Bioinformatic Investigation of the Peculiarities of Long Intron Splicing in *Hominidae*

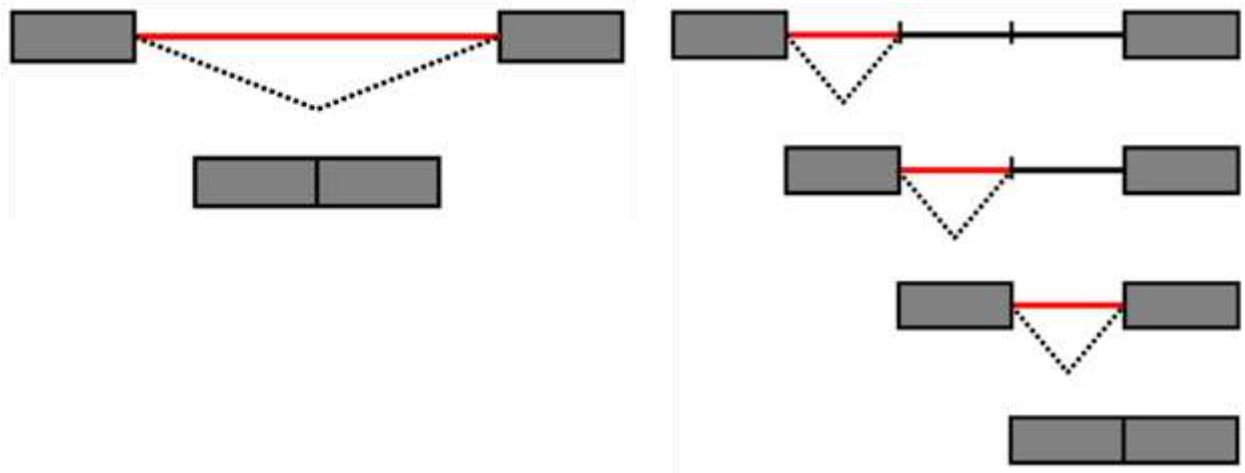
Saniya J. Gaitonde

W. Tresper Clarke High School

Introduction

RNA transcripts are usually matured by intron excision in a single lariat (rope-like) unit, followed by the joining of exon sequences, during a two-step catalytic reaction. Recursive splicing, a process by which intronic sequences are removed from pre-mRNA transcripts in multiple distinct segments, has been observed in a small subset of *Drosophila melanogaster* introns through novel computational approaches. The detection of recursive splicing requires observation of splicing intermediates that are inherently unstable, making it difficult to study by other laboratory means. Recursive sites were found in most very long (> 40 kb) fly introns, including many genes involved in morphogenesis and development, and tend to occur near the midpoints of introns. Suggesting a possible function for recursive splicing, it was observed that fly introns with recursive sites are spliced more accurately than comparably sized non-recursive introns (Pai et al., 2018).

Figure 1: Conventional splicing (left) vs. recursive splicing (right) / Hubé et al., 2015

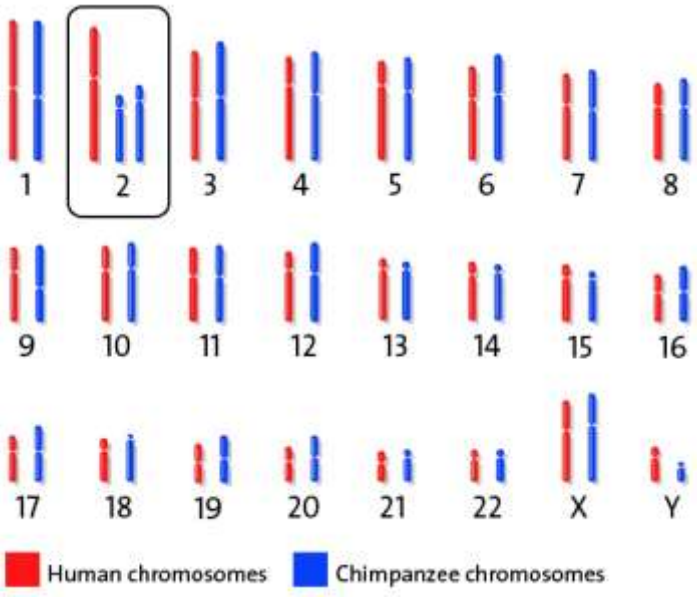


In the predecessor to this investigation, a detailed bioinformatic analysis of the recursive motifs mediating multi-step RNA splicing in several members of the *Drosophila* genus, (including *D. melanogaster*, *D. yakuba*, *D. simulans*, and *D. pseudoobscura*), revealed that the

conservation of recursive splicing is generally consistent with evolutionary relationships, with recursive sequences found most abundantly in introns.

Many basic biological, physiological, and neurological properties are conserved among *D. melanogaster* and mammals, and nearly 75% of human disease-causing genes are believed to have a functional homolog in the fly (Pandey et al., 2011). DNA evidence suggests that the bonobo (*Pan paniscus*) and common chimpanzee (*Pan troglodytes*) species diverged from each other approximately 0.86 to 0.89 million years ago. Both *Pan* species, cladistically, are equally close to humans (*Homo sapiens*), sharing approximately 98% of genes. The *Pan* line split from the last common ancestor shared with humans approximately 6 to 7 million years ago, during which it is theorized that human chromosome 2 was formed by the end-to-end fusion of two smaller ancestral chromosomes, 2a and 2b, based on identified remnants of a second centromere and excess vestigial telomeres on human chromosome 2. Orangutans, gorillas, chimpanzees, and bonobos all have 24 chromosome pairs, while humans and our closest extinct ancestors (Neanderthals and Denisovans) have 23 (Mooney, 2017).

Figure 2: Chromosomes of human and chimpanzee, aligned for comparison / NCBI / (Mooney, 2017)



A study of over 1500 microRNAs to identify intronic/non-coding variation between humans and other great apes found that two microRNAs which are highly expressed in brain tissue and may exert effects on genes with neural functions, in addition to two other microRNAs which have a role in development, were specific in sequence and length to humans (Gallego et al., 2016).

The purpose of this investigation is to evaluate the potential of recursive splicing in *Homo* and *Pan* genomes in order to gain novel insight about long intron splicing in *Hominidae* and the genetic basis of developmental variation among humans and other Great Apes. It was hypothesized that motifs for recursive sites, which mediate multi-step RNA splicing, are enriched and conserved in the introns of genes with key developmental functions, consistent with evolutionary relationships.

Methods

Recursive splicing can be detected by computationally parsing genomic DNA for recursive sequences. A python program was written to scan and score genomic sequences for matches to recursive motifs, using known regulatory sites as controls. The script accepts a position-specific scoring matrix (PSSM) file along with any organismal genome in FASTA file format and outputs data in the form of a BED file which can be used for analysis. A PSSM is a numerical representation of a motif with each column representing the probability that each position has a given base.

Figure 3: Example PSSM

	pos1	pos2	pos3	pos4
A	0.1	0.5	0	1
C	0.1	0	0.5	0
G	0.1	0.5	0	0
T	0.7	0	0.5	0

A PSSM for known recursive sites was used, curated based on previous *Drosophila* studies. FASTA is a commonly used file type that holds the sequence of chosen genomic regions. For example, each entry in an entire genome FASTA represents a chromosome.

Figure 4: Example genome FASTA

```
>chr1
ATCGCTAAAGTCTA
>chr2
TGCTAATCGGTCAGTC
>chr3
GCTATATATCGCGGCTAA
```

Genome FASTA files for *H. sapiens*, *P. troglodytes*, and *P. paniscus* were retrieved from the UCSC database, the official database of the Human Genome Project. In this study, scaled samples of human chromosome 2, chimpanzee chromosomes 2a and 2b, and bonobo chromosomes 2a and 2b sequences were used for parsing. BED is the most widely used file format for bioinformatic genome analysis, to delineate genomic regions that are of interest. Each row is a region and each column contains pertinent information. In the outputs retrieved by the program used in this study, each row is a 26 base pair segment of the FASTA chromosome within which there is potential for the PSSM's recursive motif to fall. The last column indicates which strand of the DNA the gene was transcribed from: "+" indicates that the motif is as written in the genome file, while the "–" indicates that the motif is the reverse complement to that location in the genome.

Figure 5: Example BED file

```
column1: chromosome
column2: start of the region
column3: end of the region
column4: name of the region (user determined)
column5: score of the region (often "." when no score is needed)
column6: strand off of which the region is transcribed (see above)
```

After the program was executed with the recursive PSSM for each sample, column 5, delineating the score of each 26 base pair region's match to the recursive sequence, was used to determine the average match score for each sample. The BED files for chimpanzee chromosomes 2a and 2b were then consolidated into one file to make for a scaled comparison to human chromosome 2, and the same was completed for bonobo chromosomes 2a and 2b. WebLogo was then used to materialize the appearance of the recursive sequences in order to determine the extent to which the sequences found in each species resemble each other and the recursive PSSM. The "sequence" column from each BED file was pulled and inputted into the software, (with adjustments made as specified in the formatting guide), which then produced a motif sequence logo with the letters at each position indicating the nucleotides that are observed at the given position, and the height indicating the probability (in bits) of observing that nucleotide. As a control and basis for comparison, a motif sequence logo was also created for the recursive PSSM input.

Figure 6: Sequence logo of recursive PSSM, built by WebLogo

LiftOver analysis was then employed to investigate whether the location of the recursive sites found in each species has been conserved. The original BED outputs were directly uploaded to the online LiftOver tool, which converted the genomic coordinates of each species' sequences to a common reference, in order to access whether there is evidence of overlap. The converted BED files were then inputted into a pre-programmed python LiftOver script, which outputted the total

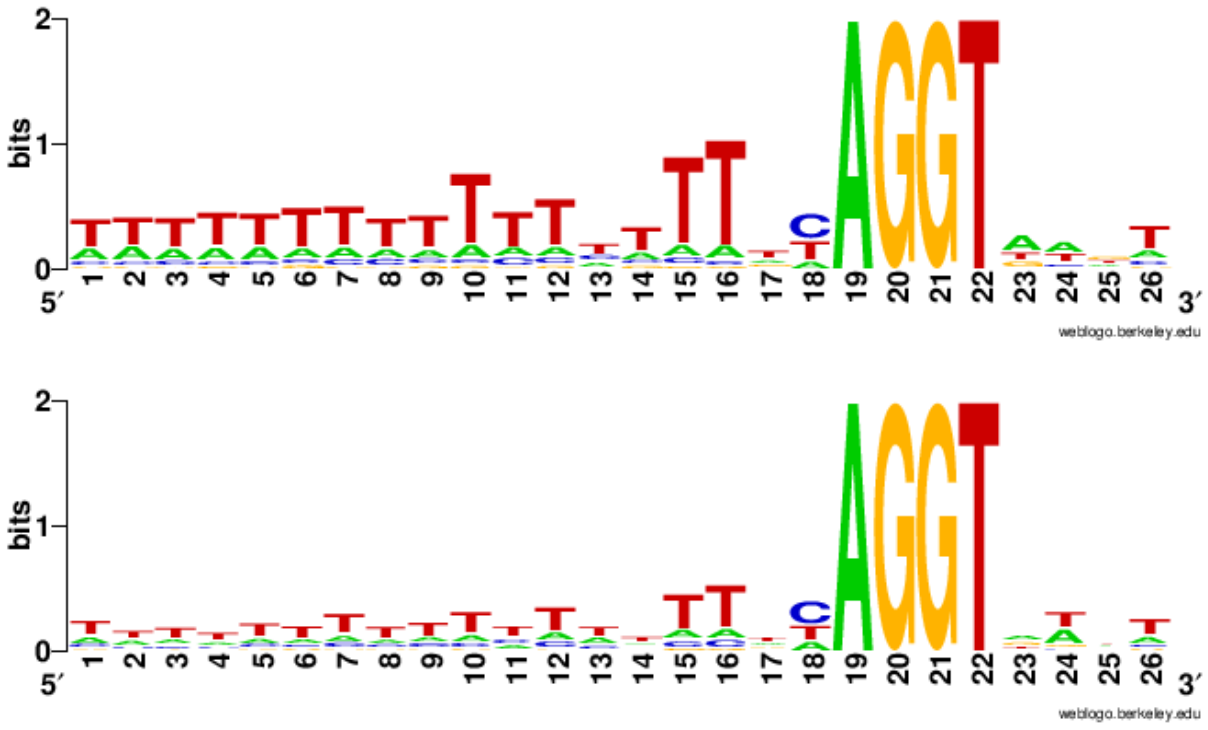
number of motif matches in each species and how many overlapped in both directions for each combination of two species.

Results

Table 1: Average match scores of each sample's 26 base pair sequences to the recursive PSSM

Sample	Average match score of 26 base pair sequences to recursive PSSM
D. melanogaster full genome (control)	12.68
Human chromosome 2	12.56
Chimpanzee chromosome 2a	12.58
Chimpanzee chromosome 2b	12.36
Chimpanzee 2a + 2b	12.47
Bonobo chromosome 2a	12.58
Bonobo chromosome 2b	12.36
Bonobo 2a + 2b	12.45

Figure 7-10: Motif WebLogos for recursive sequences (top to bottom) D. melanogaster, human chr 2, chimpanzee chr 2a + 2b, bonobo chr 2a + 2b



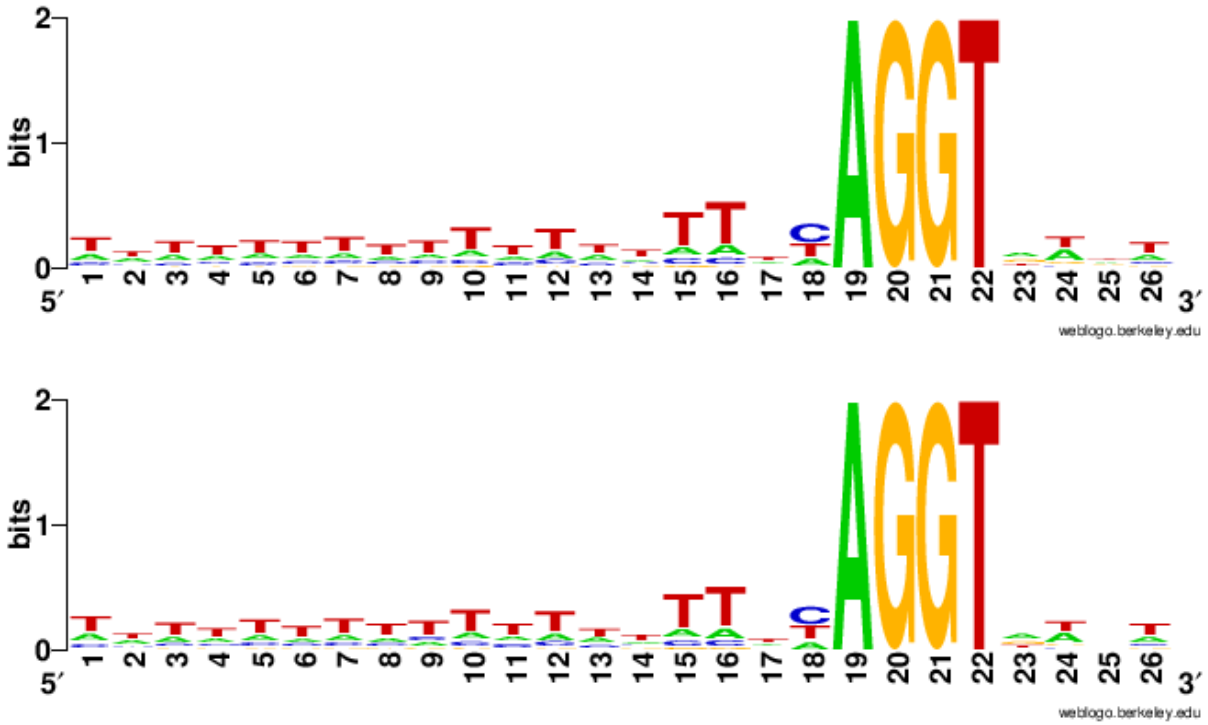
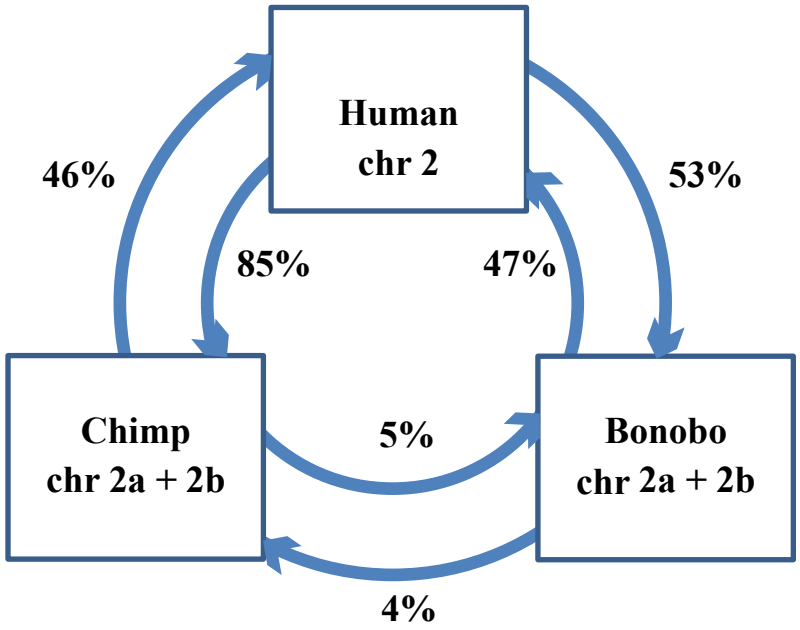


Figure 11: LiftOver percentages



The results displayed in Table 1 show that all the samples parsed in this investigation had extremely similarly scoring sequences to each other and *Drosophila*, whose key developmental genes have known recursive sites enriched in introns, indicating that the potential for similar

patterns in Hominidae is promising. Bonobo chromosome 2a had the highest match score of its 26 base pair sequences to the recursive PSSM (~12.580), followed by chimpanzee chromosome 2a (~12.575). In contrast, bonobo chromosome 2b had a relatively low average score of ~12.358, as did chimpanzee chromosome 2b, which had an average score of ~12.364. The sample from human chromosome 2 had an average score of ~12.556, which most closely resembled the average scores of chimpanzee chromosome 2a and 2b together (~12.472), and bonobo chromosomes 2a and 2b together (~12.451). The average score for the complete *Drosophila melanogaster* genome was included in the table to serve as a basis for comparison as it contains known recursive sites; this score was calculated in the predecessor of this investigation to be (~12.679). Overall, this data is largely consistent with known evolutionary relationships, supporting that the relationship between *Drosophila* and the *Pan* genus is more closely linked than that which exists between *Drosophila* and *Homo*. The finding that chimpanzee and bonobo chromosomes 2a had the highest average scores suggests that the majority of recursive sites on human chromosome 2 were contributed by the ancestral chromosome 2a during the end-to-end fusion event. Perhaps, even, the fusion event could have been favored by natural selection partly to offset the significant difference in concentration of recursive sites between chromosomes 2a and 2b, supporting further the hypothesis that recursive splicing functions to provide a more accurate mechanism for pre-mRNA transcript maturation.

The WebLogo sequence analysis revealed that although subtle differences exist, the recursive sequences of the three *Hominidae* species resemble each other far more closely than they resemble the *Drosophila* recursive sequences and the recursive PSSM. This is most apparent at nucleotide position 24, where the probability is highest of observing an Adenine base in *D. melanogaster*, and a Thymine base in chimpanzee, bonobo, and human. When critically

analyzing the motif sequence logos, there are several instances of variation among the species that stray from what may be preconceived. For example, at position 7, the human sequence logo resembles the *D. melanogaster* logo more closely than the chimpanzee and bonobo sequence logos do, and at position 25, the patterns for human and bonobo exhibit the most similarity. Holistically, the three *Hominidae* sequence logos resembled the *D. melanogaster* logo far more than they resembled the PSSM, which, likewise, was far more enriched in *D. melanogaster*. The latter is because the PSSM used was created and generalized based on known *Drosophila* recursive sequences, and itself used *Drosophila* regulatory sites as control. Therefore, it is likely alternative splicing sites may exist in *Hominidae* that are independent from those in *Drosophila* and therefore those returned by analysis with the PSSM that was used in this investigation. In support of this hypothesis, studies suggest that recursive sites in invertebrates such as *Drosophila* are characterized by consecutive splicing from the 5′-end at a series of combined donor-acceptor splice sites called RP-sites, while vertebrates may lack the proper enrichment of RP-sites in their large introns, and, therefore, require some other mechanism to aid splicing (Shepard et al., 2009).

The LiftOver analysis returned that 589 out of 690 (~85%) motif matches found in human chromosome 2 were also found in chimpanzee chromosome 2a or 2b, and 586 out of 1,262 (~46%) motif matches found in chimpanzee chromosome 2a or 2b were also found in human chromosome 2. 370 out of 690 (~53%) motif matches found in human chromosome 2 were also found in bonobo chromosome 2a or 2b and 368 out of 785 (~47%) motif matches found in bonobo chromosome 2a or 2b were also found in human chromosome 2. Interestingly, only 32 out of 661 (~5%) motif matches found in chimpanzee chromosomes 2a and 2b were also found in bonobo chromosomes 2a and 2b, and similarly, only 31 out of 760 (~4%) motif matches found in bonobo chromosomes 2a and 2b were also found in chimpanzee chromosomes 2a and

2b. This reveals that chimpanzee chromosomes 2a and 2b had the highest number of total motif matches to the recursive PSSM, and that the coordinates of the recursive sites found in human chromosome 2 are most similar to those of the common chimpanzee. Coupled with the evidence that both chromosomes 2a had the highest average score for the match of their 26 base pair sequences to the recursive PSSM, this makes for a strong case that chromosome 2a is the richest in recursive sites, many of which were naturally selected for *Homo sapiens* during an ancestral fusion event causing the diversion from the last common ancestor shared with the *Pan* line. The significantly low LiftOver percentages for chimpanzee and bonobo indicate that, unless due to a processing error, the coordinates of the recursive sites in the species' chromosomes 2a and 2b are vastly different, which contradicts evidence of their close cladistic relationship. LiftOver did not permit analyses between the *Hominidae* species and *Drosophila* due to genomic incompatibility.

Serving as a limitation to this experiment, in addition to the fact that the PSSM used was developed based on *Drosophila* patterns as discussed, is the very small sample size for each *Hominidae* species that was accessible. The *Drosophila melanogaster* genome is a mere 5% of the length of the human genome, and therefore, in order to scan any entire *Hominidae* genome, the script would have to be optimized for a personal computer with limited processing power and memory. Thus, the investigation was limited to chromosome 2, due to the evolutionary basis for comparison throughout the taxonomic family. However, chromosome 2 is the second-largest human chromosome, spanning more than 242 million base pairs and representing almost eight percent of the total DNA in human cells, so only scaled samples of human chromosome 2, chimpanzee chromosomes 2a and 2b, and bonobo chromosomes 2a and 2b were used. This makes for a very limited dataset to draw conclusions about the pervasiveness of recursive splicing in entire *Hominidae* genomes, but still serves to illustrate its potential through

extrapolation and evolutionary development. Additionally, in order to determine whether the recursive sequences found are enriched in introns, exons, or non-coding regions, the BED files for each species would have had to be run in the python annotationoverlap.py program which requires a second input file delineating which regions of the genome FASTA files correspond to introns and exons; these inputs were not accessible for the *Hominidae* species differentiable by chromosome. However, several such analyses with entire Drosophila genomes have indicated that recursive sites are most enriched at the midpoints of long introns of genes with key developmental functions, and the finding that *Drosophila* and *Hominidae* recursive sites had very similar match scores to the recursive PSSM supports that enrichment patterns are most likely consistent as well.

To expand the sample size to include full chromosome FASTAs or entire genomes, an optimized version of the program used or a high-speed processing computer would be necessary. The data collected can be used to develop a PSSM which is more suited to alternative splicing in humans, analysis with which can provide novel insight into disorders resulting from RNA implications and erroneous protein synthesis. Performing a similar analysis on human DNA from disorders characterized by implicated protein synthesis could reveal a relationship between the origins of such disorders and alternative splicing.

References

Gallego, A., Melé, M., Balcells, I., García-Ramallo, E., Torruella-Loran, I., Fernández-Bellon, H., … Espinosa-Parrilla, Y. (2016, April 22). Functional Implications of Human-Specific Changes in Great Ape microRNAs. Retrieved January 18, 2020, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154194

Mooney, C. (2017, June 24). You share 98.7 percent of your DNA with this sex-obsessed ape. Retrieved from https://www.motherjones.com/politics/2014/02/evolution-creationism-bonobos-neanderthals-denisovans-chromosome-two/

Pai, A. A., Paggi, J. M., Yan, P., Adelman, K., & Burge, C. B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. *PLOS Genetics*, *14*(8). doi: 10.1371/journal.pgen.1007588

Pandey, U. B., & Nichols, C. D. (2011). Human Disease Models in Drosophila melanogaster and the Role of the Fly in Therapeutic Drug Discovery. *Pharmacological Reviews*, *63*(2), 411–436. doi: 10.1124/pr.110.003293

Shepard S, McCreary M, Fedorov A (2009) The Peculiarities of Large Intron Splicing in Animals. PLoS ONE 4(11): e7853. https://doi.org/10.1371/journal.pone.0007853