Probing the mechanisms of intron creation in a fast-evolving mite

Scott William Roy
Department of Biology
San Francisco State University
1600 Holloway Ave
San Francisco, CA 94132.

Available genomic sequences from diverse eukaryotes attest to creation of millions of spliceosomal introns throughout the course of evolution, however the question of how introns are created remains unresolved. Resolution of this question has been thwarted by the fact that many modern introns appear to be hundreds of millions of years old, obscuring the mechanisms by which they were initially created. As such, analysis of lineages undergoing rapid intron creation is crucial. Recently, Hoy et al. reported the genome of the predatory mite *Metaseiulus occidentalis*, revealing generally rapid molecular evolution including wholesale loss of ancestral introns and gain of new ones. I sought to test several potential mechanisms of intron creation. BLAST searches did not reveal patterns of similarity between intronic sequences from different sites or between intron sequences and non-intronic sequences, which would be predicted if introns are created by propagation of pre-existing intronic sequences or by transposable element insertion. To test for evidence that introns are created by any of multiple mechanisms that are expected to lead to duplication of sequences at the two splice boundaries of an intron, I compared introns likely to have been gained in the lineage leading to *M. occidentalis* and likely ancestral introns. These comparisons did initially reveal greater similarity between boundaries in *M. occidentalis*-specific introns, however this excess appeared to be largely or completely due to greater adherence of newer introns to the so-called 'protosplice' site, and therefore may not provide strong evidence for particular intron gain mechanisms. The failure to find evidence for particular intron creation mechanisms could reflect the relatively old age of even these introns, intron creation by variants of tested mechanisms that do not leave a clear sequence signature, or by intron creation by unimagined mechanisms.

The ubiquity of spliceosomal introns in eukaryotic nuclear genes and the diversity of intron positions across eukaryotic diversity attests to a huge number of intron creation events in the history of eukaryotes (Rogozin et al. 2003). In 2003, Alexei Fedorov and myself performed near-genome-scale comparisons of intron positions in orthologs, in hopes of identifying recently-created introns (Roy et al. 2003). Much to our surprise (and chagrin), we found a remarkably small number of changes: among 10,020 intron positions studied in a comparison of species diverged 80 million years ago (human and mouse), we found only five that were not shared between the species. Furthermore, all five of these were shared with outgroups, indicating intron loss and not gain (Roy et al. 2003). Other genome-wide and

smaller-scale studies have confirmed the general finding of striking degrees of intron position sharing suggestive of little intron gain in many different lineages over a variety of phylogenetic depths (Stajich and Dietrich 2005; Roy and Hartl 2006; Roy et al. 2006; Rogozin et al. 2003; Yang et al. 2013).

This finding of small numbers of intron gains in many lineages has greatly hindered our understanding of the mechanisms, phenotypic impacts and population genetics of intron-exon structures. However, a growing handful of studies have begun to fill in these long-standing gaps. Li et al. (2009, 2014) and Omilian et al. (2008) discovered dozens of recent intron creations in the water flea *Daphnia pulex* which appear to have arisen by imprecise double strand break repair (DSBR), and Farlow et al. (2011), Yenerall et al. (2011) and Sun et al. (2014) have provided evidence that a similar mechanism may create introns in species of *Drosophila* and *Neurospora.* Alex Worden's group and others have probed widespread creation of introns in the green alga *Micromonas pusilla* by propagation of a cryptic transposable element (Worden et al. 2009; van Baren et al. 2016; Simmons et al. 2016; Verhelst et al. 2013), and multiple groups have reported on a seemingly similar case in a clade of fungi (Collemare et al. 2013, 2015; van der Burgt et al. 2012). Curtis and Achibald (2010) reported a single intron gained by insertion of a non-intronic portion of mitochondrial DNA. We previously reported a variety of intron gain mechanisms in the fast-evolving chordate *Oikopleura*, including local propagation of short intron sequences by unknown mechanisms and intron creation by transposable element insertion (Denoeud et al. 2010). We also previously reported creation of introns by 'intronization,' i.e. splicing out of internal portions of ancestral exonic sequences (Irimia et al. 2007). Hellsten et al. 2011 reported intron creation by internal duplication of exonic sequencing and usage of two resulting copies of an AG|GT containing motif (Hellsten et al. 2011).

However, even while revealing much about intron creation, the diversity of mechanisms reported by these studies increases the importance of finding more cases, in order to understand the relative incidence and determinants of these different mechanisms across species. Here, I study the recently-reported genome of the predatory mite *M. occidentalis*, which has undergone widespread intron gain over long evolutionary timescales (Hoy et al. 2016).

Sequence similarity searches do not support intron creation by intron propagation or transposable element insertion

To test for the possibility of insertions of introns at new sites by propagation of pre-existing introns to new sites, I used BLAST comparisons between regions including intronic sequences and their flanking exonic sequence. The expected signature of intron propagation would be sequence similarity including the entire intron sequence, but not flanking exonic sequences. I downloaded the *M. occidentalis* genome and annotations from Genbank (GCF_000255335.1_Mocc_1.0), and extracted every annotated intronic sequence within the *M. occidentalis* genome along with 20 exonic nucleotides on either side (for 52,196 introns in total). I performed all-against-all BLASTN searches (with filtering of repetitive sequences turned off in order to promote extensions of BLAST hits through low-complexity regions) between these sequences, and identified hits that began and ended within

10 nucleotide positions of the splice boundaries for both the query and subject sequences. This identified only a single case (a pair of introns in genes XP_003744432.1 and XP_003744441.1), which upon manual alignment was revealed to involve longer homologous regions extending far beyond the ends the introns (Figure 1a). Thus this search revealed no case of inter-intron similarity that would be suggestive of intron propagation.

I next tested for evidence of intron creation by insertion of another genomic sequence (in particular a transposable element, although the case reported by Curtis and Archibald (2010) raises the possibility of a copy of a non-mobile sequence being inserted into a new genomic locus). Again, the expected sequence signature in this case would be a region sequence similarity corresponding closely to the boundaries of the intron. I performed a BLASTN search of all of the sequences generated above (intron plus 20 nucleotides of flanking exonic sequence) against the entire *M. occidentalis* genome and, as above, identified hits that began and ended within 10 nucleotide positions of both the query and subject sequences. This identified a total of 10 introns with genomic hits (several with multiple hits). As with the intron-intron case, for each of the 10 cases, manual scrutiny and alignment revealed multiple genomic copies extending well into the annotated exonic sequences, indicating that these cases do not reflect intron creation from a genomic sequence newly inserted into the interior of an exon. In addition, specific BLASTN searches against the mitochondrial genome revealed no similarity to nuclear intronic sequences.

Novel introns have stronger protosplice sites but do not have extended shared motifs at their boundaries

I next tested a prediction of several different hypotheses, namely similarities between the sequences spanning or flanking the two splice sites (donor and acceptor). Such similarities are expected by multiple models including DSBR, in cases involving 'sticky' end breakage and repair (Li et al. 2009), and internal duplication of exonic sequencing and usage of two resulting copies of an AG|GY containing motif (Rogers 1989; Venkatesh et al. 2009; Hellsten et al. 2011). Testing for similarity between the two boundaries of an intron is complicated by the lack of a clear null expectation, particularly given that introns in general in several species (including ancestral introns) exhibit sequence preferences at the termini of the flanking exons that match the corresponding intronic sequence (AG|gt donor and ag|GT acceptor).

This complication can be circumvented by specifically testing for an excess of donor-acceptor similarity in putatively more recently created introns relative to putatively more ancestral introns, since the former are more likely to retain sequence signatures betraying their mechanism of creation. I compared intron positions in putatively orthologous genes between (i) *M. occidentalis*, (ii) the tick *Ixodes scapularis*, the closest relative for which a genome is available; and (iii) *Homo sapiens*, chosen because of its retention of a large fraction of the ancestral metazoan intron complement (see e.g., Srivastava et al. 2008). The *I. oxodes* and *H. sapiens* genomes and annotations were downloaded from Genbank, intron-exon structures were extracted, and orthologs were defined by reciprocal BLAST searches at the

protein level. Genes were aligned at the protein level in Clustalw2 with standard parameters, and conserved protein-coding regions identified were defined as positions with ≥40% amino acid identity and no gaps for windows of 10 amino acid positions on both sides of a position, for all three pairs of species individually (use of a variety of less stringent criteria did not qualitatively change any of the following results; data not shown). Intron positions were then mapped onto the protein-level alignment and each *M. occidentalis* intron position was identified as novel or as shared (i.e., an intron position at which an intron in *I. scapularis* and/or *H. sapiens* is found at the exact homologous position in the alignment), as in Rogozin et al. (2003) or Roy et al. (2003). This identified 3355 novel and 2509 shared *M. occidentalis* introns. Notwithstanding the ongoing debate about whether shared intron positions reflect actual shared ancestral introns (Li et al. 2014), for brevity these will be referred to as shared/novel introns (in place of the more precise 'introns at shared/novel positions').

Figure 1b shows overall sequence logos of splice boundaries for the two sets of introns (Crooks et al. 2004). Interestingly, this comparison reveals a clearly much stronger preference for the so-called 'protosplice site' in new introns ($P < 10^{-10}$ for each of the -1, -2 and +1 bases individual, by simulation). This finding is consistent with previous results (Sverdlov et al. 2003; Qiu et al. 2004); however, inference of intron age in the previously analyzed datasets was a notoriously difficult endeavor (Rogozin et al. 2003; Roy and Gilbert 2005; Csuros 2005; Carmel et al. 2007), thus observation of the same pattern on this simpler dataset is a comforting confirmation.

To test for similarity between donor and acceptor sites of the same intron, I performed three similar but distinct tests. First, for each intron I performed an ungapped alignment of the 10 nucleotides straddling the donor and acceptor boundaries (5 nucleotides on either side of the splice boundary) and counted the number of nucleotide matches. Second, I counted the maximum number of such matches in a row within this same region. Third, I compared the maximum length of shared motif between donor and acceptor site regions within an extended 20 nucleotide region (five exonic nucleotides plus 15 intronic nucleotides from donor and acceptor). All three comparisons did show differences in the predicted direction, with clearly more matches, clearly longer runs of matches, and very slightly more overall longer shared motifs in novel than shared introns (top row of Figure 1c, left, middle and right, respectively; $P < 0.01$ for each test). However, because the AG|G protosplice site matches the corresponding intron positions (AG|g…ag|G), it is possible that this excess simply reflects greater similarity at the boundary due to the strong protosplice site in novel introns (which may or may not reflect the mechanism of intron creation; see below). Therefore I ran equivalent tests excluding the four positions flanking the splice sites (the NNgt and agNN sites; thus positions -7 through -3 and +3 through 7 were compared for total matches and run of matches, 7-to-3 for exonic side and 17-to-3 for intronic side for longest total shared motif). These tests showed a very different pattern (bottom row of Figure 1c). All three tests showed no clear difference between novel and shared introns: thus there is no evidence for greater similarity between 5' and 3' boundaries for novel introns outside of the stronger protosplice site .

Concluding remarks

Examination of intronic sequences in a metazoan with highly divergent intron-exon structures for the most part failed to reveal expected signatures of several tested mechanisms. The one clear difference between novel and shared introns was the greater strength of the protosplice site. The implications of this finding are not clear. Greater protosplice site character is exactly as expected from a variety of mechanisms that lead to duplication of the exonic insertion site; however, greater protosplice character could also reflect greater success of newly-created introns that insert into optimal exonic contexts (e.g., the upstream exonic AG can participate in basepairing to the U1 snRNA, a key step in spicing).

The lack of clear sequence signatures of intron creation mechanisms is also difficult to interpret. While the divergent character of the *M. occidentalis* genome relative to other available metazoan genomes attests to a large degree of change, no genomic sequences are available for close relatives of *M. occidentalis –I. scapularis*, the closest relative with a genome sequence, is from a different order – thus the timing of these changes is unknown. Most of the intron creation in the history of *M. occidentalis* could date to many million of years ago, in which case it would not be surprising that the sequence signatures of intron creation are no longer detectable. Alternatively, the introns in *M. occidentalis* could have been gained by mechanisms without clear sequence signatures. For instance, in the data of Li et al. (2009, 2014) some newly-gained introns do not show clear tandem repeats at the intron boundaries, which could reflect intron creation by repair of a blunt end double strand break (which is not expected to result in tandem duplication of a sequence at the borders). Still another possibility is suggested by the extremely rapid sequence evolution of new introns in *D. pulex* discovered also reported by Li et al. (2014), which could rapidly obscure the origins of new introns. Finally, introns in *M. occidentalis* could be created by a mechanism not yet imagined (or at least not tested here). Genomic sequences from closer relatives of *M. occidentalis* will hope to clarify the evolutionary history of this intriguing transformed genome.

References

Carmel L., Wolf Y.I., Rogozin I.B., Koonin E.V. (2007) Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 17, 1034-44.

Collemare J., Beenen H.G., Crous P.W., de Wit P.J., van der Burgt A. (2015) Novel Introner-Like Elements in fungi Are Involved in Parallel Gains of Spliceosomal Introns. PLoS One. 10, e0129302.

Collemare J., van der Burgt A., de Wit P.J. (2013) At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi. Commun Integr Biol. 6, e23147.

Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. (2004) WebLogo: a sequence logo generator. Genome Res. 14, 1188-90.

Csűrös, M. (2005) Likely scenarios of intron evolution. *Comparative genomics.* Springer Berlin Heidelberg. 47-60.

Curtis B.A., Archibald J.M. (2010) A spliceosomal intron of mitochondrial DNA origin. Curr Biol. 20, R919-20.

Denoeud F., Henriet S., Mungpakdee S., Aury J.M., Da Silva C., Brinkmann H., Mikhaleva J., Olsen L.C., Jubin C., Cañestro C., Bouquet J.M., Danks G., Poulain J., Campsteijn C., Adamski M., Cross I., Yadetie F., Muffato M., Louis A., Butcher S., Tsagkogeorga G., Konrad A., Singh S., Jensen M.F., Huynh Cong E., Eikeseth-Otteraa H., Noel B., Anthouard V., Porcel B.M., Kachouri-Lafond R., Nishino A., Ugolini M., Chourrout P., Nishida H., Aasland R., Huzurbazar S., Westhof E., Delsuc F., Lehrach H., Reinhardt R., Weissenbach J., Roy S.W., Artiguenave F., Postlethwait J.H., Manak J.R., Thompson E.M., Jaillon O., Du Pasquier L., Boudinot P., Liberles D.A., Volff J.N., Philippe H., Lenhard B., Roest Crollius H., Wincker P., Chourrout D. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science. 330, 1381-5.

Farlow A., Meduri E., Schlötterer C. (2011) DNA double-strand break repair and the evolution of intron density. Trends Genet. 27, 1-6.

Hellsten U., Aspden J.L., Rio D.C., Rokhsar D.S. (2011) A segmental genomic duplication generates a functional intron. Nat Commun. 2, 454.

Hoy M.A., Waterhouse R.M., Wu K., Estep A.S., Ioannidis P., Palmer W.J., Pomerantz A.F., Simão F.A., Thomas J., Jiggins F.M., Murphy T.D., Pritham E.J., Robertson H.M., Zdobnov E.M., Gibbs R.A., Richards S. (2016) Genome sequencing of the phytoseiid predatory mite Metaseiulus occidentalis reveals completely atomised Hox genes and super-dynamic intron evolution. Genome Biol Evol. *In press.*

Irimia M., Rukov J.L., Penny D., Vinther J., Garcia-Fernandez J., Roy, S.W. (2008). Origin of introns by 'intronization' of exonic sequences. Trends Genet 24: 378-381.

Li W., Kuzoff R., Wong C.K., Tucker A., Lynch M. (2014) Characterization of newly gained introns in Daphnia populations. Genome Biol Evol. 6, 2218-34.

Li W., Tucker A.E., Sung W., Thomas W.K., Lynch M. (2009) Extensive, recent intron gains in Daphnia populations. Science. 326, 1260-2.

Ma M.Y., Che X.R., Porceddu A., Niu D.K. (2015) Evaluation of the mechanisms of intron loss and gain in the social amoebae Dictyostelium. BMC Evol Biol. 15, 286.

Omilian A.R., Scofield D.G., Lynch M. (2008) Intron presence-absence polymorphisms in Daphnia. Mol Biol Evol. 25, 2129-39.

Qiu W.G., Schisler N., Stoltzfus A. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol Biol Evol. 2004 J.ul;21(7):1252-63 (2004) Epub 2004 Mar 10. Erratum in: Mol Biol Evol. 21, 1252-63.

Rogers J.H. (1989) How were introns inserted into nuclear genes. Trends Genet. 5, 213-6.

Rogozin I.B., Wolf Y.I., Sorokin A.V., Mirkin B.G., Koonin E.V. (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13, 1512-7.

Roy S.W., Fedorov A., Gilbert W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci U S A. 100, 7158-62.

Roy S.W., Gilbert W. (2005) Complex early genes. Proc Natl Acad Sci U S A. 102, 1986-91.

Roy S.W., Hartl D.L. (2006) Very little intron loss/gain in Plasmodium: intron loss/gain mutation rates and intron number. Genome Res. 16, 750-6.

Roy S.W., Irimia M., Penny D. (2006) Very little intron gain in Entamoeba histolytica genes laterally transferred from prokaryotes. Mol Biol Evol. 23, 1824-7.

Simmons M.P., Bachy C., Sudek S., van Baren M.J., Sudek L., Ares M J.r, Worden A.Z. (2015) Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations. Mol Biol Evol. 32, 2219-35.

Srivastava M., Begovic E., Chapman J., Putnam N.H., Hellsten U., Kawashima T., Kuo A., Mitros T., Salamov A., Carpenter M.L., Signorovitch A.Y., Moreno M.A., Kamm K., Grimwood J., Schmutz J., Shapiro H., Grigoriev I.V., Buss L.W., Schierwater B., Dellaporta S.L., Rokhsar D.S. (2008) The Trichoplax genome and the nature of placozoans. Nature. 454, 955-60.

Stajich J.E., Dietrich F.S. (2006) Evidence of mRNA-mediated intron loss in the human-pathogenic fungus Cryptococcus neoformans. Eukaryot Cell. 5, 789-93.

Sun Y., Whittle C.A., Corcoran P., Johannesson H. (2015) Intron evolution in Neurospora: the role of mutational bias and selection. Genome Res. 25, 100-10.

Sverdlov A.V., Rogozin I.B., Babenko V.N., Koonin E.V. (2003) Evidence of splice signal migration from exon to intron during intron evolution. Curr Biol. 13, 2170-4.

Venkatesh B., Ning Y., Brenner S. (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. Proc Natl Acad Sci U S A. 96, 10267-71.

Verhelst B., Van de Peer Y., Rouzé P. (2013) The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. Genome Biol Evol. 5, 2393-401.

Worden A.Z., Lee J.H., Mock T., Rouzé P., Simmons M.P., Aerts A.L., Allen A.E., Cuvelier M.L., Derelle E., Everett M.V., Foulon E., Grimwood J., Gundlach H., Henrissat B., Napoli C., McDonald S.M., Parker M.S., Rombauts S., Salamov A., Von Dassow P., Badger J.H., Coutinho P.M., Demir E., Dubchak I., Gentemann C., Eikrem W., Gready J.E., John U., Lanier W., Lindquist E.A., Lucas S., Mayer K.F., Moreau H., Not F., Otillar R., Panaud O., Pangilinan J., Paulsen I., Piegu B., Poliakov A., Robbens S., Schmutz J., Toulza E., Wyss T., Zelensky A., Zhou K., Armbrust E.V., Bhattacharya D., Goodenough U.W., Van de Peer Y., Grigoriev I.V. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. 324, 268-72.

Yang Y.F., Zhu T., Niu D.K. (2013) Association of intron loss with high mutation rate in Arabidopsis: implications for genome size evolution. Genome Biol Evol. 5, 723-33.

Yenerall P., Krupa B., Zhou L. (2011) Mechanisms of intron gain and loss in Drosophila. BMC Evol Biol. 11, 364.

van Baren M.J., Bachy C., Reistetter E.N., Purvine S.O., Grimwood J., Sudek S., Yu H., Poirier C., Deerinck T.J., Kuo A., Grigoriev I.V., Wong C.H., Smith R.D., Callister S.J., Wei C.L., Schmutz J., Worden A.Z. (2016) Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. BMC Genomics. 17, 267.

van der Burgt A., Severing E., de Wit P.J., Collemare J. (2012) Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Curr Biol. 22, 1260-5.

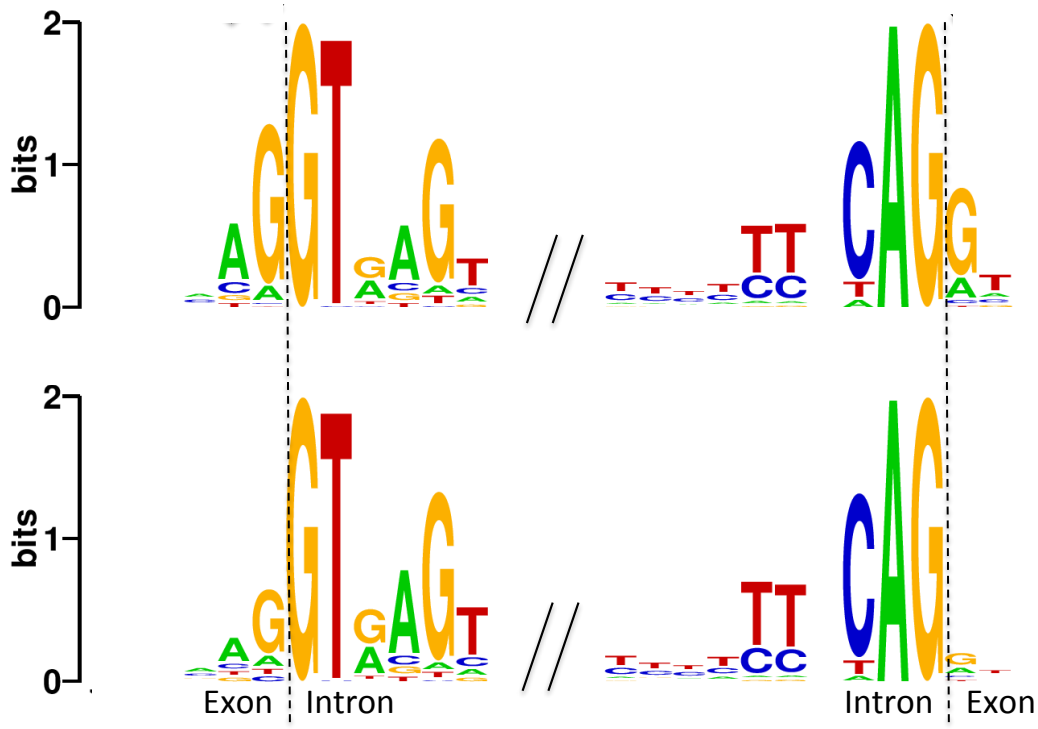
A

```
XP_003744432.1  CTCCATATTGTGCAAGGATCTTCTAAGAAGGACGTATGCACCGTTTCCGGCgtaagtgaataccgtggaccaatattctgat
XP_003744441.1  CTCCATATTGTGCAAGGATCTTCCAAGAAGGACGTATGCATAATTTCTGGCgtaagtgaataccgtggaccaatattctaat
                ************************ ****************   **** ******************************** **

XP_003744432.1  cattcaaaatgccgtcgctcctttacctacagattccaaacgtcccgtatgaattctccgaaccgattaaccgtcctcctcc
XP_003744441.1  cgatcagagggccgtcgctcctttccaacagattccaaacgtcccgtatgaattctccgaaccgattaaccgtccttctac
                *  *** *  ************** ** ************************************************** ** *

XP_003744432.1  tcctctcgcagCTCACCGCAGTTTCGACGATCCTCGACGTCAAGAAGAAGATCTATGCCACGA
XP_003744441.1  tcctctcgcagCTCACCGCACTCTCAACGATACTCGATGTCAAAAAAAAGATTCATGCCACGA
                ******************** * ** ***** ***** ***** ** *****  *********
```

B



C

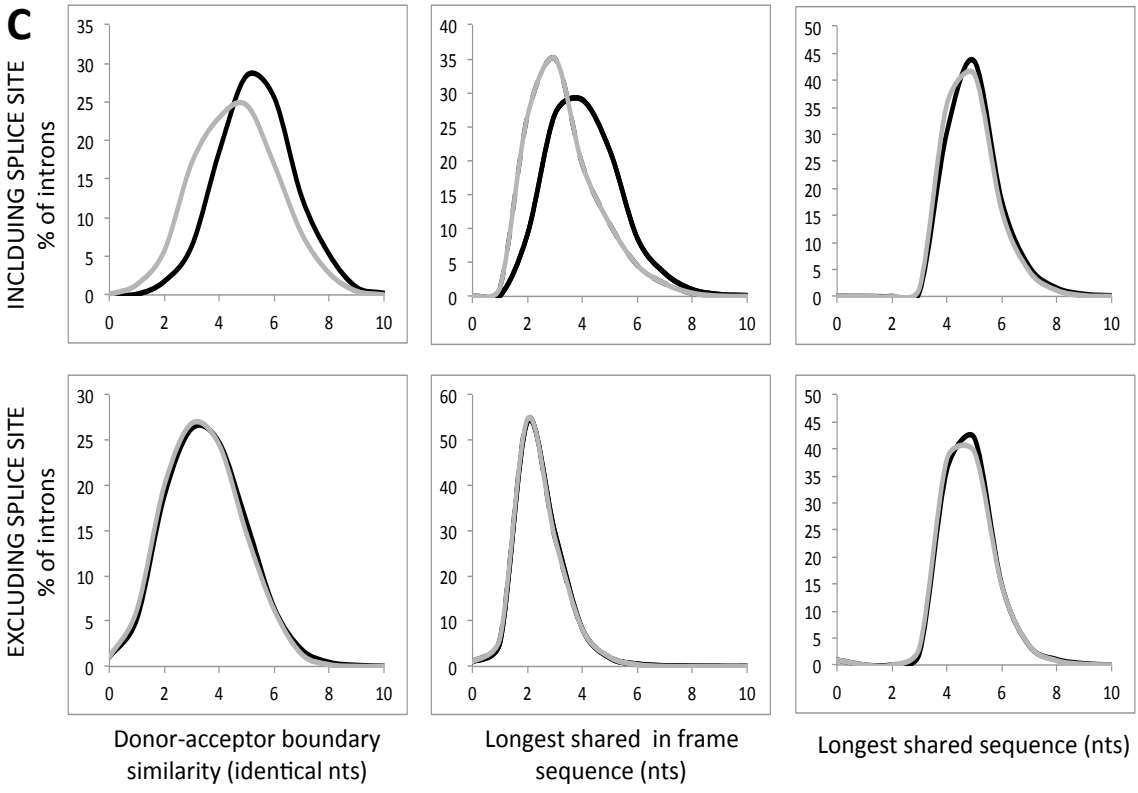


Figure 1. Search for signatures of intron creation mechanism in *M. occidentalis* reveals no clear patterns. A. The single pair of introns for which initial screening for signatures of intron propagation instead represents an extended region of homology. Clustalw2 alignment of similar introns with flanking exons within genes XP_003744432.1 and XP_003744441.1 are shown. Upper/lowercase indicates exonic/intronic nucleotides. B. Sequence logo plots for novel (top) and shared (bottom) introns, showing greater protosplice character at the two flanking nucleotide sites of each exon. C. Comparison of sequence similarity between donor and acceptor sites for novel (black) and shared (gray) introns. Left: number of nucleotide identities within ungapped alignment of 10 nucleotides spanning donor and acceptor boundaries for each intron. Middle: Longest number of shared nucleotides within ungapped alignment of 10 nucleotides spanning donor and acceptor boundaries for each intron. Right: Longest total shared motif between extended donor and acceptor regions (including 5 exonic and 15 intronic nucleotides). Results are given either including (top) or excluding (bottom) 2 terminal nucleotide positions for each exon/intron.