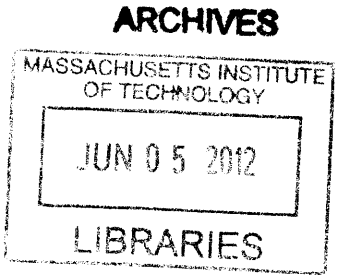
Conservation of Exon Scrambling in Human and Mouse

by

Monica L. Hamilton

B.S., Biology
Duke University, 2010

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Biology



at the

Massachusetts Institute of Technology

June 2012

Signature of Author: _____
Department of Biology
May 25, 2012

Certified by: _____
Christopher B. Burge
Professor of Biology and Biological Engineering
Thesis Supervisor

Accepted by: _____
Stephen P. Bell
Professor of Biology
Co-Director, Graduate Committee

Conservation of Exon Scrambling in Human and Mouse

by

Monica L. Hamilton

Submitted to the Department of Biology
on May 25, 2012 in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Biology

ABSTRACT

Exon scrambling is a phenomenon in which the exons of an mRNA transcript are spliced in an order inconsistent with that of the genome. In this thesis, I present a computational analysis of scrambled exons in human and mouse. RNA-seq data was mapped to the genome and all unaligned reads were subsequently mapped to a database of all possible exon-exon junctions. Eight conserved genes were found to undergo scrambled splicing in both species. In several cases, not only the gene was conserved, but the particular exons involved were conserved as well. Reading frame was preserved in just over half of the events, indicating that although some transcripts may be translated into protein, some may be non-functional or may play a regulatory role. The introns flanking scrambled exons were significantly longer than average, providing clues to the mechanism for this abnormal splicing pattern. The results of this study demonstrate that presence of scrambled transcripts in the cell is infrequent, but can be conserved over tens of millions of years of evolution, suggesting it has a biological function.

Thesis Supervisor: Christopher B. Burge
Title: Professor

Acknowledgments

I would like to thank my advisor, Chris Burge, for both his scientific mentorship and his support of my goals. I would also like to thank all of the members of the Burge lab. I appreciate not only their willingness to answer questions and lend advice, but also their camaraderie during this past year.

Table of Contents

Introduction

Splicing, the removal of introns and joining of exons to form a mature transcript[1,2], is a critical step in the processing of RNA in higher eukaryotes. Precise excision of introns is crucial to an organism's viability; if a mutation occurs in a splice site or other splicing regulatory sequence, it can lead to disease[3,4]. In fact, it is thought that between 50% and 60% of disease-causing mutations affect splicing[5]. Additionally, *trans*-acting mutations that disrupt the splicing machinery can have drastic consequences on the health of an individual, such as myotonic dystrophy[6]. For these reasons, splicing must be heavily regulated.

The cellular machinery responsible for catalyzing this exact intron removal and exon joining is the spliceosome, a large complex of RNA and protein[7,8]. The recognition of exon-intron boundaries is aided by three sequences that are present in all introns from yeast to humans: the 5' splice site, the 3' splice site, and the branch point sequence (Figure 1). Additionally, metazoans possess a polypyrimidine tract upstream of the 3' splice site that contributes to splicing. If splicing were a completely constitutive process, these sequences might provide sufficient instructions for the cell; however, the situation is more complex.

The average human gene contains 10.4 exons, yet the average mature transcript contains only 9.1 exons[9]. This discrepancy can be explained by the phenomenon of alternative splicing (AS), in which the exons of a gene can be joined together in various combinations. AS can be categorized into several classes (Figure 2)[10–13]. The most prevalent type of AS is exon skipping (also known as a cassette exon), in which an exon is not spliced into the mature transcript and is instead removed alongside its two flanking

introns. In the case of mutually exclusive exons, one of two neighboring exons is included while the other is excluded; the exons can be neither both included nor both excluded. In addition to events that include or exclude an entire exon, there are two types of events that allow an exon to be partially skipped: alternative 5' splice sites and alternative 3' splice sites. Finally, intron retention is the rarest type of alternative splicing, in which an intron is included in the final transcript.

Alternative splicing enables organisms to diversify their expression without increasing the size of their genomes. The *Drosophila* gene DSCAM, the poster child for alternative splicing, has the potential to form more than 38,000 isoforms, due to mutually exclusive splicing of its 95 exons[14,15]. This figure only considers the isoforms that can be formed when splicing occurs in an order consistent with the genome. If genomic sequence were not a constraint on splicing and instead exons were allowed to join in any order, the number of potential isoforms that could be generated would be astronomical.

In fact, there are several known exceptions to this sequential splicing rule. One such example is *trans*-splicing during which exons from two different RNA molecules are spliced together to form a single mature transcript. This type of splicing was first observed in trypanosomes. Trypanosome transcripts undergo SL (spliced leader) *trans*-splicing in which one of the transcripts contains a common 39-base sequence (termed the spliced leader) that is donated to the 5' of end of the other variable transcript[16]. In this way, every trypanosome transcript contains the same first exon. Later, many other organisms from diverse phyla were found to exhibit SL *trans*-splicing including euglenozoa, cnidarians, nematodes, platyhelminthes, and tunicates[17,18].

Though SL *trans*-splicing has not been observed in arthropods or vertebrates, other forms of non-sequential splicing have been found. For example, the *Drosophila* genes *mod(mdg4)*[19,20] and *lola*[21] exhibit *trans*-splicing between distinct RNA molecules. In the case of the *mod* gene, exons from both the sense and antisense strands are incorporated into a single transcript; with *lola,* the exons are all derived from the same strand, but interallelic complementation studies have provided evidence that exons from both copies of the gene are utilized. This mechanism differs significantly from SL *trans*-splicing in which the donated exon is non-coding and therefore does not contribute to protein functionality.

Finally, since the early 1990s, a phenomenon dubbed "exon scrambling" has been observed at low frequency[22–40]. In exon scrambling, exons from a single gene are spliced together using consensus splice sites, but in an order that differs from that of the genome. This novel type of splicing was first described by Nigro and coworkers in the tumor suppressor *DCC*[22]. Two sets of exon junction-specific primers were designed, one for each possible orientation of two putative neighboring exons, and both PCR products were observed, with the non-genomic orientation present at $1/1000^{th}$ the level of the genomic orientation. Importantly, 90% of the scrambled transcripts showed no evidence of a polyA tail, suggesting that the molecules were circular. Therefore, it was proposed that such transcripts are likely mistakes in the normal splicing process, as opposed to a meaningful biological phenomenon.

The first evidence for polyadenylation of transcripts came from Caldas *et al.* in 1998, demonstrating that not all scrambled transcripts are circular[26]. Due to the presumed linear nature of the transcripts, it seemed probable that exon scrambling may be the result

of *trans*-splicing of two different mRNA molecules. Polyadenylation of scrambled transcripts has been upheld in a number of subsequent publications[28,29,33,35,40]. This debate is far from over, however, with several other studies still providing support for circular transcripts[23–25,30].

Another significant finding associated with exon scrambling was the observation that the introns flanking the scrambled junctions tended to be especially long[23]. This finding has been upheld by several other reports[25,28,37,39], perhaps offering some mechanistic insight into the formation of these abnormal transcripts.

A different mechanistic explanation for exon scrambling was proposed by Zaphiropoulos[24,25]. After observing a correlation between scrambled exons and their reciprocal exon skipping events, it was suggested that exon scrambling and exon skipping may in fact be interrelated. This interrelatedness was questioned, though, by Caldas *et al.* in their study of the *MLL* gene[26], in which no such correlation was seen. No further evidence has been provided to support this reciprocal occurrence, though it would certainly provide a reasonable explanation for the existence of the abnormal transcripts.

Of all the exon scrambling instances identified to date, only one corresponding protein product has been catalogued[27]. Through Western blotting, two isoforms of the rat carnitine octanoyl-transferase (COT) protein were immunolocalized, one corresponding to the normal protein and the other corresponding to a protein with exons two and three repeated. Whether or not this protein is able to perform its normal function in the cell is still a question to be answered. It will be interesting to see if this represents a unique event or if other translated products can be found.

More recently, large-scale computational analyses have attempted to determine the frequency of scrambled events[38–40]. While these studies may have a higher rate of false positives due to their high throughput nature, they have been able to confirm some overarching themes, such as proximity to large introns.

Scrambled exons have long held a controversial position in the field of splicing, as to whether they represent a biologically significant variation or whether they are simply a consequence of noise in the splicing process. In this thesis, I aim to establish whether exon scrambling represents a meaningful splicing variation by examining conservation of scrambled events between human and mouse.

Materials and Methods

Transcriptome Sequencing

Human RNA-seq data was obtained from the Illumina Human Body Map 2.0 Project, which is available in the ENA archive with accession number ERP000546 (http://www.ebi.ac.uk/ena/data/view/ERP000546). The Body Map 2.0 data were generated by the Expression Applications R&D group at Illumina using a standard polyA-selected Illumina RNA-Seq protocol from total RNA obtained commercially (Ambion) using the HiSeq 2000 system. Reads were 50 bp in length.

Mouse RNA was collected from three individuals and in nine tissues: brain, colon, heart, kidney, liver, lung, skeletal muscle, spleen, and testes. Paired-end RNA-sequencing was performed using a standard polyA-selected on Illumina Hi-Seq. Reads were sequenced to a length of 36, 40, 50, 75, or 80 bp. Only 36 and 50 bp reads were used in this analysis, which were derived from two of the three mice.

Identification of Scrambled Events

Reads were aligned to the UCSC hg19 and mm9 genomes using the Bowtie aligner[42]. Only the best mappings were reported with a maximum of two mismatches. All reads that did not align to the genome were then aligned to the junction database using Bowtie with the same parameters as above. The junction database was generated to include all possible exon-exon junctions for each gene. Junctions were classified as forward (an upstream exon followed by a downstream exon), reverse (a downstream exon followed by an upstream exon), or same (the repetition of an exon).

Results

Human and mouse paired-end Illumina RNA sequencing reads were mapped to their respective genomes using Bowtie. All reads that failed to align were then mapped to a custom database of all permutations of exon-exon junctions, including exons whose order was rearranged relative to the genome and repeated exons. Junctions were considered scrambled if the 3' end of an exon was followed by the 5' end of a genomically upstream exon or by the 5' end of the same exon. Reads that mapped to the junction database were filtered to have at least 8 matching bases on each side of the exon-exon boundary. Additionally, both reads of the pair must have mapped to the same chromosome to be considered for analysis.

We found 6,975 reads mapping to scrambled junctions in mouse, representing 3,496 events and 1,916 genes. In human, we found 2,632 scrambled reads, with 626 events and 479 genes. As a quality filter, the analysis was limited to events that had at least one library with at least two supporting reads, reducing these numbers substantially. Out of the 14,773 genes that share orthology between mouse and human, 68 genes showed evidence of scrambling in mouse, 178 genes showed evidence of scrambling in human, and 11 genes show evidence of scrambling in both human and mouse. However, after visual inspection and manual alignment of the reads to the genome using UCSC's BLAT[42], only eight of these appear to be valid exon scrambling events (Figures 4 – 11). Using the hypergeometric test on these values yields a statistically significant p-value of 1.5×10^{-6}.

One of the conserved scrambled genes is Arid1a, an AT-rich interactive domain-containing protein. For both human and mouse, a single supporting read is shown in

Figure 4. In each species, the first half of the read maps to the 3' end of the fourth exon of Arid1a, while the second half of the read maps to the 5' end of the second exon. BLAT was used for validation of these mappings: each segment of the read yielded a single hit corresponding to the location found in the junction database[43]. By visual inspection, it appears as though the exons participating in the scrambled splicing (and not just the genes) are orthologous; this was confirmed by look-up in a gene-oriented exon orthology dataset by Fu *et al.*[44].

In addition to exons spliced out of order, we witnessed exons spliced to themselves. For example, in Slc8a1, a gene belonging to the sodium/calcium solute carrier family, the first half of the read maps to the 5' end of the second exon and the second half of the read maps to the 3' end of the same exon (Figure 5). Two possible explanations can account for this read: *trans*-splicing has occurred between two transcripts to yield a mature mRNA with a repeated second exon, or intra-transcript splicing has yielded a circular transcript (Figure 3). Because libraries were prepared with polyA selection, circular RNA molecules should be substantially under-represented in (or absent from) the sequencing results.

As stated above, out of the eleven scrambled genes shared between human and mouse, three appear to be false positives. One such case is obscurin (Obscn), a cytoskeletal calmodulin and titin-interacting RhoGEF. Obscn is a very large protein (about 800 kDa) present in striated muscle that is encoded by about 100 exons[45,46]. Most of the exons contain immunoglobulin- and fibronectin-like domains and undergo a large amount of alternative splicing. With such a highly expressed gene containing many

similar exons, it would not be surprising if some reads mapped to scrambled junctions instead of the junctions they were truly derived from due to sequencing errors.

The two other highly suspect scrambled genes are Tpm1 and Abi3bp. In both cases, the junction position is skewed toward the end in all supporting reads and there are one or two mismatches in the shorter overhang. It is difficult to estimate the number of false positives without manually examining the supporting reads for each potential exon scrambling event. One way to approximate this error rate is to look at the overall distribution of where in the read the junction falls. The more bases that accurately map to each side of the junction, the more likely the read represents a true exon scrambling event. Theoretically, the position of a junction within a read should be uniformly distributed. However, in the case of mouse, there is a large peak of junctions towards one end of the read (Figure 12). When reads are binned by number of mismatches, this peak is especially pronounced with two mismatches (the maximum number of mismatches allowed by the filter). This peak likely corresponds to a high level of false positives due to the less stringent criteria: two mismatches and just ten bases of overhang.

The exon-specific orthology analysis described for Arid1a was performed on all eight conserved genes (Table 1). In four of the genes (Arid1a, Rsf1, Slc8a1, and Man1a2), both exons involved in scrambling were orthologous between mouse and human. In three genes (Strn3, Zc3h6, and Anp32b), only one of the two exons was orthologous, and in one gene (App), neither exon was orthologous. Those genes that conserved the exon-exon junction are more likely to represent functional products than those for which just the gene is conserved.

Several previous studies had indicated that exon scrambling was often found next to large introns[23,25,30,38,47]. This appears to be consistent with the eight conserved scrambled genes as well. The intron upstream of the acceptor splice site had a mean length of 27,553 bp in mouse and 31,021 bp in human while the intron downstream of the donor splice site had a mean of 32,713 bp in mouse and 42,873 bp in human (Table 1). According to a z-test, these lengths are significantly greater than the overall mean intron length, documented at 2,874 bp in mouse and 3,749 bp in human[48]. This tendency towards longer introns flanking scrambled exons may be indicative of an underlying mechanism for exon scrambling.

In order to evaluate whether the eight conserved scrambled events have any functional consequences in the cell, each event was tested for preservation of reading frame. Reading frame was considered preserved if the reading frame of the scrambled 3' exon was the same as that of the normal 3' exon; for example, in a scrambling event in which exon 5 is followed by exon 2, the reading frames of exon 6 and exon 2 would be compared. In all, ten out of the sixteen events (mouse and human counted separately) exhibited reading frame preservation. Those events that introduced frame shifts would most likely produce mRNAs that would be degraded by the nonsense-mediated mRNA decay pathway. This may imply that the scrambled events have no functional role in the cell since they are unlikely to form stable protein. Alternatively, this may suggest a gene regulatory role for the exon scrambling.

Discussion

The splicing of non-sequential exons is not a novel concept in biology. *Trans*-splicing was demonstrated *in vitro* in 1985[47,48]. It also occurs ubiquitously in trypanosomes and other organisms in which a common spliced leader exon is spliced to the beginning of every transcript[16]. More recently, it has become clear that *trans*-splicing can occur between two transcripts of the same gene (homotypic trans-splicing), as opposed to between transcripts of different genes (heterotypic trans-splicing).

Exon scrambling, in which exons of the same gene are spliced in an order inconsistent with the genome, has been described in a number of higher eukaryotic species. The origin of these shuffled transcripts has been a matter of debate for some time. Two plausible explanations have been offered: a scrambled junction may represent a circular RNA molecule that is the by-product of the normal splicing process[22–25,30] or it may be the product of splicing between two transcripts of the same gene[26,28,29,33,35,40] (Figure 3). In this study, because polyadenylation selection was performed during library preparation, it is likely that the scrambled transcripts were derived from *trans*-splicing.

One limitation faced was the overall low number of scrambled reads, which made applying meaningful filters a challenge. Only events for which at least two supporting reads were found in the same library were considered for this study. Ideally a more stringent filter would have been applied, perhaps requiring multiple libraries or a larger number of reads per event, but due to the rarity of exon scrambling this was not possible. Additionally, since the reads were not barcoded, it was not possible to determine if identical reads were an artifact of PCR amplification of the same original mRNA fragment or if they represented unique events.

Another difficulty in using computational methods to identify potentially scrambled transcripts is the high level of false positives. The suspiciously large number of reads with a short overhang and two mismatches indicates that these may represent mis-mappings of normal exon-exon junctions rather than truly scrambled junctions (Figure 12). As with other computationally-based studies[40], manual curation was used to screen all putative conserved events for repetitive or suboptimal matches.

Consistent with several other reports[23,25,28,37,39], it was observed that exons involved in scrambling are adjacent to longer-than-average introns. This may possibly have a mechanistic basis. Splicing generally occurs co-transcriptionally, such that by the time transcription is finished, the pre-mRNA molecule has been processed into a mature transcript[49,50]. However, if an exceptionally large intron is being transcribed, then there may be a substantial period during which a 5' donor splice site is waiting to be spliced. The longer the intron, the more likely it may be that the free donor site will be spliced to another nearby nascent transcript. In fact, this hypothesis has been experimentally tested by Takahara *et al.*[37]. By inserting various sequences to pause the trajectory of RNAPII, they demonstrated that delay of 3' splice site transcription increases the frequency of *trans*-splicing.

In order to analyze whether exon scrambling may have functional consequences in the cell, putative scrambling events were tested for reading frame conservation. Although 6 of 16 events did not preserve reading frame and thus are likely subject to nonsense-mediated decay (NMD)[51–53], these transcripts may still play a regulatory role. By splicing together exons that result in a frameshift mutation, the effective amount of protein synthesized is decreased while maintaining the same level of transcription.

Of the eight scrambled genes that share orthology between human and mouse, one was also identified as a top hit in a separate large-scale search for scrambled exons[41]. As in this study, the authors found *Man1a2* to have many supporting reads and to occur in multiple samples. Additionally, of the three scrambled exon pairs reported in *Man1a2*, one pair (exon 5 joined to exon 2) matched the junction conserved by human and mouse in this study (Figure 6). Furthermore, in their experimental test for conservation in mouse, the *Man1a2* 5-2 PCR product was observed in all tissues analyzed. Finally, a quantitative PCR analysis showed that the 5-2 scrambled transcript was present at levels on par with the normal 5-6 junction. With such strong evidence in support of this event, it is tempting to speculate that it is translated into a functional protein product. This is certainly plausible, as the scrambled event preserves the normal reading frame. Because the protein has not yet been crystallized, it is difficult to predict what effect the repetition of exons 2, 3, 4, and 5 would have on the protein's structure and activity; nevertheless, it is the most promising example of scrambling with functionality.

Finally, the fact that there is such limited overlap between the eight conserved, manually curated scrambled genes from this study and the experimentally validated genes of other studies suggests that that there are still many more events to be documented. Clearly the screen for exon scrambling has not been saturated and it will require more data and more analysis before we have a complete understanding of the full extent of exon scrambling.

Whether referred to as *trans*-splicing[27,31,33,37], exon repetition[28,29,34,36], post-transcriptional exon shuffling[40], rearrangements or repetition in exon order (RREO)[39], or exon scrambling[22,26,32], transcripts with unordered exons still hold a controversial position

in the field of splicing. While this study has helped to identify several scrambled genes conserved between mouse and human, it is still unclear what role these transcripts may play in the cell. Though it is uncertain if they represent intentional splicing variations or simply noise in the cell's splicing machinery, the conservation of some of these events for the tens of millions of years separating human from mouse suggests that some exon scrambling play a biologically important role.

Acknowledgments

We thank Jason Merkin for his work in generating the mouse RNA-seq data sets and Alex Robertson for his analysis of the human data.

References

1. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 3171–3175 (1977).
2. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977).
3. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics* **3**, 285–298 (2002).
4. Evsyukova, I., Somarelli, J. A., Gregory, S. G. & Garcia-Blanco, M. A. Alternative splicing in multiple sclerosis and other autoimmune diseases. *RNA Biol* **7**, 462–473 (2010).
5. Wang, G.-S. & Cooper, T. A. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics* **8**, 749–761 (2007).
6. Sicot, G., Gourdon, G. & Gomes-Pereira, M. Myotonic dystrophy, when simple repeats reveal complex pathogenic entities: new findings and future challenges. *Human Molecular Genetics* **20**, R116–R123 (2011).
7. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).
8. Will, C. L. & Lührmann, R. Spliceosome Structure and Function. *Cold Spring Harb Perspect Biol* **3**, (2011).
9. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
10. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology* **6**, 386–398 (2005).
11. Sammeth, M., Foissac, S. & Guigó, R. A General Definition and Nomenclature for Alternative Splicing Events. *PLoS Comput Biol* **4**, (2008).
12. Wang, Z. & Burge, C. B. Splicing Regulation: From a Parts List of Regulatory Elements to an Integrated Splicing Code. *RNA* **14**, 802–813 (2008).
13. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* **11**, 345–355 (2010).
14. Schmucker, D. *et al.* Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671–684 (2000).
15. Graveley, B. R. Mutually Exclusive Splicing of the Insect Dscam Pre-mRNA Directed by Competing Intronic RNA Secondary Structures. *Cell* **123**, 65–73 (2005).
16. Van der Ploeg, L. H. Discontinuous transcription and splicing in trypanosomes. *Cell* **47**, 479–480 (1986).
17. Hastings, K. E. M. SL trans-splicing: easy come or easy go? *Trends in Genetics* **21**, 240–247 (2005).
18. Pettitt, J., Harrison, N., Stansfield, I., Connolly, B. & Müller, B. The evolution of spliced leader trans-splicing in nematodes. *Biochem. Soc. Trans.* **38**, 1125–1130 (2010).
19. Dorn, R., Reuter, G. & Loewendorf, A. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9724–9729 (2001).

20. Labrador, M. *et al.* Protein encoding by both DNA strands. *Nature* **409**, 1000 (2001).
21. Horiuchi, T., Giniger, E. & Aigaki, T. Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, lola. *Genes Dev.* **17**, 2496–2501 (2003).
22. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64**, 607–613 (1991).
23. Cocquerelle, C., Daubersies, P., Majérus, M. A., Kerckaert, J. P. & Bailleul, B. Splicing with inverted order of exons occurs proximal to large introns. *EMBO J.* **11**, 1095–1098 (1992).
24. Zaphiropoulos, P. G. Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6536–6541 (1996).
25. Zaphiropoulos, P. G. 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 (1997).
26. Caldas, C. *et al.* Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene. *Gene* **208**, 167–176 (1998).
27. Caudevilla, C. *et al.* Natural Trans-Splicing in Carnitine Octanoyltransferase Pre-mRNAs in Rat Liver. *PNAS* **95**, 12185–12190 (1998).
28. Frantz, S. A. *et al.* Exon Repetition in mRNA. *PNAS* **96**, 5400–5405 (1999).
29. Finta, C. & Zaphiropoulos, P. G. The Human CYP2C Locus: A Prototype for Intergenic and Exon Repetition Splicing Events. *Genomics* **63**, 433–438 (2000).
30. Surono, A. *et al.* Circular Dystrophin RNAs Consisting of Exons That Were Skipped by Alternative Splicing. *Hum. Mol. Genet.* **8**, 493–500 (1999).
31. Akopian, A. N. *et al.* Trans-splicing of a voltage-gated sodium channel is regulated by nerve growth factor. *FEBS Letters* **445**, 177–182 (1999).
32. Crawford, J. *et al.* ThePISSLREGene: Structure, Exon Skipping, and Exclusion as Tumor Suppressor in Breast Cancer. *Genomics* **56**, 90–97 (1999).
33. Takahara, T., Kanazu, S.-I., Yanagisawa, S. & Akanuma, H. Heterogeneous Sp1 mRNAs in Human HepG2 Cells Include a Product of Homotypic Trans-Splicing. *J. Biol. Chem.* **275**, 38067–38072 (2000).
34. Hide, W. A., Babenko, V. N., Van Heusden, P. A., Seoighe, C. & Kelso, J. F. The Contribution of Exon-Skipping Events on Chromosome 22 to Protein Coding Diversity. *Genome Res.* **11**, 1848–1853 (2001).
35. Flouriot, G., Brand, H., Seraphin, B. & Gannon, F. Natural Trans-Spliced mRNAs Are Generated from the Human Estrogen Receptor-A (hERα) Gene. *J. Biol. Chem.* **277**, 26244–26251 (2002).
36. Rigatti, R., Jia, J.-H., Samani, N. J. & Eperon, I. C. Exon Repetition: A Major Pathway for Processing mRNA of Some Genes Is Allele-Specific. *Nucl. Acids Res.* **32**, 441–446 (2004).
37. Takahara, T., Tasic, B., Maniatis, T., Akanuma, H. & Yanagisawa, S. Delay in Synthesis of the 3′ Splice Site Promotes trans-Splicing of the Preceding 5′ Splice Site. *Molecular Cell* **18**, 245–251 (2005).
38. Shao, X., Shepelev, V. & Fedorov, A. Bioinformatic Analysis of Exon Repetition, Exon Scrambling and Trans-Splicing in Humans. *Bioinformatics* **22**, 692–698 (2006).

39. Dixon, R. J., Eperon, I. C., Hall, L. & Samani, N. J. A Genome-Wide Survey Demonstrates Widespread Non-Linear mRNA in Expressed Sequences from Multiple Species. *Nucl. Acids Res.* **33**, 5904–5913 (2005).
40. Al-Balool, H. H. *et al.* Post-Transcriptional Exon Shuffling Events in Humans Can Be Evolutionarily Conserved and Abundant. *Genome Res.* **21**, 1788–1799 (2011).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
42. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
43. Fu, G. C.-L. & Lin, W. Identification of gene-oriented exon orthology between human and mouse. *BMC Genomics* **13**, S10 (2012).
44. Fukuzawa, A., Idowu, S. & Gautel, M. Complete human gene structure of obscurin: implications for isoform generation by differential splicing. *Journal of Muscle Research and Cell Motility* **26**, 427–434 (2005).
45. Kontrogianni-Konstantopoulos, A., Ackermann, M. A., Bowman, A. L., Yap, S. V. & Bloch, R. J. Muscle Giants: Molecular Scaffolds in Sarcomerogenesis. *Physiol Rev* **89**, 1217–1267 (2009).
46. Hong, X., Scofield, D. G. & Lynch, M. Intron Size, Abundance, and Distribution Within Untranslated Regions of Genes. *Mol Biol Evol* **23**, 2392–2404 (2006).
47. Konarska, M. M., Padgett, R. A. & Sharp, P. A. Trans splicing of mRNA precursors in vitro. *Cell* **42**, 165–171 (1985).
48. Solnick, D. Trans splicing of mRNA precursors. *Cell* **42**, 157–164 (1985).
49. Reed, R. Coupling transcription, splicing and mRNA export. *Current Opinion in Cell Biology* **15**, 326–331 (2003).
50. Han, J., Xiong, J., Wang, D. & Fu, X.-D. Pre-mRNA splicing: where and when in the nucleus. *Trends in Cell Biology* **21**, 336–343 (2011).
51. McGlincy, N. J. & Smith, C. W. J. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends in Biochemical Sciences* **33**, 385–393 (2008).
52. Rebbapragada, I. & Lykke-Andersen, J. Execution of nonsense-mediated mRNA decay: what defines a substrate? *Current Opinion in Cell Biology* **21**, 394–402 (2009).
53. Nicholson, P. & Mühlemann, O. Cutting the nonsense: the degradation of PTC-containing mRNAs. *Biochemical Society Transactions* **38**, 1615 (2010).

Figure 1

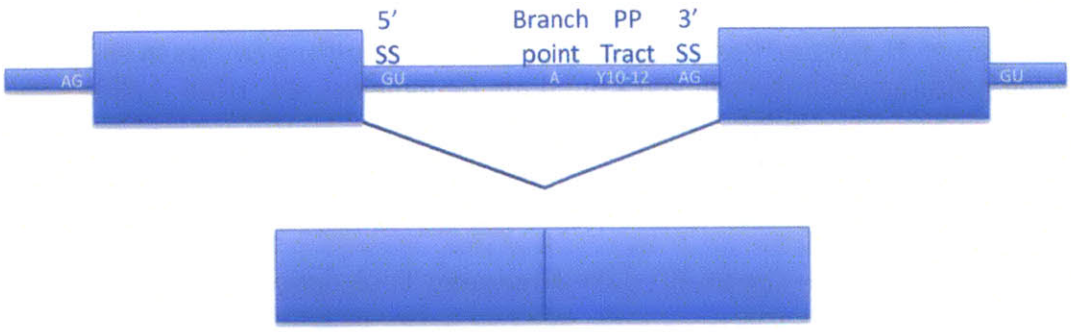


Figure 1. Canonical splicing.
There are three core splicing signals present in all introns: the 5' splice site (GU), the branch point (A) and the 3' splice site (AG). Additionally, a polypyrimidine tract located between the branch point and the 3' splice site aids in splice site recognition in higher eukaryotes.

Figure 2

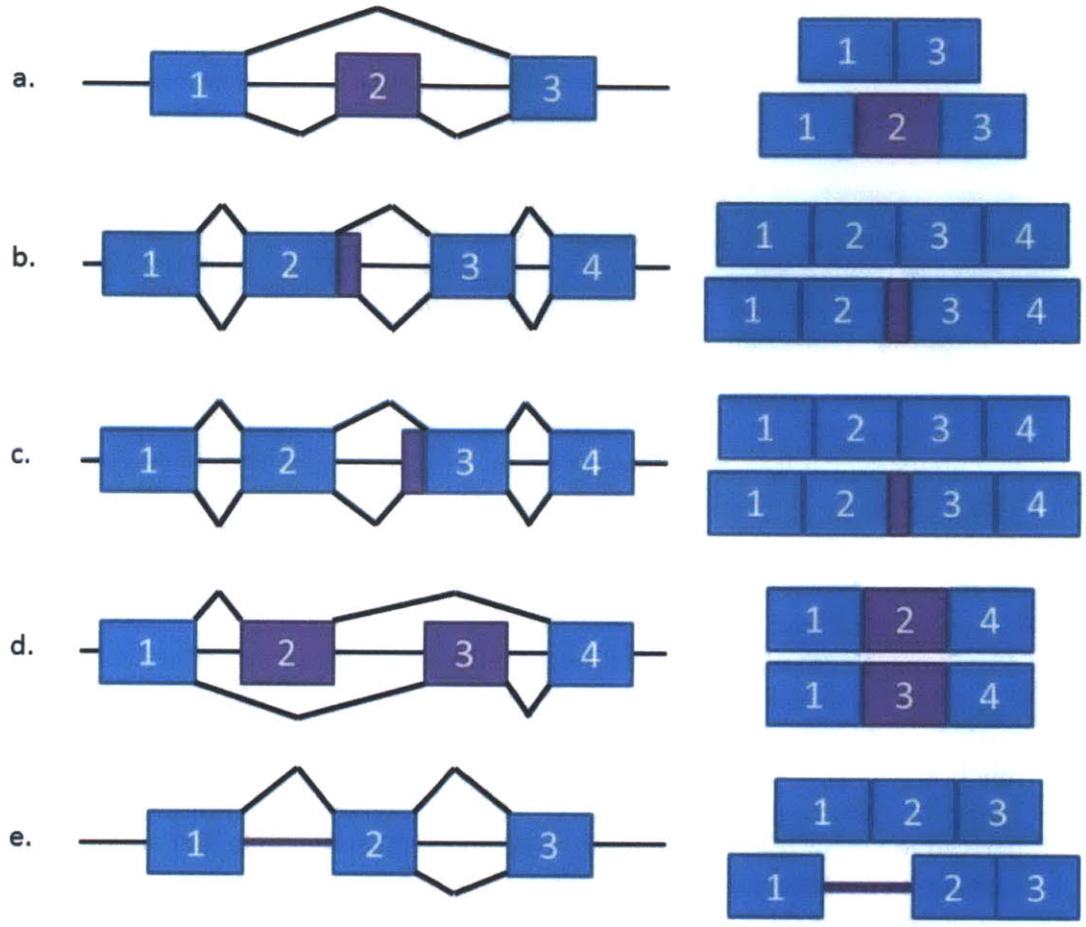


Figure 2. Alternative splicing falls into five categories:

a. Exon skipping, the most prevalent form of alternative splicing, results in the exclusion of an exon from the mature transcript.
b. Alternative 5' splice sites allow exclusion of just a terminal portion of an exon.
c. Alternative 3' splice sites regulate the inclusion or exclusion of the initial segment of an exon.
d. Mutually exclusive exons allow either one or the other (but not both or neither) of two neighboring exons to be included in the mature mRNA.
e. Intron retention causes an intronic sequence to be expressed in the final transcript.

Figure 3

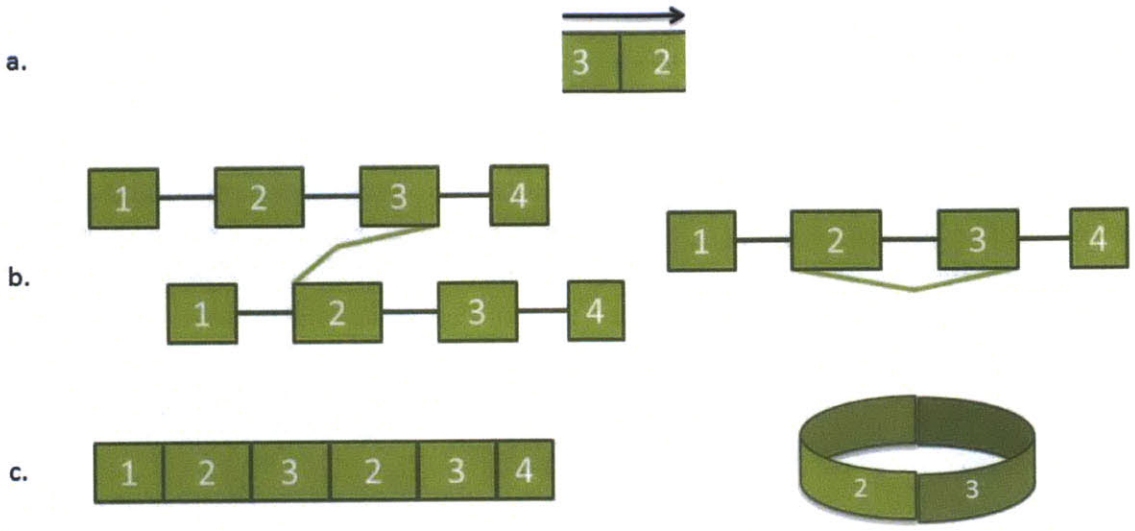


Figure 3. Exon scrambling.

a. A read maps across a scrambled junction from exon 3 to exon 2.
b. There are two possible splicing events that can explain the read: two separate pre-mRNAs are *trans*-spliced (left) or a single pre-mRNA is spliced in the reverse junction (right).
c. There are two possible mature mRNAs from which the read could have been derived: a longer transcript with exons 2 and 3 appearing in tandem duplication (left) or a circular read (right).

Figure 4

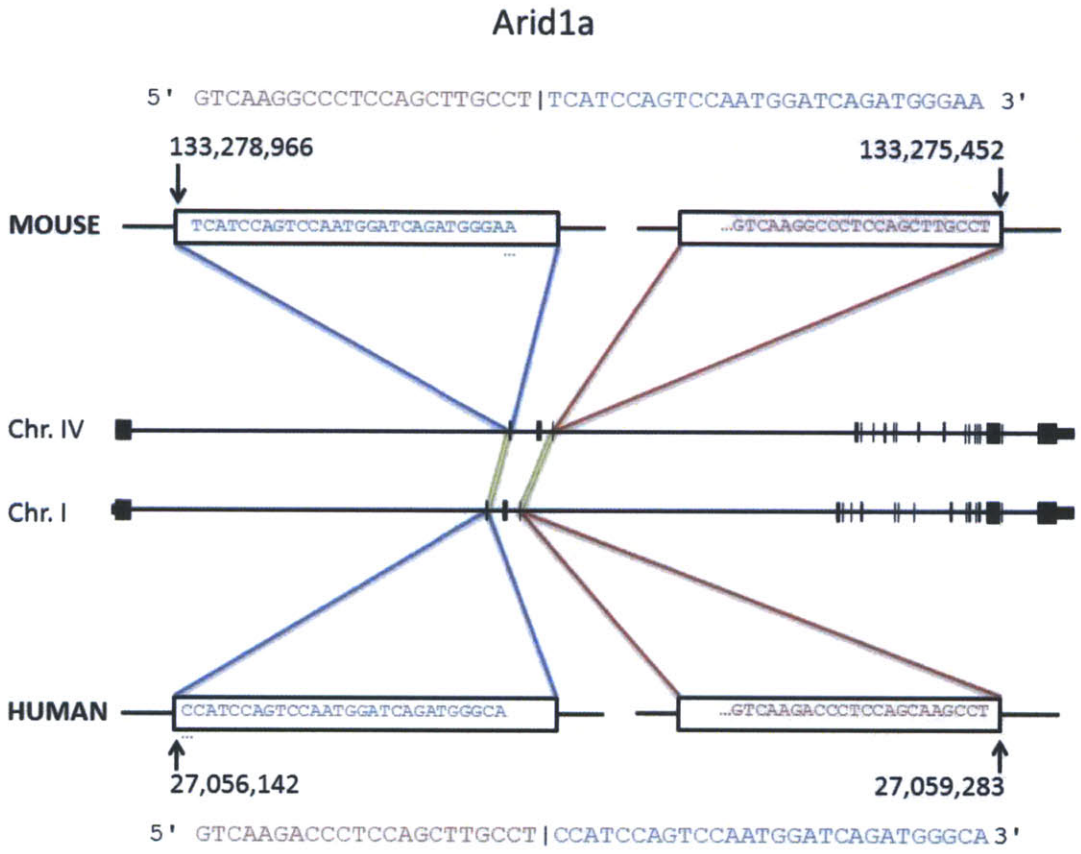


Figure 4. Arid1a, AT rich interactive domain 1A
One of the eight conserved genes exhibiting exon scrambling is Arid1a. In both mouse (top) and human (bottom), the first half of the read maps to the 3' end of exon four and the second half of the read maps to the 5' end of exon two. Both exons involved are conserved (indicated by the double green lines). Neither read contains a mismatch and the exon-exon junctions falls fairly close to the middle, thus increasing our confidence in this event.

Figure 5

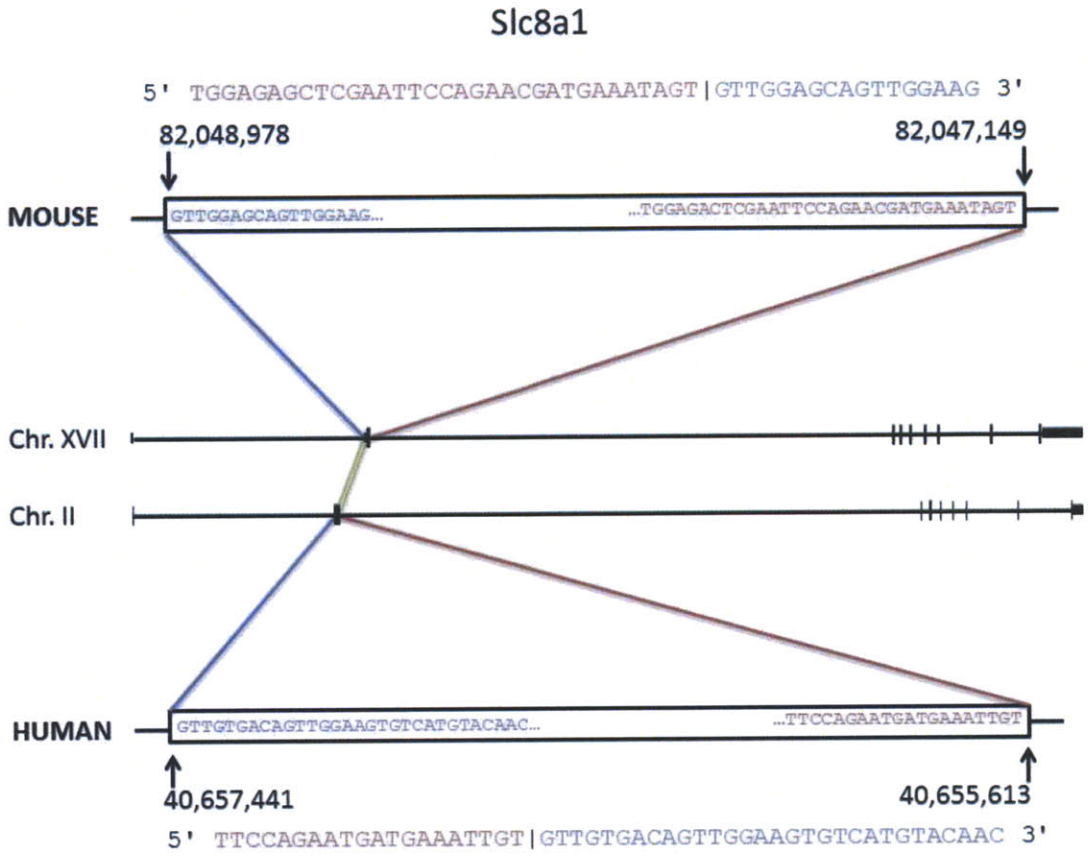


Figure 5. Slc8a1, Solute carrier family 8 (sodium/calcium exchanger) member 1
One of the eight conserved genes exhibiting exon scrambling is Slc8a1. In both mouse (top) and human (bottom), the first half of the read maps to the 3' end of exon two and the second half of the read maps to the 5' end of the same exon two. Exon two is orthologous between human and mouse (indicated by the double green lines).

Figure 6

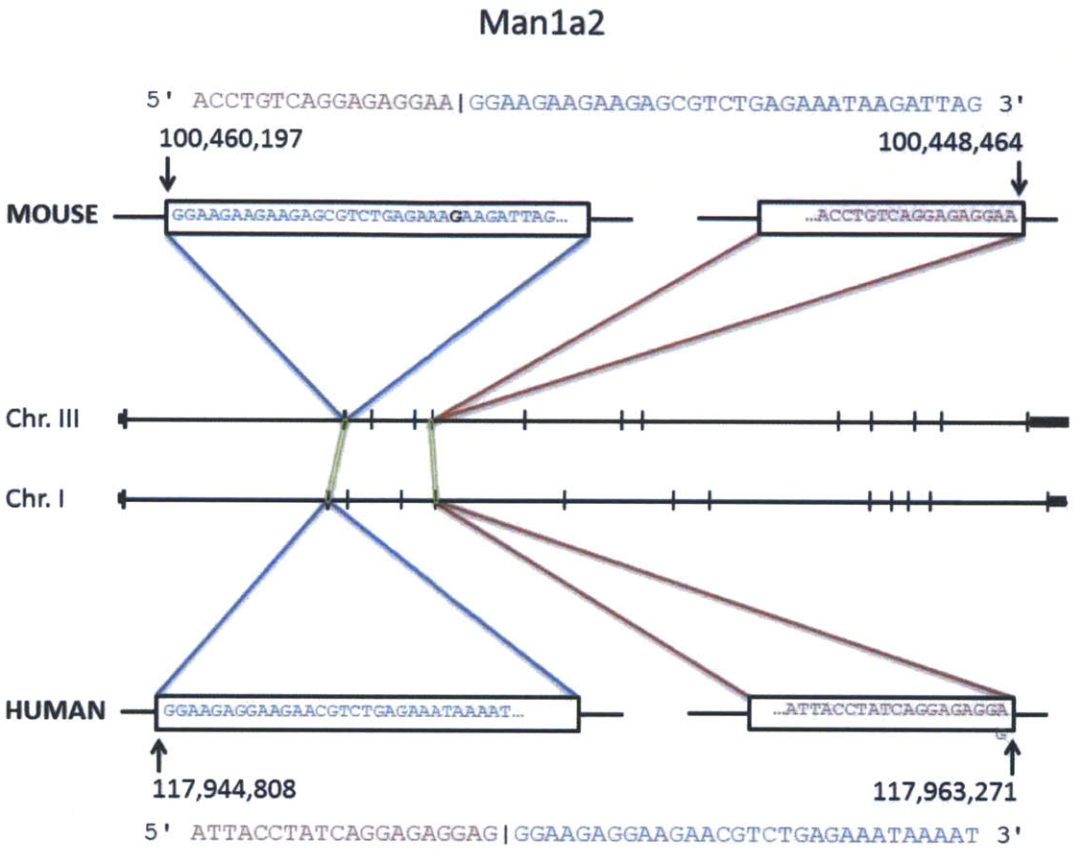


Figure 6. Man1a2, Mannosidase alpha, class 1A, member 2
One of the eight conserved genes exhibiting exon scrambling is Man1a2. In both mouse (top) and human (bottom), the first half of the read maps to the 3' end of exon five and the second half of the read maps to the 5' end of exon two. Both exons involved are orthologous between mouse and human (indicated by the double green lines).

Figure 7

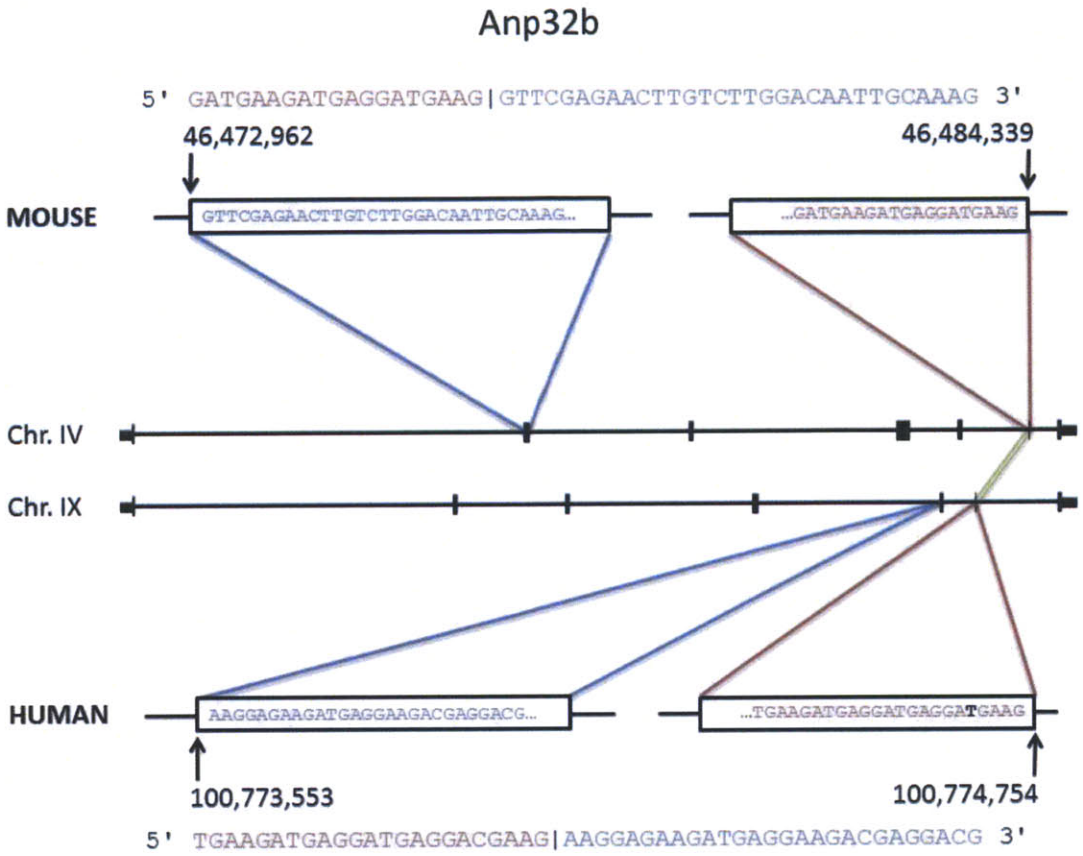


Figure 7. Anp32b, Acidic (leucine-rich) nuclear phosphoprotein 32 family, member B
One of the eight conserved genes exhibiting exon scrambling is Anp32b. Only one half of the scrambled exon-exon junction is conserved between mouse and human (indicated by the double green line). In both mouse (top) and human (bottom), the first half of the read maps to the 3' end of exon six; however, the second half of the mouse read maps to the 5' end of exon two, whereas the second half of the human read maps to the 5' end of exon five.

Figure 8

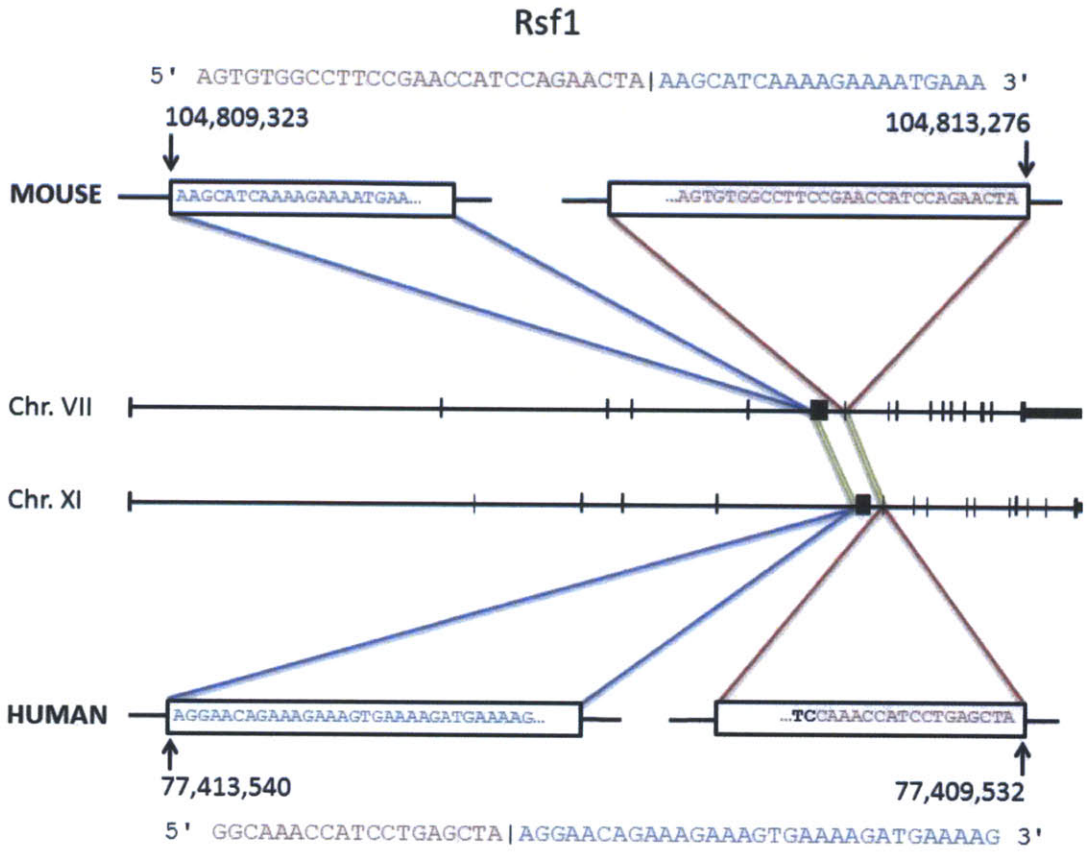


Figure 8. Rsf1, Remodeling and spacing factor 1
One of the eight conserved genes exhibiting exon scrambling is Rsf1. In both mouse (top) and human (bottom), the first half of the read maps to the 3' end of exon seven and the second half of the read maps to the 5' end of exon six. Both exons involved are conserved (indicated by the double green lines).

Figure 9

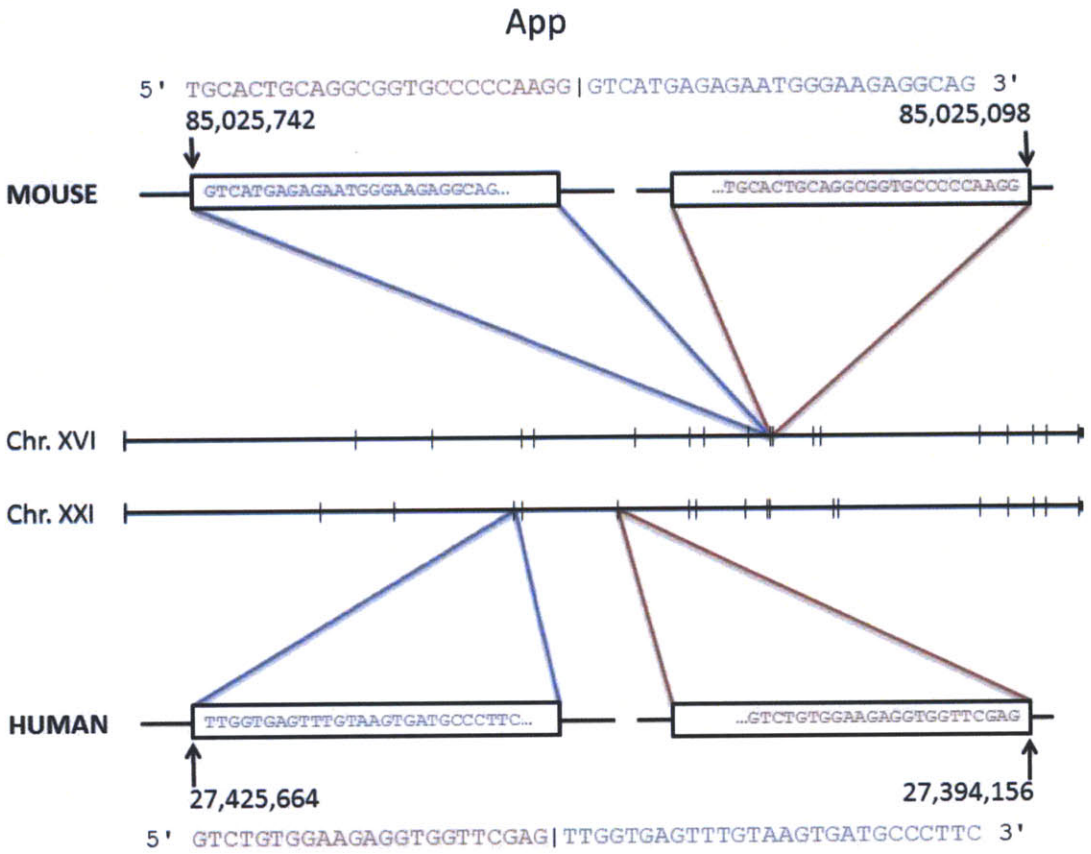


Figure 9. App, Amyloid beta (A4) precursor protein
One of the eight conserved genes exhibiting exon scrambling is App. Neither half of the scrambled exon-exon junction shares orthology between mouse and human. In mouse (top), the first half of the read maps to the 3' end of exon eleven and the second half of the read maps to the 5' end of exon ten; in human (bottom), the first half of the read maps to the 3' end of exon six and the second half of the read maps to the 5' end of exon four.

Figure 10

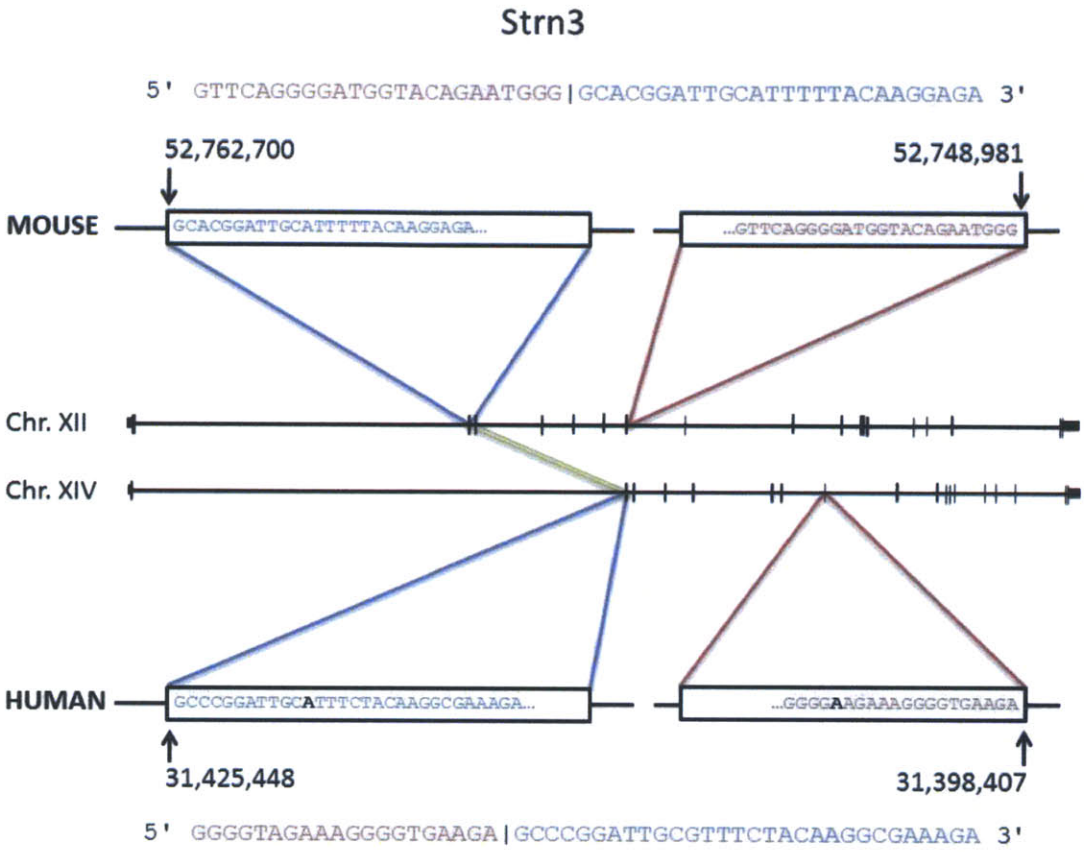


Figure 10. Strn3, Striatin, calmodulin binding protein 3
One of the eight conserved genes exhibiting exon scrambling is Strn3. Only one half of the scrambled exon-exon junction is conserved between mouse and human (indicated by the double green line). In both mouse (top) and human (bottom), the second half of the read maps to the 5' end of exon two; however, the first half of the mouse read maps to the 3' end of exon seven, whereas the second half of the human read maps to the 3' end of exon eight.

Figure 11

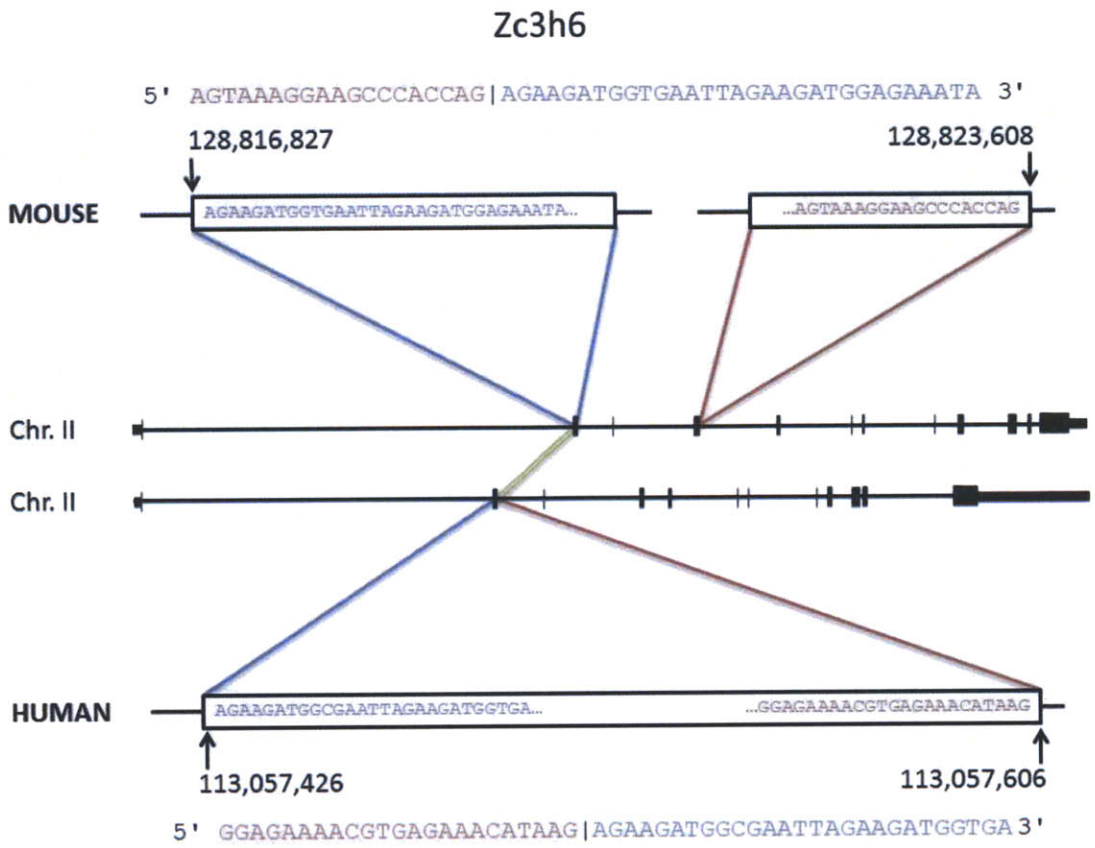


Figure 11. Zc3h6, Zinc finger CCCH type containing 6
One of the eight conserved genes exhibiting exon scrambling is Zc3h6. Only one half of the scrambled exon-exon junction is conserved between mouse and human (indicated by the double green line). In both mouse (top) and human (bottom), the second half of the read maps to the 5' end of exon two; however, the first half of the mouse read maps to the 3' end of exon four, whereas the second half of the human read maps to the 3' end of exon two.

Figure 12

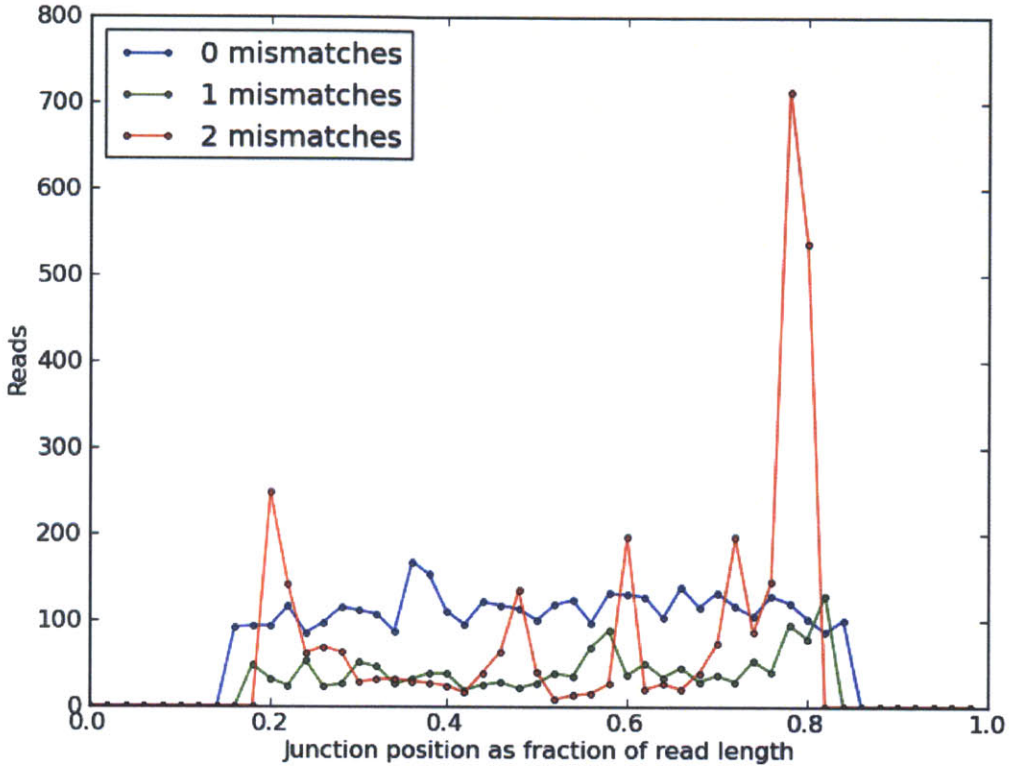


Figure 12. Estimating false positives.
In order to estimate the number of false positives in our putative scrambled exon set, the junction positions were plotted to see where they fell along the reads. Reads were binned by number of mismatches (two mismatches was the maximum allowed). A particularly high peak can be seen at about 80% of the read length in the two mismatches bin (red). This peak probably consists of false positives since the occurrence of two mismatches in ten bases of overhang is likely to be the result of a mis-mapping.

Table 1

	MOUSE							HUMAN						
Name	ID	Chr	+/-	5' site	Intron	Intron	3' site	RF	ID	Chr	+/-	5' site	Intron	Intron	3' site	RF
Arid1a	ENSMUSG00000007880	4	-	**133275452**	23381	29418	**133278966**	Y	ENSG00000117713	1	+	**27059283**	28064	32111	**27056142**	Y
Rsf1	ENSMUSG00000035623	7	+	**104813276**	5003	7606	**104809323**	Y	ENSG00000048649	11	-	**77409532**	4876	23050	**77413540**	N
Slc8a1	ENSMUSG00000054640	17	-	**82047149**	202216	88671	**82048978**	Y	ENSG00000183023	2	-	**40655613**	250618	23104	**40657441**	N
Man1a2	ENSMUSG00000008763	3	-	**100448464**	11943	28287	**100460197**	Y	ENSG00000198162	1	+	**117963271**	21582	33701	**117944808**	Y
Strn3	ENSMUSG00000020954	12	-	52748981	4780	29709	**52762700**	N	ENSG00000196792	14	-	31398407	10095	69662	**31425448**	N
Zc3h6	ENSMUSG00000042851	2	+	28823608	4270	23255	**128816827**	N	ENSG00000188177	2	+	113057606	3200	23823	**113057426**	Y
App	ENSMUSG00000022892	16	-	85025098	9493	4735	85025742	Y	ENSG00000142192	21	-	27394156	21659	36595	27425664	Y
Anp32b	ENSMUSG00000028333	4	+	**46484339**	620	8741	46472962	N	ENSG00000136938	9	+	**100774754**	2892	6118	100773553	Y

Table 1. Conserved scrambled genes.
Genes with exon scrambling events that passed the filters in both mouse and human are listed above. The first column lists the gene name. Next is listed the ENSEMBL gene ID. Following this are the junction position identifiers: chromosome number, strand, 5' splice site genomic coordinate, length of intron downstream of 5' splice site, length of intron upstream of 3' splice site, and 3' splice site genomic coordinate. The last column tells whether the reading frame has been preserved between the normal downstream junction and the scrambled upstream junction ('Y' = preserved, 'N' = not preserved). Orthology of exons between human and mouse is indicated by boldface type. For example, for the gene Anp32b, only the 5' exon involved in each scrambling event is orthologous.