Transsplicing fusion from MSMB to NC0A4

Thomas D. Schneider *

version = 1.01 of transsplice.tex 2011 Nov 4

An RNA fusion was found in Mike Dean's lab on chromosome 10 between the end of exon 2 in gene MSMB to the beginning of exon 2 of NC0A4. How was it created?

```
ATGAATGTTCTCCTGGGCAGCGTTGTGATCTTTGCCACCTTCGTGACTTTATGCAATGCATCATGCTATT
TCATACCTAATGAGGGAGTTCCAGGAGATT
(fusion)
CAACCAGGAGAGCAGTGAGGAGAATGAATACCTTCCAAGACCAGAGTGGCAGCTCCAGTAATAGAGAACC
CCTTTTGAGGTGTAGTGATGCACGGAGGGACTTGGAGCTTGCTATTGGTGGAGTTCTCCGGGCTGAACAG
CAAATTAAAGATAACTTGCGAGAG
```

I used the UCSC Genome browser (`http://genome.ucsc.edu/`) blat program (`http://genome.ucsc.edu/cgi-bin/hgBlat`[1]) with the two parts of the sequence to identify the locations in the human genome. The reported fusion junction is after base 51555827 (hg19 coordinates), supposedly this is the end of MSMB exon 2. The first sequence gives one hit at chr10:51555730-51555827:

```
cDNA YourSeq
atGAATGTTC TCCTGGGCAG CGTTGTGATC TTTGCCACCT TCGTGACTTT  50
ATGCAATGCA TCATGCTATT TCATACCTAA TGAGGGAGTT CCAGGAGATT  100
```

The lowercase 'at' indicates that the blat program locates intronic sequence there. Also blat reports that the sequence ends on the last base at a putative splice junction.

The second part of the sequence gives 6 hits one of which is at chr10:51579128-51579282. Surprisingly, other 5 have identities of 97.5% 97.5% 90.3% 93.4% and 85.8% identical. The predicted cDNA does not match exactly on the 5′ end:

```
cDNA YourSeq
caaccaggaG AGCAGTGAGG AGAATGAATA CCTTCCAAGA CCAGAGTGGC  50
AGCTCCAGTA ATAGAGAACC CCTTTTGAGG TGTAGTGATG CACGGAGGGA  100
CTTGGAGCTT GCTATTGGTG GAGTTCTCCG GGCTGAACAG CAAATTAAAG  150
ATAACTTGCG AGAG
```

*National Institutes of Health, National Cancer Institute at Frederick, Gene Regulation and Chromosome Biology Laboratory, P. O. Box B, Frederick, MD 21702-1201. (301) 846-5581, email: schneidt@mail.nih.gov http://alum.mit.edu/www/toms/

[1]You can click on this link in the PDF to get the blat program in your browser. Be sure to set the genome to hg19.

Human donor and acceptor site models built by Pete Rogan from 111772 and 108079 sequences respectively were scanned over the sequences around the two junction points as shown in Fig. 1 [1, 2]. Piece 1, the region around the fusion juction of MSMB shows clearly that there is no sequence walker for a donor site at the junction between 51555827 and 51555828 (marked by a vertical bar). However, a decent 6.0 bit site is at 51555837 (marked by a second vertical bar). In addition, setting the genome browser to chr10:51,555,800-51,555,864 reveals that exon 2 ends with the amino acid sequence S-T-R. Finally, the sequence between the two vertical bars is caaccagga which is ***exactly*** the same as the extra sequence on the 5′ of NCOA4 mentioned above. In other words, the location of the fusion is at the second bar, not the first bar as initially reported.

Piece 2 in the figure shows a strong branch point site (unpublished model) at 51579090 and a spectacularly strong acceptor site at 51579127 in NCOA4. Splicing between the donor on MSMB and this acceptor on NCOA4 recreates the observed fusion mRNA. (the green bars are the zero coordinates of the sequence walkers and they are defined to be on the intron side of the junction.)

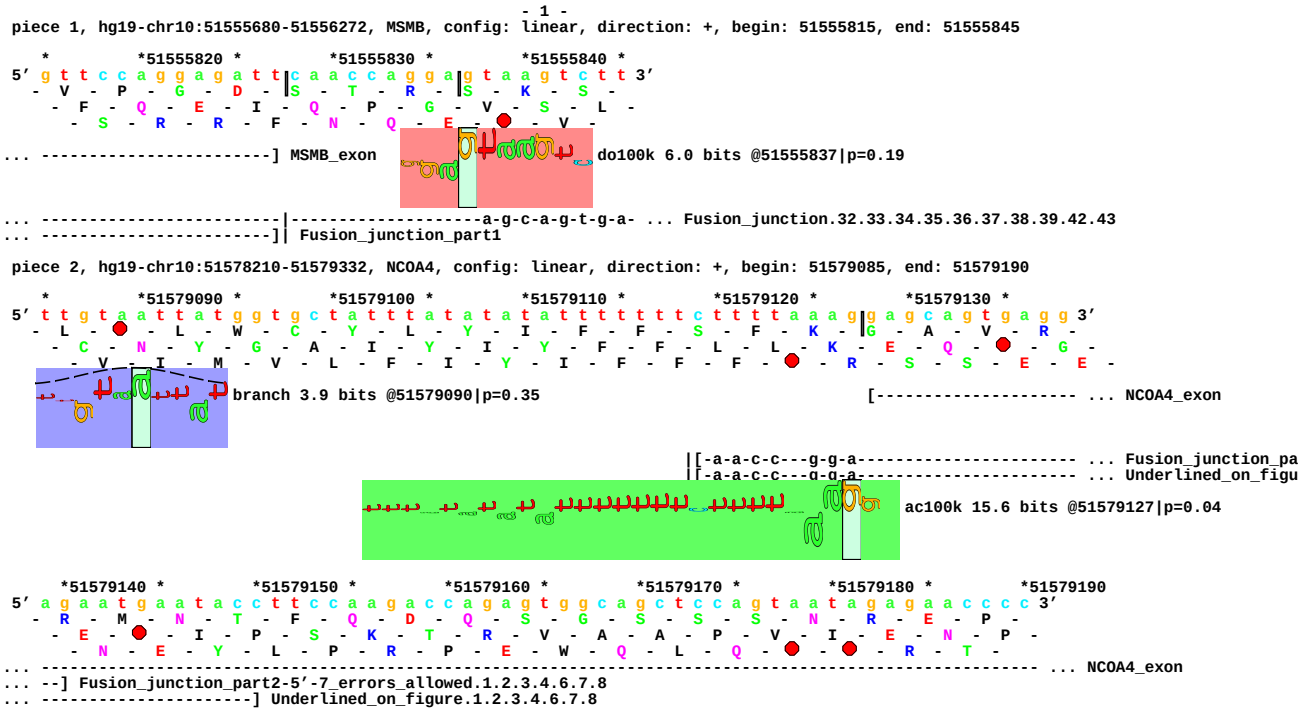

Figure 1: Lister map of the regions around the transsplicing fusion between the end of exon 2 in MSMB and the beginning of exon 2 of NC0A4.

Fig. 2 shows that the 5′ end of MSMB exon 2 has a strong acceptor of 10.0 bits with a superb branch point of 7.3 bits. Since the spliceosome first binds the acceptor and then scans to the donor according to the exon definition model [3, 4, 5, 6, 7, 8, 9, 10, 11], this exon should be favored for splicing to the strong acceptor of NCOA4.

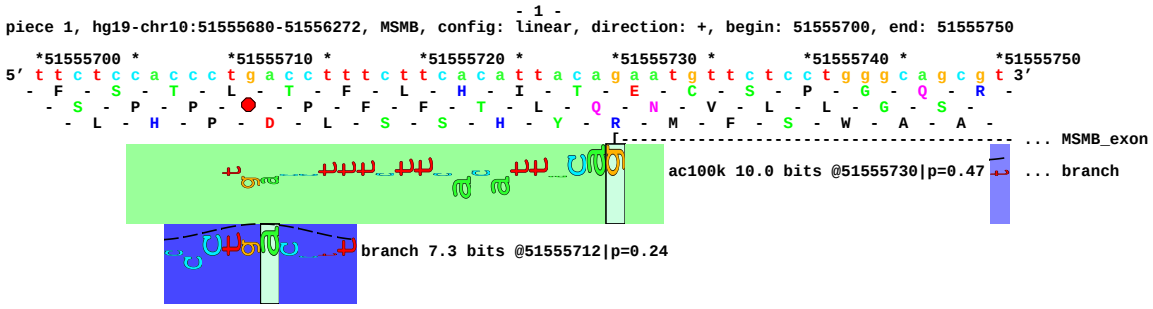

Figure 2: Lister map at the start of MSMB exon 2.

These observations establish that the fusion is likely to be between a strongly ‘activated’ donor and a strong acceptor. However, the distance between the two is 23290 bases. How can they get together? The observation mentioned above that there are 5 other sequences on the genome that strongly match to the NCOA4 exon 2 Could those places be involved in pairing the mRNA? I hypothesized that there is an mRNA that pairs NCOA4 exon 2 with the region downstream of exon2 in MSMB. So I used BLAT to look for the region downstream of MSMB exon2:

`gtaagtcttggcttttcaatgtttattatgttattgcagcc`

Using `http://genome.ucsc.edu/cgi-bin/hgBlat` gives, with splicing, ***5 other exact 100.0% matches in the genome!*** None are at NC0A4. To push this further I looked 200 bases downstream of MSMB exon 2:

```
gtaagtcttggcttttcaatgtttattatgttattgcagcctggtagatggacctgtctg
cagatgaaagcctttgtgtttctgtttgtttgtttctttgtttttttgagatagtcatgct
ctgtctcccaggctggagtgcagtggcaccatttcagctcactgcaaactcctcctcccg
```

This gives a stunning 192 hits with a mean of 90.5 ± 3.4% identity! These are all SINE elements. How to compare this to the sequence of NCOA4? I took more of the second part of the fusion:

```
CAACCAGGAGAGCAGTGAGGAGAATGAATACCTTCCAAGACCAGAGTGGCAGCTCCAGTAATAGAGAACCCCTTTTGAGGTGT
AGTGATGCACGGAGGGACTTGGAGCTTGCTATTGGTGGAGTTCTCCGGGCTGAACAGCAAATTAAAGATAACTTGCGAGAG
```

5 more hits. The second is also a SINE.

Examining repeats in these regions: chr10:51,555,827-51,556,354 shows that AluSx (chr10:51555915-51556222) is just downstream of MSM exon 2 and AluJr (chr10:51578260-51578557) is just upstream of NCOA4.

How similar are the two Alus? google: ‘ncbi blast two sequence alignment’ gives BLAST 2 Sequences - NCBI - National Institutes of Health (NIH) `http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi` Using "Somewhat similar sequences (blastn)":

```
>lcl 2703
Length=298

 Score =  172 bits (190),  Expect = 9e-48
 Identities = 216/292 (74%), Gaps = 7/292 (2%)
 Strand=Plus/Plus

Query  20   ttgtt-ttttgAGATAGTCATGCTCTGTCT--CCCAGGCTGGAGTGCAGTGGCACCATTT  76
            |||||  ||||||||:||   | | |: |||  |||||| ||||||||||  |||  ||:
Sbjct  11   TTGTTGTTTTGAGACAGAGTTTCACCCTCTTACCCAGGGTGGAGTGCAGGTGCA--ATCA  68

Query  77   CAGCTCACTGCAAACTCCTCCTCCCGGGTTCCAGTGATTCTACCTCCTCAGCCTCCCAAG  136
            |:||||||||||: : |   ||||:|||:|| :||||||:| || :||||||||  |||||
Sbjct  69   CGGCTCACTGCAGCTGCGAACTCCTGGGCTCAGGTGATCTTCCCATCTCAGCCTGCCAAG  128

Query  137  TAGCTGGGACTACAGGTGTGTGCCATCATGCCTGGCTAATTTTTGTATTTTTGGTAAAGA  196
            |||||||||||||||:||::::|||:||::: ||||||||||||:||||||  |||:|||
Sbjct  129  TAGCTGGGACTACAGATGCACACCACCACATGTGGCTAATTTTTATATTTT--GTAGAGA  186

Query  197  TGGGGCTTTGCCACGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTCATCTGCCCG  256
            :||||::| ||:|: ||| |||||||||||:||||||||||:  |||:|| |||::|||:
Sbjct  187  CGGGGTCTGGCTATTTTGCCCAGGCTGGTCCCGAACTCCTGGAATCAAGTGATCCACCCA  246

Query  257  CCACAGCCTCCCAAAGTGTTGGGATTACAGGTGTGAGCCACTGTGCCTGGCC  308
            || ::|||||||||||||::|:|:|||||||||||||||||||:|||||:||||
Sbjct  247  CCTTGGCCTCCCAAAGTACTAGAATTACAGGTGTGAGCCACCGTGCCCGGCC  298

| means an exact match.
: means a T to C pair or A to G which correspond to a U:G bond in RNA.
Including these, the sequences match to 89%.
```

Trans Splicing Model

A model for bringing the two exons together is a reverse transcript of an Alu element (not necessarily either of the ones involved here, but it could be) that base pairs with the two mRNA transcripts. There can also be a second reverse reverse transcript that also binds to form a complete helix with a crossover as in a Holliday junction. The junction would be in the middle of the repeat element sequences and could move back and forth. The two sequences would look like a χ symbol as shown in Fig. 3.

A prediction of this model is that there could be a reverse promoter driving either one or both of the Alu element copies. If they are strong and local, they would match one sequence perfectly and match the other well enough to more or less frequently bring the two exons together. This promoter might be identified and sequenced by qPCR with the appropriate primers for the Alu sequences. That is, one could determine if the reverse transcripts come from one or the other of these genes or from somewhere else on the chromosome. Indeed, if there were a SNP anywere

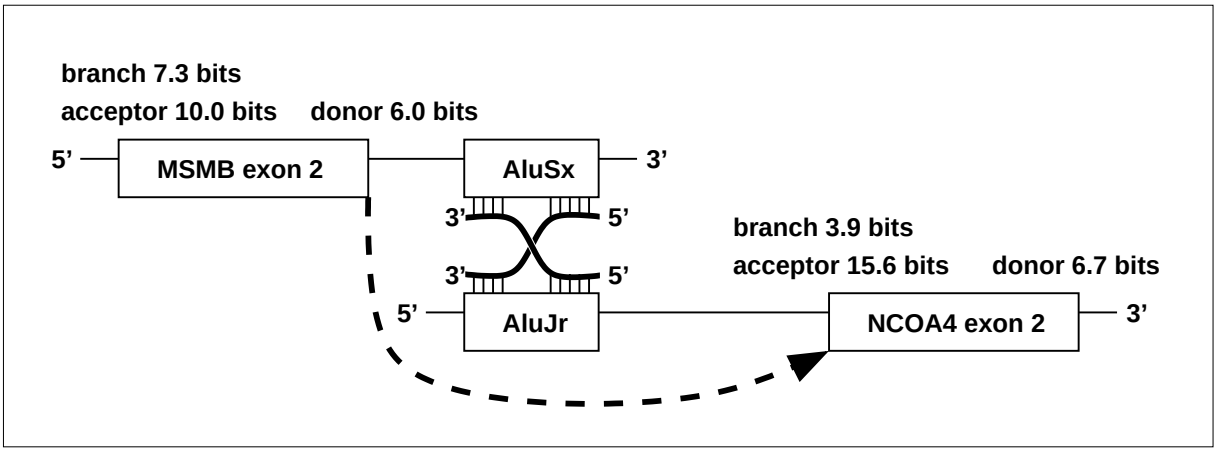

Figure 3: χ model of Alu RNAs could bring two distant exons together.

that increased transcription of an Alu that could hybridize to AluSx and AluJr it could cause the abnormal fusion.

An alternative hypothesis is that the RNA polymerase reading the AluSx dislodges and starts reading AluJr since the sequences are similar. Then the splicing proceeds as before to fuse MSMB exon 2 to NCOA4 exon 2. In this case there will be a continuous mRNA with a transition somewhere inside the Alu sequences. Again, this could be detected by qPCR and sequenced. Interestingly, the primers for this test are on the complement of the primers for testing the Alu transcript hypothesis.

Misha Kashlev pointed out that that RNA Pol II could indeed skip from one DNA to the other and he suggested that it could be enhanced to do so by a CTCF site that promotes pausing.

google: 'transsplicing' `http://en.wikipedia.org/wiki/Trans-splicing` Trans splicing is found in normal cells.

google: 'trans-splicing alu' Second hit is [12]. They say:

> Another feasible way to achieve trans-splicing of mRNAs by the tRNA endonuclease is to exploit the vast repertoire of repetitive sequences present in eukaryotic organisms. In the human genome, about half of the nucleotide sequence consists of repetitive elements (41). The short repeats (SINEs, short interspersed nuclear elements) are related to tRNA genes or other RNA Polymerase III-transcribed genes, and their number ranges from a few hundred to ≈500,000 for the MIR (mammalian interspersed repeats), which is a tRNA-derived family (42). Moreover, there are more than one million copies of the Alu elements, the most abundant family of repeats typical of humans and other primates, that by themselves comprise ≈10% of the whole genome (43). In the past, repetitive sequences were called "junk" DNA. Nowadays, however, considerable evidence points to a more complex picture, where repetitive elements can be recruited to reshape the genome and promote its evolution. Repetitive sequences thus

constitute a large reservoir of potential regulatory elements, functioning, for instance, in alternative splicing, RNA editing, transcription and translation regulation (43-45).

The second Wikipedia reference is [13]; they investigate trans splicing.

I have not found the sequences yet so don't know if this one could use Alu sequences to bring the parts together.

Some additional notes and speculations:

1. Since there are so many Alu sequences in the genome, this many not be the only or entire mechanism that brings the two exons together.
2. I believe Misha Kashlev mentioned that Alu sequences have transcription in both directions, so more complex structures are possible.
3. SNPs that induce transcription in an Alu could start the transsplicing.
4. Jeff Strathern pointed out that current sequencing technologies use short reads and so transspliced sequences tend to be thrown out of the data sets. So this phenomena may be more common than anticipated.
5. It is possible that high Alu transcription initiates two DNA regions coming close together. Once two regions are close, continuing transcription could maintain the state. If transcription is shut off, the two DNA regions could separate. In other words, Alu transcription might provide for a memory mechanism, and this mechanism could be common.
6. The brain uses a lot of alternative splicing. Could Alu transcription be used to create memories?

References

[1] R. M. Stephens and T. D. Schneider. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.*, 228:1124–1136, 1992. http://alum.mit.edu/www/toms/papers/splice/.

[2] P. K. Rogan, B. M. Faux, and T. D. Schneider. Information analysis of human splice site mutations. *Human Mutation*, 12:153–171, 1998. Erratum in: Hum Mutat 1999;13(1):82. http://alum.mit.edu/www/toms/papers/rfs/.

[3] M. R. Green. Pre-mRNA splicing. *Annu. Rev. Genet.*, 20:671–708, 1986.

[4] P. A. Sharp. Splicing of messenger RNA precursors. *Science*, 235:766–771, 1987.

[5] M. R. Green. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu. Rev. Cell Biol.*, 7:559–599, 1991.

[6] B. L. Robberson, G. J. Cote, and S. M. Berget. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.*, 10:84–94, 1990.

[7] M. Talerico and S. M. Berget. Effect of 5′ splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol.*, 10:6299–6305, 1990.

[8] M. Niwa, C. C. MacDonald, and S. M. Berget. Are vertebrate exons scanned during splice-site selection? *Nature*, 360:277–280, 1992.

[9] L. P. Eperon, J. P. Estibeiro, and I. C. Eperon. The role of nucleotide sequences in splice site selection in eukaryotic pre-messenger RNA. *Nature*, 324:280–282, 1986.

[10] M. Ohno, H. Sakamoto, and Y. Shimura. Preferential excision of the 5′ proximal intron from mRNA precursors with two introns as mediated by the cap structure. *Proc. Natl. Acad. Sci. USA*, 84:5187–5191, 1987.

[11] M. Niwa and S. M. Berget. Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes Dev.*, 5:2086–2095, 1991.

[12] G. Di Segni, S. Gastaldi, and G. P. Tocchini-Valentini. Cis- and trans-splicing of mRNAs mediated by tRNA sequences in eukaryotic cells. *Proc. Natl. Acad. Sci. USA*, 105:6864–6869, 2008.

[13] H. Li, J. Wang, G. Mor, and J. Sklar. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, 321:1357–1361, 2008.