambling splicing event is even more elusive, if it exists at all (0 cases of whole ExScram among 3.2 million ESTs). When rearrangements occurred with exon pieces, ESTs with partial ExRep- and ExScram-like structures, were detected in 0.023% of the studied sequences. However, alternative explanations such

as cDNA rearrangements or reverse polymerase slippage could explain these sequencing abnormalities.

Reproducible cases of trans-splicing that occurred with at least two independent ESTs have not been detected among 3.2 million human mRNA molecules. All 15 ESTs with TransSpl-like structures were detected in cancer cells and probably appeared due to genomic rearrangements. For example, we found the origin of a novel human gene in an amelanotic melanoma by recombining the 5' end of the phosphomannomutase 2 gene and the 3' end of the farnesyl diphosphate synthase gene. Therefore, even though TransSpl may occur, it is extremely rare and does not play an important role in humans.

#### ACKNOWLEDGEMENTS

We would like to thank Dr Robert Blumenthal, Medical University of Ohio, for discussion and valuable suggestions on our manuscript. Support for this work was provided by the Medical University of Ohio Foundation and the Stranahan Foundation, through the Program in Bioinformatics and Proteomics/Genomics.

*Conflict of Interest:* none declared.

#### REFERENCES

- Akopian, A.N. *et al.* (1999) Trans-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Lett.*, **445**, 177–182.
- Benson, D.A. *et al.* (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
- Boguski, M.S. *et al.* (1993) dbEST—database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.
- Bonen, L. (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J.*, **7**, 40–46.
- Caldas, C. *et al.* (1998) Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene. *Gene*, **208**, 167–176.
- Caudevilla, C. *et al.* (2001) Localization of an exonic splicing enhancer responsible for mammalian natural trans-splicing. *Nucleic Acids Res.*, **29**, 3108–3115.
- Caudevilla, C. *et al.* (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNA in rat liver. *Proc. Natl Acad. Sci. USA*, **95**, 12185–12190.
- Church, D.M. *et al.* (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nat. Genet.*, **6**, 98–105.
- Cocquerelle, C. *et al.* (1993) Mis-splicing yields circular RNA molecules. *FASEB J.*, **7**, 155–160.
- Crawford, J. *et al.* (1999) The PISLLRE gene: structure, exon skipping, and exclusion as tumor in breast cancer. *Genomics*, **56**, 90–99.
- Dorn, R. *et al.* (2001) Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in *Drosophila*. *Proc. Natl Acad. Sci. USA*, **98**, 9724–9729.
- Fedorov, A. *et al.* (1998) Influence of exon duplication and shuffling on intron phase distribution. *J. Mol. Evol.*, **46**, 263–271.
- Finta, C. and Zaphiropoulos, P.G. (2000) The human CYP2C locus: a prototype for intergenic and exon repetition splicing events. *Genomics*, **63**, 433–438.
- Flouriou, G. *et al.* (2002) Natural trans-spliced mRNAs are generated from the human estrogen receptor- $\alpha$  (hER $\alpha$ ) gene. *J. Biol. Chem.*, **277**, 26244–26251.
- Frantz, S.A. *et al.* (1999) Exon repetition in mRNA. *Proc. Natl Acad. Sci. USA*, **96**, 5400–5405.
- Harrington, J.J. *et al.* (2001) Creation of genome-wide protein expression libraries using random activation of gene expression. *Nat. Biotechnol.*, **19**, 440–445.
- Hide, W.A. *et al.* (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res.*, **11**, 1848–1853.
- Hirano, M. and Noda, T. (2004) Genomic organization of the mouse Msh4 gene producing bicistronic, chimeric and antisense mRNA. *Gene*, **342**, 165–177.
- Istrail, S. *et al.* (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA*, **101**, 1916–1921.
- Kondrashov, F.A. and Koonin, E.V. (2001) Origin of alternative splicing by tandem exon duplication. *Hum. Mol. Genet.*, **10**, 2661–2669.
- Letunic, I. *et al.* (2002) Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.*, **11**, 1561–1567.

- Li,B.L. *et al.* (1999) Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes. *J. Biol. Chem.*, **274**, 11060–11071.
- Maniatis,T. and Reed,R. (2002) An extensive network of coupling among gene expression machines. *Nature*, **416**, 499–506.
- Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.
- Nigro,J.M. *et al.* (1991) Scrambled exons. *Cell*, **64**, 607–613.
- Rabbitts,T.H. (1994) Chromosomal translocations in human cancer. *Nature*, **372**, 143–149.
- Rigatti,R. *et al.* (2004) Exon repetition: a major pathway for processing mRNA of some genes is allele-specific. *Nucleic Acids Res.*, **32**, 441–446.
- Romani,A. *et al.* (2003) Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res.*, **31**, e17.
- Saxonov,S. *et al.* (2000) EID: The Exon–Intron Database: an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
- She,X. *et al.* (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, **431**, 927–930.
- Stamm,S. (2002) Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. *Hum. Mol. Genet.*, **11**, 2409–2416.
- Takahara,T. *et al.* (2000) Heterogeneous Sp1 mRNAs in human HepG2 cells include a product of homotypic trans-splicing. *J. Biol. Chem.*, **275**, 38067–38072.
- Zaphiropoulos,P.G. (1996) Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. *Proc. Natl Acad. Sci. USA*, **93**, 6536–6541.
- Zaphiropoulos,P.G. (1997) Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol. Cell. Biol.*, **17**, 2985–2993.
- Zhang,Y.J. *et al.* (2001) Reverse transcription slippage over the mRNA secondary structure of the LIP1 gene. *Biotechniques*, **31**, 1286–1290.