BRIEFINGS IN BIOINFORMATICS. VOL 11. NO 2. 198–209
Advance Access published on 2 December 2009

doi:10.1093/bib/bbp041

Detection of human interchromosomal *trans*-splicing in sequence databanks

Roberto Hirochi Herai and Michel E. Beleza Yamagishi**

Submitted: 14th June 2009; Received (in revised form): 8th August 2009

Abstract

Trans-splicing is a common phenomenon in nematodes and kinetoplastids, and it has also been reported in other organisms, including humans. Up to now, all *in silico* strategies to find evidence of *trans*-splicing in humans have required that the candidate sequences follow the consensus splicing site rules (spliceosome-mediated mechanism). However, this criterion is not supported by the best human experimental evidence, which, except in a single case, do not follow canonical splicing sites. Moreover, recent findings describe a novel alternative tRNA mediated *trans*-splicing mechanism, which prescinds the spliceosome machinery. In order to answer the question, 'Are there hybrid mRNAs in sequence databanks, whose characteristics resemble those of the best human experimental evidence?', we have developed a methodology that successfully identified 16 hybrid mRNAs which might be instances of interchromosomal *trans*-splicing. Each hybrid mRNA is formed by a *trans*-spliced region (TSR), which was successfully mapped either onto known genes or onto a human endogenous retrovirus (HERV-K) transcript which supports their transcription. The existence of these hybrid mRNAs indicates that *trans*-splicing may be more widespread than believed. Furthermore, non-canonical splice site patterns suggest that infrequent splicing sites may occur under special conditions, or that an alternative *trans*-splicing mechanism is involved. Finally, our candidates are supposedly from normal tissue, and a recent study has reported that *trans*-splicing may occur not only in malignant tissues, but in normal tissues as well. Our methodology can be applied to 5′-UTR, coding sequences and 3′-UTR in order to find new candidates for a posteriori experimental confirmation.

Keywords: *bioinformatics; inter-chromosomal trans-splicing; non-canonical splicing sites; tRNA-mediated trans-splicing; inverted repeats*

BACKGROUND

Trans-splicing is an unusual form of RNA splicing, where distinct pre-mRNA transcripts contribute to a single mRNA formation. Although less frequent and less well understood than *cis*-splicing, *trans*-splicing has gained momentum owing to its promising applications. For example, *trans*-splicing has been applied to correct genetic defects in some species without the unwanted side-effects of unregulated expression of the target genes [1], and it has been used in a variety of human health applications reported in the specialized literature [2–10].

In 1993, Bonen [11] proposed two *trans*-splicing categories: the *spliced leader* (SL), and the *discontinuous group II intron*. The first type is common in organisms such as nematodes [12] and kinetoplastids [13], where a short leader sequence is *trans*-spliced into an untranslated mRNA 5′ region (5′UTR). The other type is found in plants, algal chloroplasts and plant mitochondria, and involves the joining of independently transcribed coding sequences. Both categories follow consensus splicing site rules which resemble conventional spliceosome-mediated *cis*-splicing. In the last few years, evidence in other organisms has shown that, although a rare

Corresponding author. Roberto Hirochi Herai, Genetics and Molecular Biology Department, Biology Institute, State University of Campinas, 13083-862 Campinas, SP, Brazil. Tel: +55 19 32115844; Fax: +55 19 32115754; E-mail: rherai@cnptia.embrapa.br

Roberto Hirochi Herai is a doctoral student. His research interests include bioinformatics applied to sequence analysis, genomic database design and analysis, the study of repetitive sequences associated with biological phenomena and the *trans*-splicing mechanism.

Michel E. Beleza Yamagishi is a researcher at the Applied Bioinformatics Laboratory, Agriculture Informatics Subdivision, Brazilian Agricultural Research Corporation, Campinas, SP, Brazil. His research interests include optimization problems, bioinformatics, data mining and the role of repetitive sequences in animal and plant genomes.

*These authors contributed equally to this work.

phenomenon, *trans*-splicing may be more widespread than believed.

Trans-splicing has been reported in *Drosophila* by Labrador *et al*. [14] and by Horiuchi and Aigaki [15], who observed that the *trans*-splicing phenomenon occurred after independently transcribed pre-mRNAs formed double-stranded RNA through complementary sequences in the *mod(mdg4)* gene. A similar mechanism [16] has been described, where the directed repetitive sequences that flanked two distinct genes approximated their respective pre-mRNAs in order to accomplish a *trans*-splicing event. Robertson *et al*. [17] found the first, and, to the best of our knowledge, only instance of *trans*-splicing in mosquitoes involving an internal gene's exons from distinct chromosomal *loci*.

Trans-splicing was reported in mammalian cells for the first time in rat liver [18]. Rigatti *et al*. [19] analyzed two rat genes that have tandem repeated exons, and, as the tandem exons were not observed in the respective genomic region, the authors proposed *trans*-splicing as a possible explanation. Evidence of intrachromosomal *trans*-splicing has been reported in cattle [20]. It is worth noting that heterologous mRNA hybrids were generated by mRNA *trans*-splicing, or the transcription of long mRNA across neighboring loci. It is interesting that independent work [21] has revealed that, in fact, the proximity of different transcripts can facilitate the occurrence of *trans*-splicing.

The diversity of *trans*-splicing phenomena in mammals is frequently associated with cancer, an association that has been studied by Chen *et al*. [22], who analyzed the structure of abnormal *MYC* mRNA genes related to tumoral cells, and concluded that the anomalous transcripts are usually generated by *trans*-splicing. Hahn *et al*. [23], using an EST library, found several fused genes from chromosome rearrangement, and unexpectedly, identified some cases that might be evidence of *trans*-splicing in normal tissues.

Canonical splicing sites worthy of note have been observed in all the studies mentioned. Rather interestingly, however, all but one reported that human experimental evidence fails to follow canonical splicing sites [13–16,24]. This may be explained either by the occurrence of infrequent splicing sites or by the existence of non-spliceosome-mediated *trans*-splicing mechanisms. The first possibility is plausible because *trans*-splicing is a rare phenomenon, and it is not unreasonable to conceive of the occurrence of non-canonical splicing sites, as has happened elsewhere [25]. The second possibility is plausible because the hybrid mRNAs might be the result of an alternative RNA processing mechanism which does not use the spliceosome machinery, such as the recently reported tRNA-mediated *trans*-splicing [26,27]. In this new mechanism, the hybrid mRNA is generated by *trans*-splicing two complementary pre-tRNA halves joined to two different pre-mRNAs. The authors speculate that 'another feasible way to achieve *trans*-splicing of mRNAs by tRNA endonuclease is to exploit the vast repertoire of repetitive sequences present in eukaryotic organisms'. Remembering that Short Interspersed Nuclear Elements (SINE) are related to tRNA genes or other RNA Polymerase III-transcribed genes, and that *Alu* elements constitute the most abundant family of short repeats (SINE), *Alu* elements may play a major role in *trans*-splicing in humans and other primates.

The first two examples of human experimental evidence have the same hybrid mRNA pattern: part of their 5′UTR came from one chromosome and the remaining sequence from another. This *trans*-splicing pattern resembles that of the SL category, although the *trans*-spliced sequence is not usually 'short' and, as already mentioned, neither example follows the splicing site rules, which means that they can not be classified as SL category.

The first human interchromosomal *trans*-splicing evidence was reported back in 1997 [28]. A truncated isoform of the Ca^{2+}/calmodulin-dependent protein kinase II (*CaM kinase II*) expressed in the human islets of Langerhans was experimentally demonstrated to be a hybrid mRNA formed from one transcript belonging to chromosome 10 and another from chromosome 18. The second example of experimental evidence was the human cholesterol acyltransferase-1 (*ACAT-1*) hybrid mRNA [29]. Its 5′UTR was mapped onto chromosome 7, while the remaining sequence was mapped onto chromosome 1. Its interesting characteristic was the presence of an unusually long 5′UTR with 1396 base pairs (bp). Note that the average 5′UTR length ranges from 100 to 200 bp [30]. Just as in the preceding case, there were no canonical splicing sites. To rule out the possibility of this hybrid mRNA being a ligation artifact produced during cDNA synthesis *in vitro*, RT–PCR experiments were performed, and the results were consistent with the *trans*-splicing hypothesis.

Projects aiming to find new genes and study their expression profiles, or identify new genetic markers such as a single nucleotide polymorphism (SNP), have produced a huge number of publicly available human transcript sequences. The development of novel efficient bioinformatics algorithms has made possible *in silico* strategies to search for human *trans*-splicing evidence in extremely large databases. The *In Silico Trans-splicing Retrieval System*—ISTReS [31] was proposed in order to meet this challenge. Its methodology is quite simple. First, the cDNA database chosen was mapped onto the human genome using BLAST. Then, only those hits that satisfied some predefined criteria were retained. One of its most restrictive criteria was to impose consensus splicing site rules. This criterion is quite controversial, because it plainly excludes both examples of human experimental inter-chromosomal *trans*-splicing evidence mentioned above. Nevertheless, the ISTReS running over the NCBI RefSeq databank found 55 hybrid mRNAs, of which only 21 were from normal tissues. The remaining 34 hybrid mRNAs were from malignant tissues, and it is known that genomic rearrangement frequently occurs in malignant tissues [32]. This last explanation seems to be more reasonable than *trans*-splicing. Therefore, ISTReS found 21 hybrid mRNAs that might be real instances of human *trans*-splicing that follow consensus splicing site rules; however, ISTReS misses those instances that do not.

In 2006, using 5 992 495 human-expressed sequence tags (EST) and ad hoc bioinformatics algorithms, several examples of exon repetition, exon scrambling and human *trans*-splicing were reported [33]. However, in this work, all 15 *trans*-splicing instances were from cancerous tissue (amelanotic melanoma), which, as the authors recognize, weakens the *trans*-splicing hypothesis.

These bioinformatics initiatives have proved to be sound and can be successfully applied to search for additional human *trans*-splicing evidence. Their main challenges are the huge number of sequences and the rarity of the *trans*-splicing phenomenon that complicates the problem. For this reason, some filtering strategy is unavoidable. Nevertheless, the filtering criteria should incorporate what we have learned so far from the experimental evidence. There are three main lessons, which can be summarized as follows. First, the focus should be on interchromosomal *trans*-splicing, because the best human experimental evidence belongs to this category. Second, hybrid mRNA is usually formed by a 5′UTR (or part of it) from one chromosome and the remaining sequence from another. Third, the occurrence of non-canonical splicing sites means that rare splicing sites may occur by chance, or might be the result of a non-spliceosome-mediated *trans*-splicing mechanism. None of the human *trans*-splicing candidate sequences reported so far satisfy those criteria simultaneously.

This work aims to show how to search for hybrid mRNAs in publicly available databases, the characteristics of which resemble those of the best human experimental *trans*-splicing evidence. It is worth noting that the *trans*-splicing fusion point may occur in coding sequences (CDS) and 3′UTR, and our methodology is intended to deal with these cases as well.

MATERIALS AND METHODS

In the following paragraphs, we describe our *in silico* methodology for searching for human *trans*-splicing evidence that considers the three lessons mentioned above. Considering that the supposed novel interchromosomal *trans*-splicing mechanism is still under investigation, we do not claim that our methodology can identify all categories of interchromosomal *trans*-splicing, although it successfully finds some instances, the characteristics of which are similar to those reported in experiments.

Database

The choice of database is no easy task, as there are plenty of options available. Each database has its own characteristics; consequently, it is critical to state which characteristics are most relevant to our point of view. Definitely, the transcript structure information, i.e. 5′UTR–CDS–3′UTR, is fundamental. Therefore, we decided to use a curated full-length cDNA (FLcDNA) library instead of an EST library, the sequences of which are shorter and significantly more redundant.

In 2004, an international collaborative project [34] built an integrated database of human genes and transcripts, called the H-Invitational Database (H-InvDB) (http://www.h-invitational.jp), which fully satisfied our criteria. It has a unique set of high-quality cDNA clones with curated annotation. H-InvDB release 5.0, based on human genome build 36.3, has 187 156 FLcDNA sequences. However, only 137 315 sequences have complete information about 5′UTR, CDS and 3′UTR regions. Moreover, the malignant tissues impose some result interpretation difficulties. For this reason, we removed all the sequences with a tissue origin positively identified as comprising malignant cells. After this preprocessing filtering, we built our own database of 113 202 sequences.

Filtering strategy

Our filtering strategy aims to minimize computational effort by reducing the number of sequences. Nonetheless, to properly define the filtering criteria, it is necessary to have a clear idea of what we are looking for. What are the characteristics that distinguish the sequences that could constitute human *trans*-splicing evidence? We have already mentioned that, considering the experimental evidence, the sequences we are looking for are hybrid mRNAs. More specifically, their 5′UTRs, or parts of them, are mapped onto one chromosome, while their CDS and 3′UTR are found in a distinct chromosome. In other words, we are looking for a particular *trans*-splicing type: interchromosomal. We do not impose splicing site rules.

Trans-splicing candidates

Our methodology is detailed in Figure 1. Initially, the H-InvDB transcripts (FLcDNA) are filtered (HInvFilter) to create a FLcDNA_FILT fasta file containing only those sequences with complete 5′UTR, CDS and 3′UTR information, collected from tissue cells without malignant labels. Next, we created three different files FLcDNA_FILT_5UTR, FLcDNA_FILT_CDS and FLcDNA_FILT_3UTR for 5′UTR, CDS and 3′UTR sequences, respectively. Using BLAST, we mapped all the 5′UTR, CDS and 3′UTR sequences onto human genome build 36.3 (BLAST parameters are detailed in Figure 1 and double-checked using BLAT [35]). The result was saved in three XML alignment files. The XML file corresponding to 5′UTR was cross-matched (GeneAlignmentAnalyser) to the CDS and the 3′UTR XML files. We retained only those sequences where TSRs were located in only a single chromosome and nowhere else, in order to avoid ambiguities in interpretation. Only these hybrid mRNAs were stored in the TransSpl file.

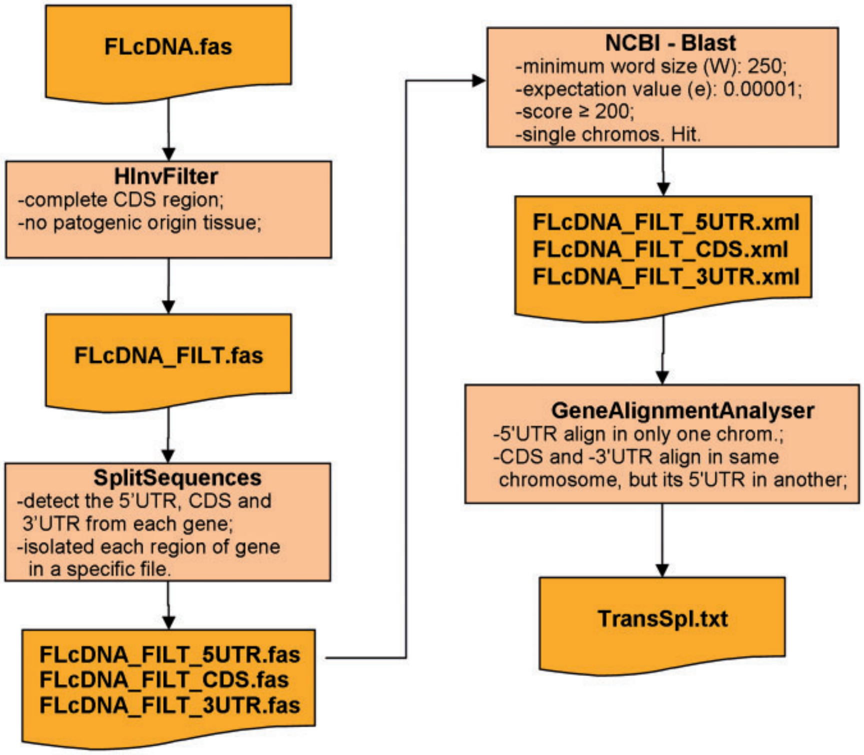
FLcDNA.fas
HInvFilter
-complete CDS region;
-no patogenic origin tissue;
FLcDNA_FILT.fas
SplitSequences
-detect the 5'UTR, CDS and 3'UTR from each gene;
-isolated each region of gene in a specific file.
FLcDNA_FILT_5UTR.fas
FLcDNA_FILT_CDS.fas
FLcDNA_FILT_3UTR.fas
NCBI - Blast
-minimum word size (W): 250;
-expectation value (e): 0.00001;
-score ≥ 200;
-single chromos. Hit.
FLcDNA_FILT_5UTR.xml
FLcDNA_FILT_CDS.xml
FLcDNA_FILT_3UTR.xml
GeneAlignmentAnalyser
-5'UTR align in only one chrom.;
-CDS and -3'UTR align in same chromosome, but its 5'UTR in another;
TransSpl.txt

Figure 1: Filtering and *trans*-splicing detection methodology fluxogram. There are four steps in the methodology, each of which is associated with an application: HInvFilter, SplitSequences, NCBI-BLAST and GeneAlignmentAnalyser.

RESULTS

We found 16 hybrid mRNAs (Table 1), the characteristics of which resemble those of the best human experimental *trans*-splicing evidence. As described in the Materials and Methods section, in order to avoid interpretational ambiguities, all the candidate hybrid mRNAs had their TSRs mapped onto only one single chromosome locus and nowhere else. Although the second case of experimental evidence suggested looking for transcripts with exceptionally long 5′UTR (like *ACAT-1* 5′UTR with 1.396 bp), our methodology found four candidate sequences with 5′UTR shorter than 400 bp.

Remarkably, there is no intersection between our candidate list and the ISTReS'. The cDNA database differences may be part of the explanation. There is, however, another reason that seems to be more significant: we did not impose canonical splice sites, and, interestingly enough, none of our candidate sequences follows them (Table 2). We speculate that this is related to the fact that our methodology was designed to look for a very specific *trans*-splicing category, i.e. interchromosomal *trans*-splicing, where the TSR belongs exclusively to 5′UTR. Perhaps this type of *trans*-splicing does not make use of canonical splicing sites or spliceosome machinery.

We double-checked the sequence quality with regard to external contamination, and found no vectorial, mitochondrial, or bacterial contamination. Unfortunately, although H-InvDB is a highly curated database, we cannot exclude the possibility that some hybrid mRNAs were produced by cloning or other experimental artifacts; however, four independent candidate sequences belong to the same H-InvDB cluster, which minimizes the probability of cloning or experimental artifacts.

We looked for evidence that the TSRs were actually transcribed. Using UCSC BLAT [35], each TSR was mapped back onto the human genome, and, using the UCSC Genome Browser, we identified annotated genes in the TSRs' genomic regions. Considering the 16 TSRs in Table 3, we arrived at two distinct cases: (i) 12 TSRs were mapped onto known gene loci (Figure 2 shows TSRs mapped onto *exons*, and Figure 3, TSRs mapped onto *introns*); (ii) four TSRs (actually, there is only one non-redundant TSR) were mapped onto a Human Endogenous Retrovirus (HERV-K) genomic region (Figure 4).

For example, the transcript with accession number [DDBJ:AK124366] is almost completely mapped onto chromosome 8. However, part of its 5′UTR came from chromosome 5, and it belongs to the 3′UTR of an *antisense* gene (PPP2CA), the accession number of which is [GenBank:NM_002715]. This sense/antisense transcript pattern resembles the *trans*-splicing mechanism mediated by repetitive sequences [16]. We developed ad hoc software to

Table 1: H-InvDB *trans*-splicing candidate transcripts

AC	TSR	5′UTR	CDS	3′UTR	mRNA	Tissue
[DDBJ:D26155] (*hsNF2a*)	[1–293]	297	4719	241	5257	Brain
[DDBJ:AL834489] (*DKFZp434F1431*)	[1–322]	324	1056	2367	3747	Testis
[DDBJ:AB023216] (*KIAA0999*)	[1–302]	437	3792	231	4460	Brain
[DDBJ:AK124366] (*FLJ42375 fis*)	[1–224]	302	255	2047	2604	Uterus
[DDBJ:AK226066] (*LAMP2*)	[1–342]	539	1236	2333	4108	Brain
[DDBJ:AF003522] (*Delta mRNA*)	[8–249]	322	2172	668	3162	N/A
[DDBJ:L33075] (*IQGAP1*)	[3–400]	467	4974	2132	7573	Placenta,liver
[DDBJ:AK130557] (*FLJ27047 fis*)	[1–578]	678	1065	1118	2861	Salivary gland
[DDBJ:L14837] (*zonula occludens ZO-1*)	[6–732]	1190	5247	1450	7887	N/A
[DDBJ:U09825] (*acid finger protein*)	[1–345]	555	1620	1420	3595	Kidney clone
[DDBJ:AB007865] (*KIAA0405*)	[1–987]	1124	1983	4420	7887	Brain
[DDBJ:AB020656] (*KIAA0849*)	[4–233]	446	2862	2106	5414	Brain
[DDBJ:AF458052] (*GRM7*)	[8–303]	451	2775	163	3389	[a]
[DDBJ:AF458053] (*GRM7*)	[8–303]	451	2736	135	3322	[a]
[DDBJ:AF458054] (*GRM7*)	[8–303]	451	2721	31	3203	[a]
[DDBJ:U92458] (*GRM7*)	[8–303]	451	2748	1113	4312	Fetal brain

[a] Brain, trachea, testis, uterus, salivary gland.

Complete: H-InvDB *trans*-splicing candidate transcripts. The 16 candidate sequences are listed with the following information: accession number (AC), TSR, 5′- and 3′-UTR, CDS, messenger ribonucleic acid transcript (mRNA) length in base pairs (bp) and tissue origin (N/A indicates that this information was not available).

Table 2: Splice sites from candidate sequences

AC	Locus	TSR Chr	TSR Sp	CDS-3′UTR Sp
[DDBJ:D26155]	9p24.3	6	**AA**-TSR-**TT**	**AG**-CDS-3′UTR
[DDBJ:AL834489]	5q35.2	12	**GG**-TSR-**AG**	**CG**-CDS-3′UTR
[DDBJ:AB023216]	11q23.3	2	**CG**-TSR-**GC**	**CT**-CDS-3′UTR
[DDBJ:AK124366]	8q24.12	5	**CT**-TSR-**CA**	**TG**-CDS-3′UTR
[DDBJ:AK226066]	Xq24	1	**GT**-TSR-**AT**	**CT**-CDS-3′UTR
[DDBJ:AF003522]	6q27	14	**TT**-TSR-**AG**	**TC**-CDS-3′UTR
[DDBJ:L33075]	15q26.1	4	**CT**-TSR-**TC**	**CA**-CDS-3′UTR
[DDBJ:AK130557]	4q13.3	19	**AT**-TSR-**TA**	**TT**-CDS-3′UTR
[DDBJ:L14837]	15q13.1	16	**TT**-TSR-**AC**	**GC**-CDS-3′UTR
[DDBJ:U09825]	6p21.33	1	**AA**-TSR-**AA**	**CA**-CDS-3′UTR
[DDBJ:AB007865]	14q31.3	4	**GA**-TSR-**TT**	**CC**-CDS-3′UTR
[DDBJ:AB020656]	16q12.1	11	**GG**-TSR-**AG**	**TC**-CDS-3′UTR
[DDBJ:AF458052]	3p26.1	19	**CA**-TSR-**TT**	**CT**-CDS-3′UTR
[DDBJ:AF458053]	3p26.1	19	**CA**-TSR-**TT**	**CT**-CDS-3′UTR
[DDBJ:AF458054]	3p26.1	19	**CA**-TSR-**TT**	**CT**-CDS-3′UTR
[DDBJ:U92458]	3p26.1	19	**CA**-TSR-**TT**	**CT**-CDS-3′UTR

Complete: Candidate sequence splicing sites. The first column contains the candidate sequence AC followed by its chromosomal map location (locus), *trans*-spliced chromosome region (TSR Chr), splicing site dinucleotides in bold letters (TSR Sp) and (CDS-3′UTR Sp), respectively.

Table 3: TSR associated transcripts

AC	AT
[DDBJ:D26155]	[GenBank:**NM.020823**] (*TMEM181*) [Exon: 1–293]
[DDBJ:AL834489]	[GenBank:**NM.175736**] (*FMNL3*) [Exon: 106–322]
[DDBJ:AB023216]	[GenBank:**NM.016552**] (*ANKMY1*) [Exon: 110–218]
[DDBJ:AK124366]	[GenBank:**NM.002715**] (*PPP2CA*) [Exon: 1–224]
[DDBJ:AK226066]	[GenBank:**NM.001821**] (*CHML*) [Exon: 1–342]
[DDBJ:AF003522]	[GenBank:**NM.021136**] (*RTN1-1*) [Exon: 1–242]
[DDBJ:L33075]	[GenBank:**BC032784**] (*CAMK2D*) [Intron]
[DDBJ:AK130557]	[GenBank:**BC136777**] (*ZNF700*) [Intron]
[DDBJ:L14837]	[GenBank:**AK124977**] (FLJ42987 fis) [Intron]
[DDBJ:U09825]	[GenBank:**NM.025106**] (*SPSB*) [Intron]
[DDBJ:AB007865]	[GenBank:**NM.147182**] (*KCNIP4*) [Intron]
[DDBJ:AB020656]	[GenBank:**BC030148**] (*ARFGAP2*) [Intron]
[DDBJ:AF458052]	[GenBank:Q9YNA8] (*HERV-K*) – 19q12
[DDBJ:AF458053]	[GenBank:Q9YNA8] (*HERV-K*) – 19q12
[DDBJ:AF458054]	[GenBank:Q9YNA8] (*HERV-K*) – 19q12
[DDBJ:U92458]	[GenBank:Q9YNA8] (*HERV-K*) – 19q12

Complete: TSR associated transcripts. When the TSR was mapped onto known genes (in bold), either it was mapped onto exons or onto introns (indicated in brackets). The corresponding gene codes are in parentheses.

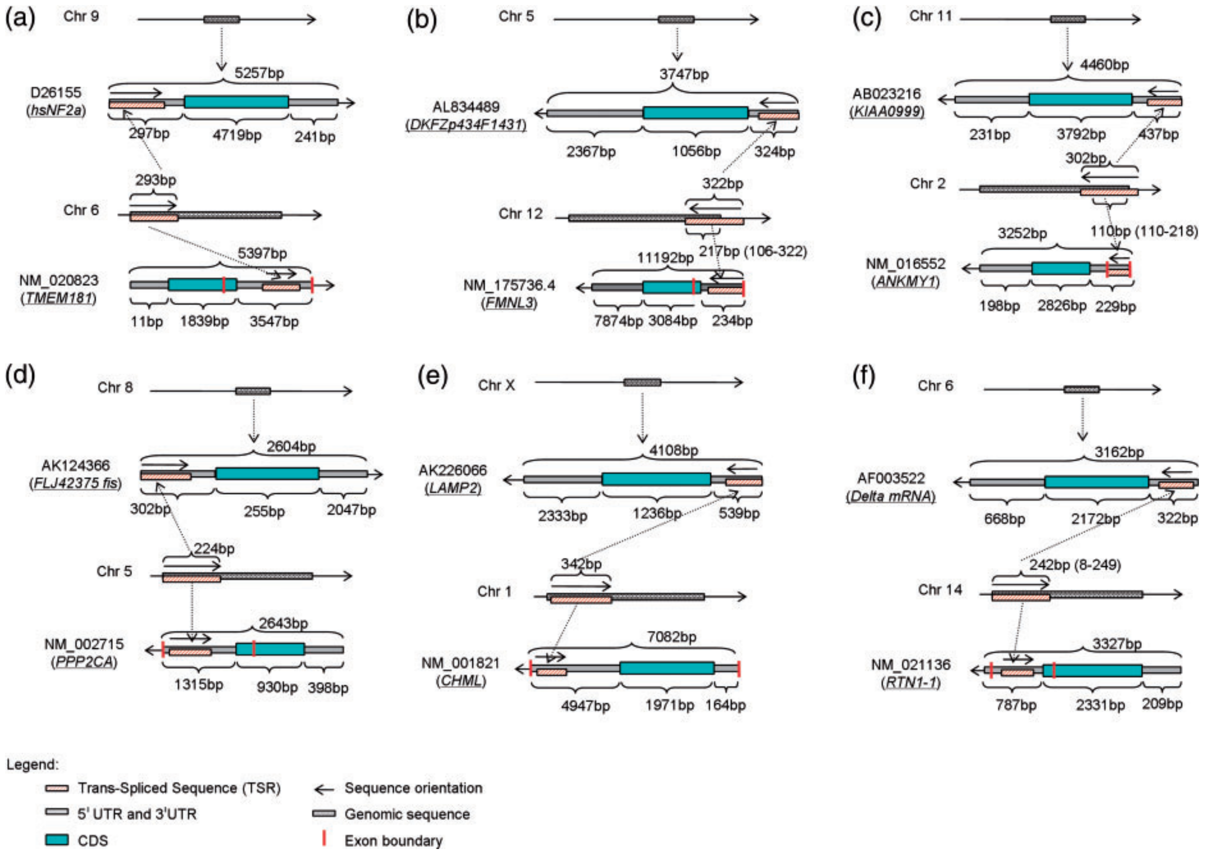

Figure 2: Gene structure (5′UTR-CDS-3′UTR, including the corresponding number of base pairs) of six *trans*-splicing candidates mapped onto exons from a different transcript. Each transcript from (a) to (f) has an annotated chromosome number, which is transcribed to generate one of our candidate transcripts formed by a TSR in its 5′UTR region. This region is transcribed by a different chromosome, within a single exon (belonging to a UTR) from another transcript. Arrows indicate the transcript and the chromosome orientation, in sense or antisense.

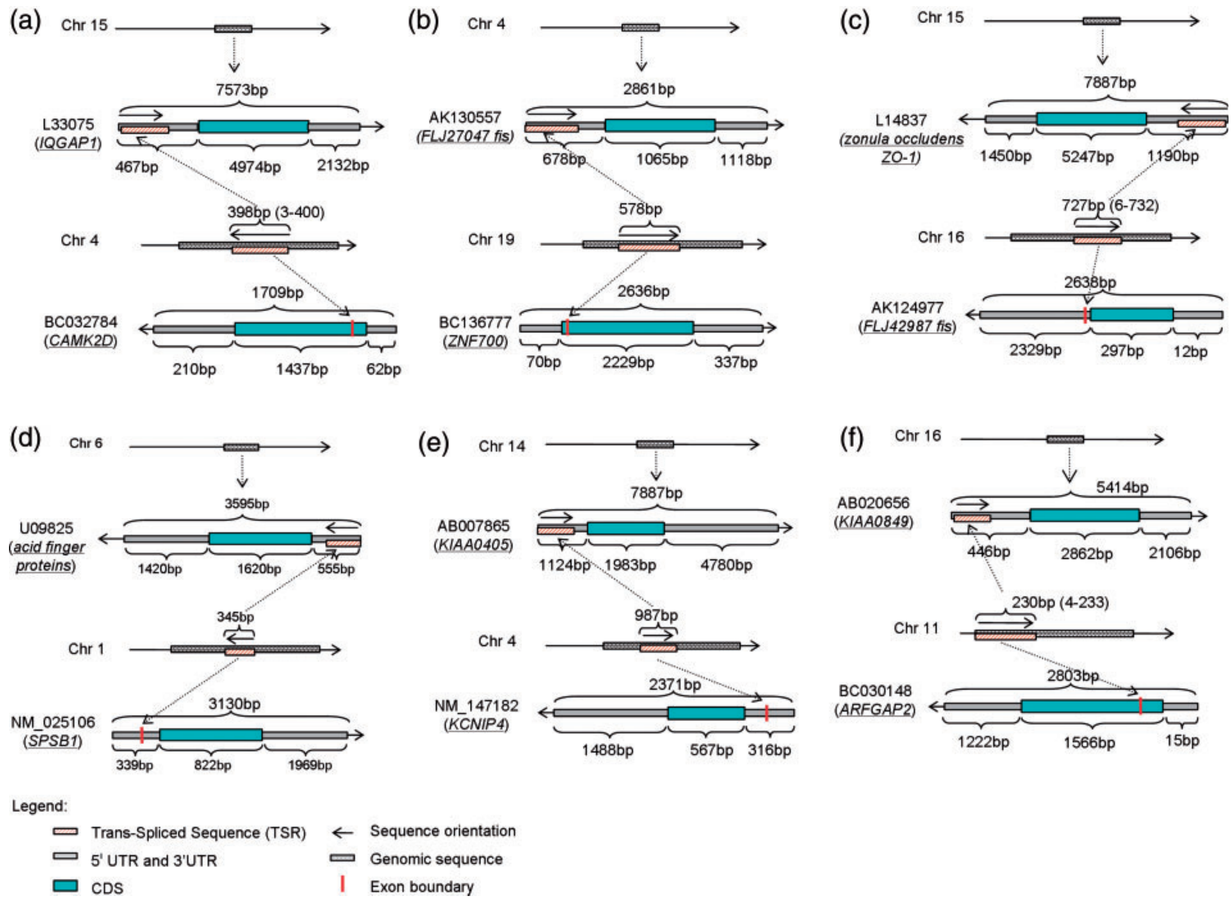

Figure 3: Gene structure (5′UTR-CDS-3′UTR, including the corresponding number of base pairs) of six *trans*-splicing candidates mapped onto the introns from a different transcript. Each transcript from (a) to (f) has an annotated chromosome, which is transcribed to generate one of our candidate transcripts that are formed by a TSR in its 5′UTR region. This region is transcribed by a different chromosome, within a single intron (in UTR or CDS) from another transcript. Arrows indicates the transcript and the chromosome orientation, in sense or antisense.

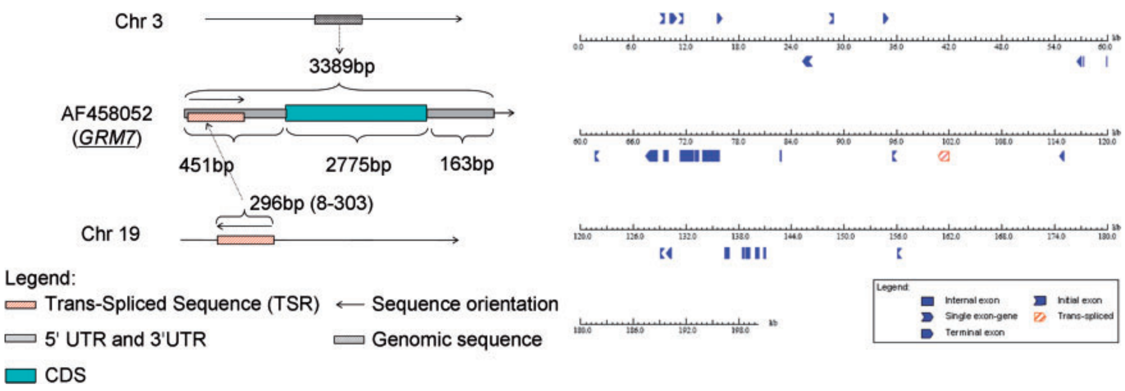

Figure 4: TSR from transcript [DDBJ:AF458052] mapped onto chromosome 19. On the left, the gene structure and its parts mapped onto chromosomes 3 (sense) and 19 (antisense). On the right, GenMark (genomic sequence analysis considering 100 kbp up-and-down stream of the TSR.) predicted genes which are highly similar to Human Endogenous RetroVirus (HERV-K).

look for short, inverted repeats that could mediate the *trans*-splicing event in both mRNA genomic regions. Although we found several almost perfect inverted repeats in all cases, in order to visualize some examples, we considered only those inverted repeats with at least 250 bp and with high sequence identity (Figure 5). We can see that, in some cases, there are so many inverted repeats that it is difficult to interpret the graphical representation. It is worth noting that most of the inverted repeats identified are actually *Alu* repeats.

The TSRs that were mapped onto a HERV-K genomic region are the most interesting and deserve special discussion. The UCSC Genome Browser indicated only LINE elements in the TSRs' locus. In order to confirm the presence of an active transcript, using BLAT, we mapped these TSRs onto the human genome, with the aim of obtaining expanded sequences (considering 100 kbp up- and down-stream of each mapped sequence). Applying validated gene prediction software, GeneMark [36], we found several putative amino acid sequences, some of which were actually similar to the retrovirus structure (*gag-pol-env*), in this case, the Human Endogenous Retrovirus K (*HERV-K*), which is an active human endogenous retrovirus active in the human organism [37]. We acknowledge that this is not strong evidence of transcription, because the information that HERV-K is an active mobile genetic element may suggest that the TSR is actually transcribed. However, as we are considering the reference genome only and HERV-K is a repetitive sequence, it is also possible that a new insertion has occurred which placed it into the existing gene, in which case, we would have no *trans*-splicing event at all. Nevertheless, as will be discussed below, two independent laboratories have indentified this hybrid mRNA. Thus, either a new insertion in the existing gene has occurred or a *trans*-splicing event has occurred, in both cases further investigation is required.

DISCUSSION

H-InvDB cluster HIX0019725

The sequences [DDBJ:AF458052], [DDBJ:AF458053], [DDBJ:AF458054] and [DDBJ:U92458] belong to the same H-InvDB cluster (number HIX0019725). This cluster has four more sequences that do not have *trans*-splicing characteristics. The four independent candidate sequences have the same *trans*-splicing pattern (Table 4), which is an unusual finding. Sequence [DDBJ:U92458] was obtained from a human fetal brain and was reported in 1998 by a Lilly Research Centre group in the USA [38], while sequences [DDBJ:AF458052], [DDBJ:AF458053] and [DDBJ:AF458054] were obtained from several human tissues, and were deposited in 2002 by the Human Genetics group at the University of Wuerzburg, Germany [39].

This cluster is associated with GPCR, family 3, a metabotropic glutamate receptor-like gene family. Their TSRs are mapped onto chromosome 19 (associated with *HERV-K*), while the remaining sequence is mapped onto chromosome 3. Their 5′UTR sequence has 451 bp, where the initial 303 bp form the TSR. However, only 296 bp from the TSR sequence were successfully mapped onto chromosome 19. The initial 7 bp were missing. Its TSR belongs to a well studied human endogenous retrovirus (*HERV-K*). These retrovirus' transcripts are found in every human tissue and they are still active. Unfortunately, the fact that HERV-K is still an active mobile element may jeopardize the hypothesis of *trans*-splicing event, because it is equally possible that a new insertion has occurred which placed the HERV-K into the GPCR genomic region in chromosome 3. Further investigation is required in order to decide which phenomenon has actually occurred.

Assuming that no new HERV-K insertion has occurred and considering the rarity of the *trans*--splicing phenomenon, the occurrence of four instances of candidate sequences in the same cluster is surprising for two reasons: first, because the sequences were deposited by two independent groups, implying that this phenomenon is recurrent and specific; second, considering the full length cDNA construction library methodology, the probability of laboratorial artifacts is reduced considerably.

A representative case: *FLRT2*

There is a single candidate sequence, accession number [DDBJ:AB007865], that most resembles the *ACAT-1 trans*-splicing pattern [29]. It is a human fibronectin leucine-rich transmembrane protein 2 (*FLRT2*) [40], the transcript structure of which is depicted in Figure 6. This transcript belongs to the H-InvDB Cluster HIX0011865, which contains six FLcDNA sequences of the *FLRT2* gene, but only sequence HIT00000122 (AC [DDBJ:AB007865])

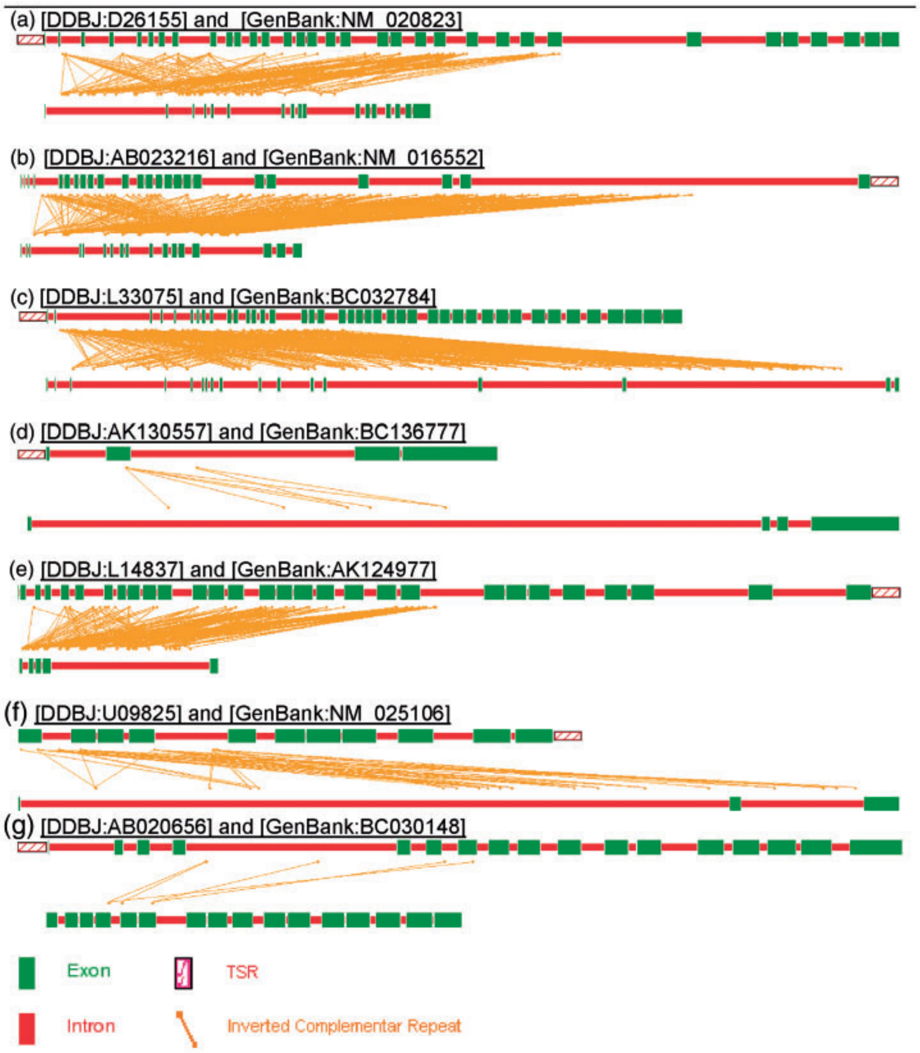
(a) [DDBJ:D26155] and [GenBank:NM_020823]
(b) [DDBJ:AB023216] and [GenBank:NM_016552]
(c) [DDBJ:L33075] and [GenBank:BC032784]
(d) [DDBJ:AK130557] and [GenBank:BC136777]
(e) [DDBJ:L14837] and [GenBank:AK124977]
(f) [DDBJ:U09825] and [GenBank:NM_025106]
(g) [DDBJ:AB020656] and [GenBank:BC030148]
Exon
Intron
TSR
Inverted Complementar Repeat

Figure 5: From (a) to (g) we show the alignments between candidate sequences and their associated transcripts (transcripts containing the TSR). For better visualization, only those alignments with at least 250 bp are shown. The majority of these repetitive sequences are actually Alu-like repeats.

Table 4: H-InvDB HIX0019725 cluster transcripts

AC	HIT	Tissue	5′UTR	CDS	3′UTR
[DDBJ:AF458052] (*GRM7*)	HIT000079970	[a]	451	2775	163
[DDBJ:AF458053] (*GRM7*)	HIT000079971	[a]	451	2736	135
[DDBJ:AF458054] (*GRM7*)	HIT000079972	[a]	451	2721	31
[DDBJ:U92458] (*GRM7*)	HIT000222625	Fetal brain	451	2748	1113

[a] Brain, trachea, testis, uterus, salivary gland.

Complete: H-InvDB cluster HIX0019725 transcripts. For each transcript, AC, identification in H-Inv (HIT), tissue origin, and the size of 5′UTR, CDS and 3′UTR.

might be an interchromosomal *trans*-splicing candidate (Figure 7).

All CDS and 3′UTRs were successfully mapped onto chromosome 14 (location 14q31.3) with high identity and coverage scores. The chromosome 14 assemblage has been completed [41], i.e. a single continuous segment with no gaps; there are 87 410 661 bps, representing its whole euchromatic portion in a single scaffold. The sequence HIT00000122 is the longest one in the H-InvDB Cluster

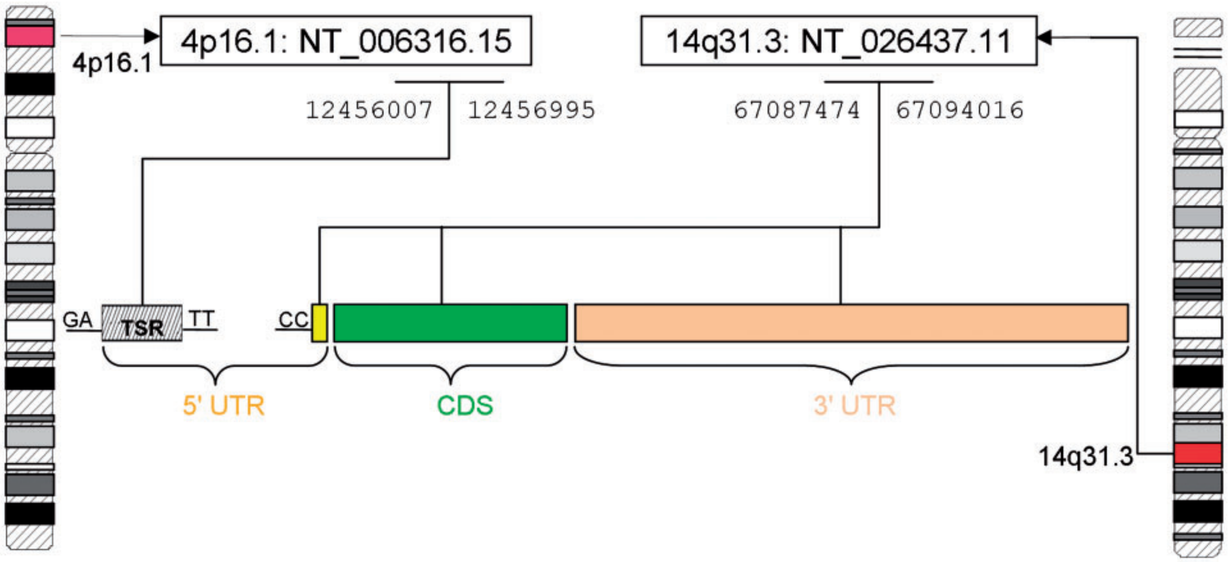

Figure 6: FLRT2 interchromosomal *trans*-splicing transcript structure. The 5′UTR has 987 bp that are *trans*-spliced from chromosome 4. The remaining sequence belongs to chromosome 14.

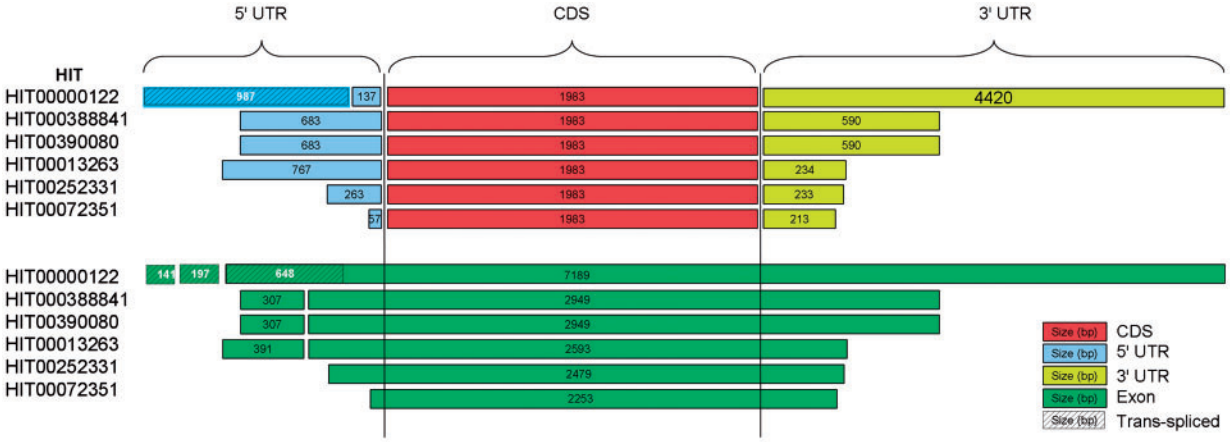

Figure 7: Cluster HIX0011865 transcripts. The transcript structure is shown. At the top, we show the six mRNAs and their 5′-UTR, CDS and 3′-UTR. At the bottom, the same set of genes and the sizes of each region considering exon–intron structures.

HIX0011865, with 7527 nucleotides. Its tissue origin is brain, with 1124 and 4420 nucleotide-long 5′UTR and 3′UTR sequences, respectively. Exactly 987 nucleotides from 5′UTR have been mapped onto chromosome 4 with high (100%) identity, and no other part of it was found in other chromosomes. As reported earlier, its TSR was found within an intron belonging to an antisense gene (*KCNIP4*).

CONCLUSION

In this work, we have shown a methodology for searching for hybrid mRNAs, the characteristics of which most resemble those of the best experimental assay. For illustrative purposes, we used a curated human FL-cDNA database, selected filtering criteria that were supported by experimental evidence, and we successfully found 16 human hybrid mRNAs. It is worth noting that in humans, except in a single case, the interchromosomal *trans*-splicing mechanism does not follow consensus splicing site rules similar to the standard *cis*-splicing mechanism; however, these rules have been extensively applied in other screening methodologies, which partially explains why none of our candidates had been previously identified.

This result may indicate that *trans*-splicing, although rare, may be more widespread than believed. Given that none of our candidate sequences follows canonical splicing sites, either infrequent splicing sites occur under special conditions or a non-spliceosome-mediated *trans*-splicing mechanism may be involved, such as the recently reported tRNA-mediated *trans*-splicing mechanism. Moreover, both genomic regions of the candidate sequences and their associated transcripts have complementary inverted repeats (predominantly *Alu* elements) that might be involved in a possible non-spliceosome-mediated *trans*-splicing mechanism, as conjectured in ref. [26].

Our *trans*-splicing candidate sequences are supposedly from normal tissues, and a recent study [42] has reported that *trans*-splicing may occur not only in malignant tissues, but in normal tissues as well, which is in agreement with our *in silico* findings.

Finally, despite the fact that our methodology was developed using human data, it can also be applied to other organisms, and the choice of 5′UTR to search for *trans*-splicing events, although supported by experimental assays, was intended for illustrative purposes only. It is clear that *trans*-splicing events may also occur in CDS or 3′UTR, and our methodology is intended to deal with these cases as well.

Key Points

- *Trans*-splicing may be mediated by tRNA without the spliceosome machinery, which means that non-canonical splicing sites may occur.
- *Trans*-splicing is not only present in malignant tissues, but in normal tissues as well.
- A screening pipeline may be easily applied to other databanks to search for *trans*-splicing candidates in other organisms.
- The bioinformatics methodologies published so far did not take these findings into account, which partially explains why none of our candidate sequences was reported earlier.

FUNDING

Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPq [grant number 140730/2008-7]; and Fundo de Amparo à Pesquisa do Estado de São Paulo—FAPESP [grant number 2008/02647-3].

References

1. Tahara M, Pergolizzi RG, Kobayashi A, *et al*. Trans-splicing repair of CD40 ligand deficiency results in naturally regulated correction of a mouse model of hyper-IgM X-linked immunodeficiency. *Nature Med* 2004;**10**:835–41.
2. Pergolizzi R, Ropper A, Dragos R, *et al*. In Vivo trans-splicing of 5′ and 3′ segments of pre-mRNA directed by corresponding DNA sequences delivered by gene transfer. *Mol Ther* 2003;**8**:999–1008.
3. Schlesinger J, Arama D, Noy H, *et al*. In-cell generation of antibody single-chain Fv transcripts by targeted RNA trans-splicing. *J Immunolog Meth* 2003;**282**:175–86.
4. Garcia-Blanco MA. Messenger RNA reprogramming by spliceosome-mediated RNA trans-splicing. *J Clin Invest* 2003;**112**:474–80.
5. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. *Nat Biotechnol* 2004;**22**: 535–46.
6. Mansfield SG, Hawkins-Clark R, Puttaraju M, *et al*. 5′ exon replacement and repair by spliceosome-mediated RNA trans-splicing. *RNA* 2003;**9**:1290–7.
7. Chao H, Mansfield SG, Bartel R, *et al*. Phenotype correction of Hemophilia A mice by spliceosome-mediated RNA trans-splicing. *Nat Med* 2003;**9**:1015–9.
8. Otto E, Temple GF, McGarrity GJ. Re-programming gene expression using spliceosome-mediated RNA trans-splicing (SMaRT$^{\text{TM}}$). *Curr Drug Discovery* 2003;**6**: 37–42.
9. Dallinger G, Puttaraju M, Mitchell LG, *et al*. Development of spliceosome-mediated RNA trans-splicing (SMaRT$^{\text{TM}}$) for the correction of inherited skin diseases. *Exp Dermatol* 2003;**12**:37–46.
10. Liu X, Jiang Q, Mansfield SG, *et al*. Functional restoration of CFTR chloride conductance in human CF epithelia by spliceosome-mediated RNA trans-splicing. *Nat Biotechnol* 2002;**20**(1):47–52.
11. Bonen L. Trans-splicing of pre-mRNA in plants, animals and protests. *FASEB J* 1993;**7**:40–6.
12. Blumenthal T. *Community TCeR. Trans-splicing and Operons*. Pasadena, CA: WormBook, The C. elegans Research Community Edition, 2005.
13. Mayer MG, Floeter-Winter LM. Pre-mRNA trans-splicing: from kinetoplastids to mammals, an easy language for life diversity. *Mem Inst Oswaldo Cruz* 2005; **100**:501–13.
14. Labrador M, Mongelard F, Plata-Rengifo P, *et al*. Protein encoding by both DNA strands. *Nature* 2001;**409**(6823): 1000.
15. Horiuchi T, Aigaki T. Alternative trans-splicing: a novel mode of pre-mRNA processing. *Biol Cell* 2006;**98**(2): 135–40.
16. Fischer SE, Butler MD, Pan Q, *et al*. Trans-splicing in C. elegans generates the negative RNAi regulator ERI-6/7. *Nature* 2008;**455**(7212):491–6.
17. Robertson HM, Navik JA, Walden KKO, *et al*. The Bursicon gene in mosquitoes: an unusual example of mRNA trans-splicing. *Genetics* 2007;**176**:1351–3.
18. Caudevilla C, Serra D, Miliar A, *et al*. Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci USA* 1998;**95**(21):12185–90.

19. Rigatti R, Jia J-H, Samani NJ, *et al*. Exon repetition: a major pathway for processing mRNA of some genes is allele-specific. *Nucleic Acids Res* 2004;**32**:441–6.

20. Roux M, Levéziel H, Amarger V. Cotranscription and intergenic splicing of the PPARG and TSEN2 genes in cattle. *BMC Genomics* 2006;**4**:7–71.

21. Viles KD, Sullenger BA. Proximity-dependent and proximity-independent trans-splicing in mammalian cells. *RNA* 2008;**14**(6):1081–94.

22. Chen C, Fossar N, Weil D, *et al*. High frequency trans-splicing in a cell line producing spliced and polyadenylated RNA polymerase I transcripts from an rDNA-myc chimeric gene. *Nucleic Acids Res* 2005;**33**(7):2332–42.

23. Hahn Y, Bera TK, Gehlhaus K, *et al*. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci USA* 2004;**101**(36):13257–61.

24. Fitzgerald C, Sikora C, Lawson V, *et al*. Mammalian transcription in support of hybrid mRNA and protein synthesis in testis and lung. *J Biol Chem* 2006;**281**:38172–80.

25. Chong A, Zhang G, Bajic VB. Information for the coordinates of exons (ICE): a human splice sites database. *Genomics* 2004;**84**(4):762–6.

26. Di Segni G, Gastaldi S, Tocchini-Valentini GP. Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells. *Proc Natl Acad Sci USA* 2008;**105**:6864–9.

27. Anderson AM, Staley JP. Long-distance splicing. *Proc Natl Acad Sci USA* 2008;**105**:6793–4.

28. Breen MA, Ashcroft SJH. A truncated isoform of Ca^{2+}/calmodulin-dependent protein kinase II expressed in human islets of Langerhans may result from trans-splicing. *FEBS* 1997;**409**:375–9.

29. Li B-L, Li X-L, Duan Z-L, *et al*. 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* 1999;**274**:11060–71.

30. Pesole G, Mignone F, Gissi C, *et al*. Structural and functional features of eukaryotic mRNA unstraslated regions. *Gene* 2001;**276**:73–81.

31. Romani A, Guerra E, Trerotola M, *et al*. Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res* 2003;**31**:17–25.

32. Rabbitts TH. Chromosomal translocations in human cancer. *Nature* 1994;**372**:143–9.

33. Shao X, Shepelev V, Fedorov A. Bioinformatic analysis of exon repletion, exon scrambling and trans-splicing in humans. *Brief Bioinformat* 2006;**22**:692–8.

34. Imanishi T, Itoh T, Suzuki Y, *et al*. Integrative annotation of 21037 human genes validated by full-length cDNA clones. *PLoS Biol* 2004;**2**:856–75.

35. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;**12**:656–64.

36. Lomsadze A, Ter-Hovhannisyan V, Chernoff Y, *et al*. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 2005;**33**(20):6494–506.

37. Flockerzi A, Ruggieri A, Frank O, *et al*. Expression patterns of transcribed human endogenous retrovirus HERV-K (HML-2) loci in human tissue and the need for a HERV Transcriptome Project. *BMC Genom* 2008;**9**: 354.

38. Wu S, Wright RA, Rockey PK, *et al*. Group III human metabotropic glutamate receptors 4, 7 and 8: molecular cloning, functional expression, and comparison of pharmacological properties in RGT cells. *Brain Res Mol Brain Res* 1998;**53**(1–2):88–97.

39. Schulz HL, Stohr H, Weber BH. Characterization of three novel isoforms of the metabotrobic glutamate receptor 7 (GRM7). *Neurosci Lett* 2002;**326**(1):37–40.

40. Lacy SE, Bonnemann CG, Buzney EA, *et al*. Identification of FLRT1, FLRT2 and FLRT3: a novel family of transmembrane leucine-rich repeat proteins. *Genomics* 1999;**62**: 417–26.

41. Heilig R, Eckenberg R, Petit JL, *et al*. The DNA sequence and analysis of human chromosome 14. *Nature* 2003;**421**: 601–7.

42. Li H, Wang J, Ma X, Sklar J. Gene fusions and RNA trans-splicing in normal and neoplastic human cell. *Cell Cycle* 2009;**8**:218–22.