Avoidance of premature splicing in short introns explains the polypyrimidine tract

Burçin Yıldırım[a,b], Claus Vogl[a,b,*]

[a] *Department of Biomedical Sciences, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*
[b] *Vienna Graduate School of Population Genetics, A-1210 Wien, Austria*

Abstract

Among eukaryotes, the major spliceosomal pathway is highly conserved. Long introns may contain additional regulatory sequences, but those in short introns seem to be nearly exclusively related to splicing. While regulatory sequences involved in splicing are well-characterized, little is known about their evolution. At the 3' end of introns, the splice signal nearly universally contains the dimer *AG*, which consists of purines. The polypyrimidine tract upstream of the 3' splice signal is characterized by over-representation of the pyrimidine bases cytosine and thymine. We hypothesize that the over-representation of pyrimidines in the polypyrimidine tract is caused by avoidance of a premature splicing signal–*AG* should be most under-represented dimer. DNA-strand asymmetry patterns in fruit flies of the genus *Drosophila* confirm this prediction. In short introns of *Drosophila*, a region between the 5' splice signal and the branch point, which at least contains the nucleotides 8-30, is considered to evolve neutrally. Comparing this presumably neutrally evolving region to the polypyrimidine tract, we infer a moderate scaled selection strength below four against 3' splicing signals. Patterns of asymmetry in other eukaryotes indicate that avoidance of premature splicing similarly affects the nucleotide composition in their polypyrimidine tracts.

Keywords: splicing motifs, polypyrimidine tract, intron evolution, selective constraint, short intron, *Drosophila*

*Corresponding author

Email addresses: burcin.yildirim@vetmeduni.ac.at (Burçin Yıldırım), claus.vogl@vetmeduni.ac.at (Claus Vogl)

January 28, 2022

Introduction

Noncoding, intergenic and intronic, DNA sequences comprise a big proportion of eukaryotic genomes. Some noncoding sequences influence the centrally important processes of chromosome assembly, DNA replication, and gene expression (Ludwig, 2002; Pennacchio & Rubin, 2001). Other noncoding sequences were identified as nearly unconstrained and have been used as a neutral reference for inference of demography and selection in many population genetic analyses (*e.g.,* Parsch et al., 2010; Lawrie & Petrov, 2014).

In *Drosophila*, Bergman & Kreitman (2001) reported that more than 20% of noncoding sequences are highly conserved, while Halligan et al. (2004) found that introns, excluding splice sites, have a higher divergence than fourfold degenerate synonymous sites. The length distribution of introns is skewed in *Drosophila*: a high mode at short lengths between 45 to 85 bp contrasts with a long tail towards very long introns (*e.g.,* Mount et al., 1992; Stephan et al., 1994). The Bergman & Kreitman (2001) dataset included only long introns, while the Halligan et al. (2004) dataset included mainly short introns. The discrepancy between these two studies is likely to have resulted from the use of data sets with different lengths of introns: Longer introns contain more functional elements (Parsch, 2003), as shown by a negative correlation between divergence and length (Haddrill et al., 2005). Subsequent studies showed that sites at positions 8-30 of introns shorter than 65 bp have a higher divergence than and similar polymorphism as fourfold degenerate sites (Halligan & Keightley, 2006; Parsch et al., 2010; Clemente & Vogl, 2012). This has been interpreted as evidence for little or no selection. In these short introns, a ratio of about $AT : GC = 2 : 1$ likely reflects mutation bias (Haddrill et al., 2005; Clemente & Vogl, 2012), while the GC content increases with increasing intron length likely due to selection (Haddrill et al., 2005). Lawrie and colleagues (Lawrie et al., 2013; Machado et al., 2020) later utilized these short intronic regions to infer directional selection on fourfold degenerate synonymous sites in the presence of mutation bias in *Drosophila*. The inferred selection coefficient is approximately similar to that inferred from the skew of the spectra of fourfold degenerate sites, showing the utility of these intronic sites as a reference for inferring weak selection (Vogl & Bergman, 2015).

Regardless of their length, all introns contain conserved sequences necessary for splicing (Green, 1986). These include 5' (or donor) splice site and 3' (or acceptor) splice site and the branch point. The 5' splice site has a consensus sequence of $G|GTRAG$; the 3' splice site a consensus sequence of $YAG|G$ (in both cases | denotes the exon-intron boundary and Y and R define pyrimidines, the bases C or T, and purines, the bases A or G, respectively) (Breathnach & Chambon, 1981; Mount, 1982). Additionally, there is a pyrimidine-rich region of variable length, upstream of the 3' end, referred to as the *polypyrimidine tract* (3PT). The pre-mRNA splicing pathway has been shown to involve two main reactions which are universal to eukaryotic organisms (Ruskin et al., 1984; Green, 1986)(fig. 1). In the first reaction, the 5' splice site is cleaved and the RNA forms a loop (lariat) by attaching to the branch point. Thereafter, splicing occurs at the 3' end, and exons are ligated (Green, 1986). Experiments characterizing the splicing intermediates have shown that the order of the splice-site cleavage is highly conserved: no 3' splice-site cleavage is observed without the cleavage of 5' splice site (Padgett et al., 1984; Grabowski et al., 1984; Ruskin et al., 1984).

The relative importance of these *cis*-acting sequences in introns has been elucidated by mutagenesis studies (for reviews see (Green, 1986; Padgett et al., 1986)). Even though both ends of the intron are marked with long consensus sequences required for maximal splicing efficiency, the GT and AG dinucleotides seem to be the most functionally important, conserved motifs for the 5' and 3' ends, respectively (Breathnach & Chambon, 1981). Mutations in these nearly invariant

dinucleotides completely inactivate the authentic splice sites and often result in the activation of cryptic (alternative) splice sites, while mutations at other positions within these signals have lesser effects (Green, 1986; Padgett et al., 1986). The sequence of the branch point is not conserved universally among eukaryotes: Generally an adenosine residue is surrounded by rather variable sequence elements (Ruskin et al., 1985). While the sequence may be variable, the distance to both 5' and 3' end seems to be strongly constrained. Decreasing the distance between the 5' splice site and branch point prevents accurate splicing or leads to the activation of an upstream cryptic 5' splice site (Green, 1986). Hence, while the total length of this 5' loop region (5LR) seems constrained, its base composition seems to evolve neutrally. Similarly, the position of the branch point relative to the 3' end, and thus the length of the 3PT, is conserved (Ruskin et al., 1985). Several splicing factors preferentially bind to the 3PT. In contrast to the 5' side, both the size and the base composition seem to be important, as some mutations in the polypyrimidine tract reduce splicing efficiency (Ruskin & Green, 1985; Coolidge et al., 1997), while increasing the pyrimidine content enhances the removal of the intron (Guo et al., 1993).

Obviously, un-spliced or mis-spliced introns may be deleterious for the cell (Jaillon et al., 2008), such that selection should maintain efficient splicing, which requires docking of the splicing machinery to specific RNA sequence motifs. Even though the main splicing signals mentioned above are the major determinants for correct splicing, other auxiliary elements are also involved (Cartegni et al., 2002). Therefore, it is crucial to detect these auxiliary elements, the cryptic splice sites, and their variation among genomes for a better understanding of the pre-mRNA splicing. Several studies have identified splicing-related sequence motifs using DNA-strand asymmetry patterns (Zhang et al., 2008; Farlow et al., 2012). According to Chargaff's second parity rule, mono- or oligonucleotides under neutral evolution should have the same frequency as their reverse complement (Mitchell & Bridge, 2006). Deviations from Chargaff's second parity rule were observed in many organisms and they were associated with either selection-driven or neutral processes. Neutral processes, such as replication or transcription-coupled asymmetries, are attributed to different mutation and repair pressures between lagging and leading strands for replication or coding and non-coding strands for transcription (Touchon et al., 2003; Green et al., 2003).

Touchon et al. (2004) observed strong strand asymmetries at the 5' and 3' ends of introns of several eukaryotic genomes and also noted lower asymmetries in the internal regions. They attributed extreme biases at the intron ends to splicing-coupled selection, but did not investigate these processes carefully. Despite the rapid advance of our understanding of the splicing mechanism, information about its effects on the evolution of nucleotide patterns is still incomplete. Probably due to the lack of neutral sequences to compare, it has been hard to quantify strength and variation within and between genomes. Yet since splicing is a conserved and widespread biological phenomenon, we would expect identification and quantification of the evolutionary forces involved in splicing to provide insights into the genome evolution patterns of most eukaryotes. In this study, we investigate the effect of splicing-coupled selection on nucleotide composition. In particular, we hypothesize that the high pyrimidine content in the 3PT is a result of selection against the splicing signal containing the highly conserved dinucleotide AG, wich consists only of purine bases.

Positive or negative selection leads to over- or under-representation of the selected motif and thus a deviation from the Chargaff's second parity rule. A comparison of counts of motifs to counts of their reverse complements may identify elements involved in splicing. On the other hand, if deviations are mainly driven by transcription-coupled repair, one would expect a constant asymmetry score along the transcribed region. As mentioned, the loop formation between 5' site and branch point does not involve base pairing. Thus selection on nucleotide composition due to

splicing in the 5' loop region (5LR) should be minimal between the 5' splice signal and the branch point, as long as a proper length is maintained. We also know from population genetics studies that short introns of *Drosophila* are minimally affected by non-splicing constraints (eg Parsch et al., 2010; Clemente & Vogl, 2012; Lawrie et al., 2013). It seems that the base composition and higher-order oligonucleotide motifs in the 5LR can be used as a neutral reference. Recent studies showed that *Drosophila* performs transcription-coupled repair (Deger et al., 2019; Törmä et al., 2020) and this might also lead to deviation from strand symmetry in the 5LR (Bergman et al., 2017). Yet it can be assumed that this effect would be similar along the intron, and as long as selection is absent in the 5' region, it could be used to control for selection effects in other parts of the intron.

From the branch point towards the 3' end of the intron, the splicing machinery will try to identify the 3' splice site, which should lead to selection against spurious signals of the 3' splice site, which might cause premature splicing in the 3PT. A spurious 5 splice signal in this region might also interfere with the splicing mechanism and could be selected against similarly. Considering the sequential nature of splicing, however, selection against the 3' splice signal should be stronger than against the 5' splice signal after the loop formation. We aim to detect and quantify selection on sequence motifs within short introns by i) comparing complementary motifs within the polypyrimidine tract in the 3' region, as has been done earlier in studies identifying sequence patterns, and by ii) comparing motifs between the 3' polypyrimidine tract (3PT) to the presumably neutral 5' loop region (5LR) of short introns.

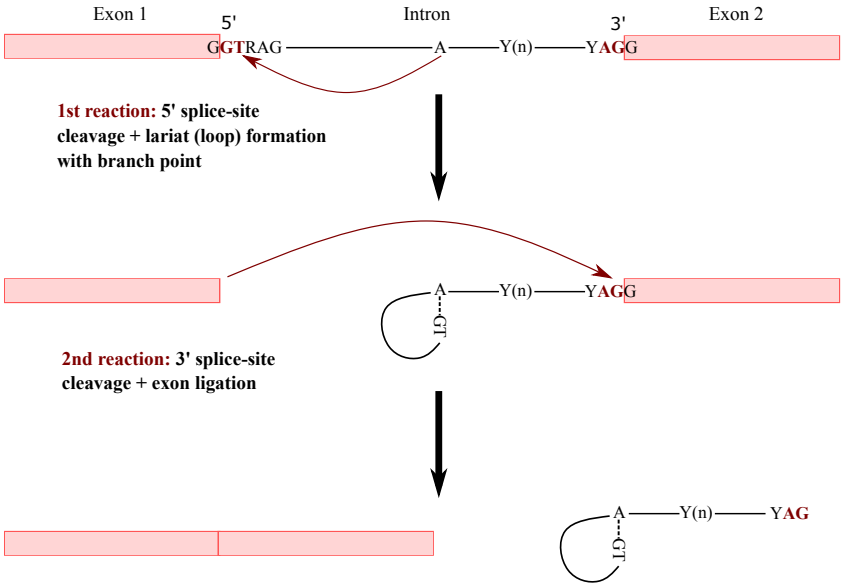


Figure 1: Representation of the pre-mRNA splicing pathway (adapted from Green, 1986)

Materials and Methods

Data used in the analyses

We analyzed whole genome data of a Zambian *D. melanogaster* population (Lack et al., 2015) and *D. simulans* populations from Madagascar (Rogers et al., 2014; Jackson et al., 2017).

For *D. melanogaster*, the dataset consists of 197 individuals for autosomes and 196 for the X chromosome. For analyses requiring comparison between chromosomes, the individual missing from the X chromosome data was also excluded from the autosomal data. Additionally, we repeated the analyses for 69 Zambian individuals reported to show no evidence of admixture with European populations according to Lack et al. (2015). The *D. simulans* dataset includes 21 individuals for each chromosome. Sequences have been obtained as consensus FASTA files. Using annotations from the reference genomes of *D. melanogaster* (r5.57 from *http://www.flybase.org/*) and *D. simulans* (Hu et al., 2013), intron coordinates were extracted and alignments of all samples were created. To avoid including the same intron sequence more than once due to alternatively spliced isoforms in the GFF file (see *https://www.ensembl.org/info/website/upload/gff.html*; last accessed November 1, 2020), we only used one representation of introns with similar coordinates. If the annotation information was coming from the non-coding strand, the sequence was reverse-complemented, so that the direction of all alignments of transcripts is from 5'-to-3'. The results presented in the main text focus on the *D. melanogaster* dataset, while the *D. simulans* dataset is used for comparison and can be found in the supplement.

While the length distribution of introns seems to fall clearly into a majority of short and a minority of long introns, the boundary between these two classes is unclear. Therefore we tried to define an upper length limit for the short intron class by binning according to length and then checking the nucleotide composition. For this analysis, we also included fifty bases from the preceding and following exons.

We differentiate i) a presumably neutral region between the 5' end and the branch point (abbreviated as 5LR) and ii) a presumably selected region through the 3' end with an excess of pyrimidines (abbreviated as 3PT). The 5LR is characterized by a ratio of approximately $AT : GC = 2 : 1$, which possibly reflects the mutation bias; the latter by a high pyrimidine content, which possibly reflects selection. Note that the branch point cannot easily be defined by a characteristic sequence pattern. To compare regions among introns of varying length, consensus positions from the 5' and 3' splice sites were defined among all length classes. Additionally, an eight-bp long region that symmetrically straddled the 3' junction (4 bp into the intron and 4 bp into the exon) was extracted.

Sequences were filtered out if they overlapped with annotated coding sequences or if they contained undefined nucleotides (N) in at least one of the individuals in the alignment. Furthermore, only alignments with full-length stretches (23 bp for the 5' loop region and 10 bp for the 3' polypyrimidine tract; see Results) were used for further analyses. Following these filtering steps, position-weight matrices and consensus sequences were created for the remaining alignments, resulting in 14,762 and 2,036 sequences for autosomal and X-linked 3PTs, respectively, and 11,938 and 1,556 sequences for autosomal 5LRs, respectively. By scanning these consensus sequences, all possible dimer ($4^2 = 16$), trimer ($4^3 = 64$), and tetramer sequences ($4^4 = 256$) were counted, separately for each region and chromosome. The expected proportion of oligomers was calculated from the base composition of the particular region. Results from autosomal introns are presented in detail in the text; results from X-linked introns in the supplement.

To evaluate whether alternative splicing influences our results, we binned introns that are in phase, length: $3n + 0$ (phase 0), and introns out of phase, length: $3n + 1$ or $3n + 2$ (non-phase 0). We then repeated analyses for datasets including i) only non-phase 0 introns in *D. melanogaster* ii) only phase 0 introns in *D. melanogaster* and excluding iii) common phase 0 introns between *D. melanogaster* and *D. simulans*, and compared the results to the full dataset.

Tests of strand symmetric evolution in the 5' loop region

Under neutrality and without transcription-associated mutation bias, counts of a motif and its reverse are expected to be identical. We examined this expectation in the 5LR with two, not mutually exclusive, tests:

- Test 1: a chi-square test to see whether forward and reverse oligonucleotide sequences are equally represented.
- Equivalence test: a test proposed by Afreixo et al. (2013), to see whether there are significant deviations from a 1:1 ratio between forward and reverse sequences. The procedure consists of obtaining confidence intervals for ratios (Katz et al., 1978) and checking if they are contained in the tolerance range, $(1/\delta, \delta)$, where δ represents a small tolerance to conclude equivalence for a ratio. If so, equivalence can be assumed.

The lower and upper values of the confidence intervals for the ratio of motifs, for a given z, are calculated as:

$$L_F = \frac{f_5(F)}{f_5(R)} e^{-z(\sqrt{\frac{1}{N_F} - \frac{1}{N} + \frac{1}{N_R} - \frac{1}{N}})} \tag{1}$$

and

$$U_F = \frac{f_5(F)}{f_5(R)} e^{z(\sqrt{\frac{1}{N_F} - \frac{1}{N} + \frac{1}{N_R} - \frac{1}{N}})}, \tag{2}$$

where $f_5(F)/f_5(R)$ is the ratio between two proportions of forward and reverse sequences from 5LR, N_F and N_R is the frequency of the forward and reverse sequences, respectively and N is the total number of oligonucleotide occurrences. To assume practical equivalence, we used a stringent $\delta = 1.1$ (Thanassoulis & Vasan, 2010).

Controlling for GC-biased gene conversion in the 5LR

A consequence of the repair of double strand breaks is gene conversion. Heteroduplex mismatches formed during the repair of double strand breaks can involve either pairing between the bases G and C (strong: S:S), A and T (weak: W:W) or between strong and weak bases, S:W. Preferential resolution of the S:W mismatches into $G : C$ rather than $A : T$ leads to GC-biased gene conversion (gBGC) (Marais, 2003). Since, gBGC only affects S:W mismatches they are referred to as GC-changing, while the others (S:S, W:W) are called GC-conservative. This categorization allows us to use unpolarized data to estimate GC bias while considering the site frequency spectra (SFS) of all six possible nucleotide pairs (Borges et al., 2019). The efficiency of gBGC (quantified by B) is directly related to effective population size (N_e), thus parametrized by the product of N_e and conversion bias (b). To infer the strength of gBGC in the 5LR, we used the maximum likelihood estimator of Vogl & Bergman (2015) and applied it to the SFS from GC-changing mutations of the 5LR. The SFS from GC-conservative mutations is considered as putatively neutral control ($B = 0$). We performed likelihood-ratio tests (LRT) to compare between the different nested models.

Calculation of asymmetry scores

Strand asymmetry of mono- and oligonucleotides for each region was calculated as

$$S = (N_F - N_R)/(N_F + N_R) \tag{3}$$

Mononucleotide asymmetries between T *vs.* A and C *vs.* G are denoted as S_{TA} and S_{CG}, respectively (Touchon et al., 2004). Additionally, at the mononucleotide level, we also calculated asymmetry scores for each position in 5LR, 3PT, 3' junction and document the change in the scores with position via regression analysis to study whether there is a dependence on distance to the splice site. The expected strand asymmetry of an oligonucleotide was predicted from the base composition of the region under consideration.

We also obtained the short introns of six additional eukaryotes (human, sea urchin, *Caenorhabditis elegans*, moss, rice, and *Arabidopsis thaliana*) from the exon-intron database, EID (Shepelev & Fedorov, 2006). The length range for the short intron class was defined separately for each species depending on the length distribution of introns and the nucleotide composition similar to that described for *Drosophila*. Inside these introns, probable polypyrimidine tracts were again characterized with high pyrimidine content and asymmetry scores of oligonucleotide motifs with lengths two and three were calculated. The number of sequences (3PT region) used for each species; Human: 12740, Sea Urchin: 21285, Rice: 24137, *Arabidopsis*: 55354, Moss: 50585, *C. elegans*: 54082.

Quantifying selection strengths

Motifs distinguished from others by their strand asymmetry in the previous step can be functionally constrained due to splicing-coupled selection, which would lead to either under- or over-representation of pairs of symmetric sequences. We quantified the selection strength causing this asymmetry on particular motifs (denoted as M) using the 5LR as background:

$$\log\left(\frac{f_3(M)/(1-f_3(M))}{f_5(M)/(1-f_5(M))}\right) = \gamma(M)\,, \tag{4}$$

where $f_3(M)$ and $f_5(M)$ represent the proportion of motifs from 3PT and 5LR, respectively. The method does not require complete strand symmetry in the 5LR, but instead assesses the selective force leading to either depletion or excess of the particular motif in the 3PT compared to the 5LR, in the presence of non-selective forces such as mutation bias.

To create confidence intervals (CIs) around the estimates, we bootstrapped by sampling introns (both 5LR and 3PT datasets) with replacement 1000 times. The size of the each bootstrap sample was approximately same with the original datasets. Selection coefficients were recalculated for each motif from these bootstrap samples, the 2.5% and 97.5% quantiles of the resulting distributions were used as lower and upper bounds of 95% CIs.

Results

Region definition inside short introns

The length distribution of *D. melanogaster* introns has a very high mode around minimal lengths, creating a "short intron" class (Mount et al., 1992) and a long tail of long introns. While the lower limit of the short introns class is given by the data, the upper limit is defined differently among studies of intron evolution (*e.g.,* Halligan & Keightley, 2006; Lawrie et al., 2013). Nevertheless all studies agree on patterns specific for the presumably neutral part of short introns: while the first seven bases at the 5' end of each intron are affected by the presence of splice sites, thereafter mutation pressure in short introns leads to an AT-rich base composition of approximately $AT : GC = 2 : 1$ (Haddrill et al., 2005; Parsch et al., 2010). This matches our findings: Bases in positions 8-30 of introns shorter than 65 bp exhibit a consistent $AT : GC$ ratio of approximately $2 : 1$. For

introns with lengths between 65 to 85 bp, this presumably neutral region stretches out until position 40. In longer introns of more than 75 bp length, the base composition in positions 8-30 approaches a ratio of $AT : GC = 1 : 1$ (fig. 2), presumably reflecting the effect of selection. We therefore use the ratio of $AT : GC = 2 : 1$ in positions 8-30 to differentiate short from long introns and include only short introns up to 75 bp in further analyses and extracted sequences between positions 8-30 as a proxy of a neutrally evolving region (fig. 4).

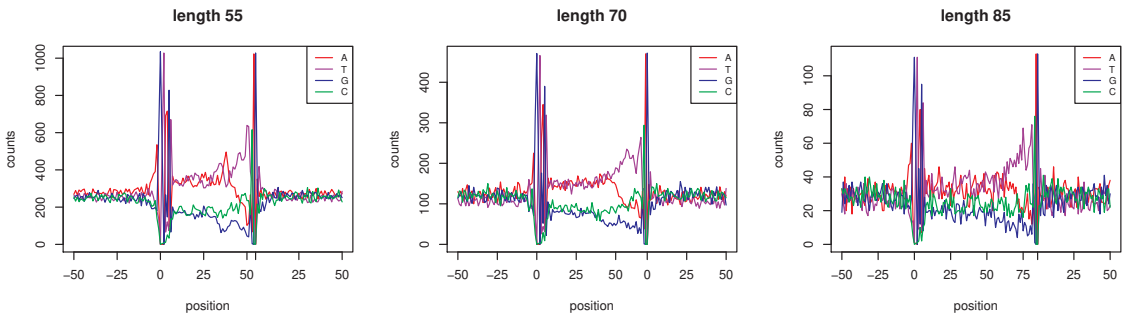


Figure 2: Nucleotide composition around exon-intron junctions for introns with length 55, 70 and 85 bp. Positions depicted as 0 correspond to the junctions between intron and exons. Only one candidate length class were chosen to visualize pattern with increasing length. Total numbers of sequences used for each length class are 1044, 474 and 117, respectively

We next searched for a region possibly under splicing-coupled selection inside short introns. The proportion of purine bases (A and G) and pyrimidine bases (C and T) is approximately equal from base eight to the midpoint of each intron (fig. 3). Then, pyrimidine bases become visibly and steadily more abundant until four bases from the 3' end of the intron. These last four bases are already part of the splice signal. The length of this pyrimidine-rich tract varies for different lengths of introns, with longer introns having longer pyrimidine-rich tracts. For consistency among length classes, we always extracted a region of 10 bp length from the 3' end after excluding the last four bases (fig. 4).

By taking the minimal consensus position range among each length class for analyzed regions, we also avoided to include the branch point, which shows significant sequence variability and does not seem in a fixed distance from either 5' or 3' ends.

Neutrality of the 5' loop region

By exploring DNA-strand asymmetry patterns, we aim to detect whether there are localized, systematic biases for particular motifs in the 3PT. If there are such motifs, the comparison with the presumably neutral 5LR may give insights about the selection strengths for these motifs. However, even in the absence of splicing-coupled selection, neutral processes might create strand asymmetry. And it is not yet clear if the evolution in the 5LR is strand symmetric and different from 3PT at all. Hence we explored strand symmetry in this region with two tests.

Both tests showed that most of the 5LR motifs deviate from DNA-strand symmetry (fig. 5A-B). Fundamentally, the null-hypothesis for the chi-squared and equivalence test is the equal representation of forward and reverse sequences. In line with this, results are consistent, *i.e.,* trimers with the lowest chi-squared p-values also have the most extreme equivalence test CI.

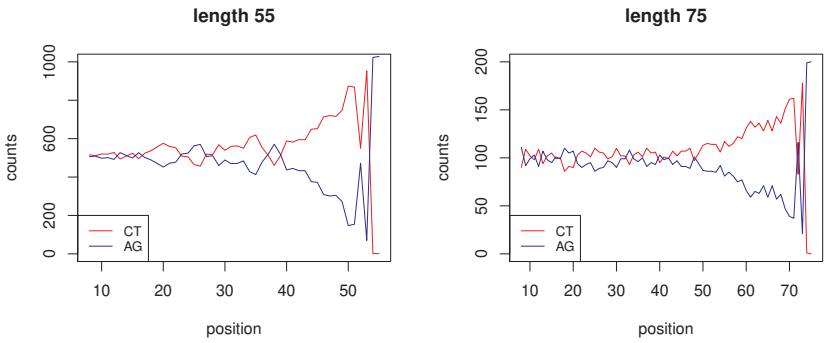


Figure 3: *AG* (purines, blue lines) and *CT* (pyrimidines, red lines) content per position in short introns with length 55 (1044 sequences) and 75 bp (205 sequences) (the same patterns has been observed for other length classes also)

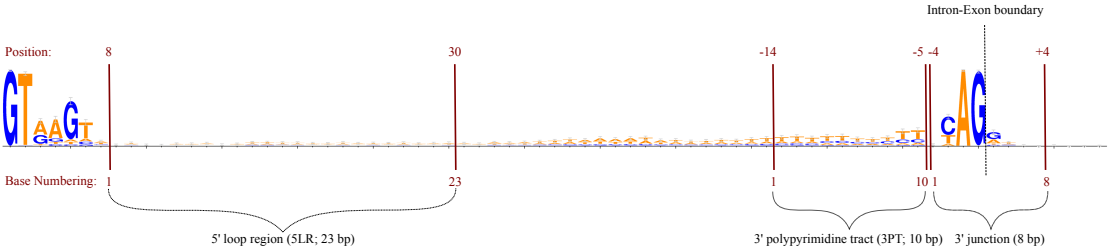


Figure 4: Representation of the regions analyzed. The base numbering of each region is ascending as going from 5' to 3'.

Given the significant deviations from strand symmetry, we can say that the 5LR is not conforming to the strand symmetric evolution. Yet, the more extreme deviations observed for all 3PT motifs (fig. S1A-B), compared to the relatively slight deviations for some 5LR motifs, indicate different causes for asymmetries in these two regions.

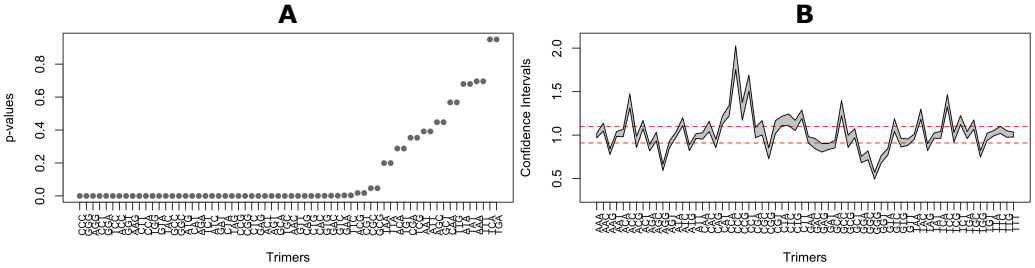


Figure 5: Tests of strand symmetric evolution in the 5LR. (A) Ordered p-values from chi-square tests, for the equality of forward and reverse complement trimers. (B) Confidence intervals for the ratios of forward and reverse complement trimers. Red dashed lines correspond to the tolerance range to assume equivalence.

Furthermore, despite the utility of the 5LR of short introns as a neutral reference in population genetic analyses, their biallelic frequency spectra deviate from the neutral prediction of symmetry, with an excess of high-frequency GC variants, both in autosomes and the X chromosome (fig. S2A-B). Nevertheless, in short introns, evidence for the presence of directional selection is equivocal (Parsch et al., 2010; Halligan & Keightley, 2006; Clemente & Vogl, 2012; Vogl & Bergman, 2015), with no plausible biological explanation for a selective constraint at the base composition level. This directional force favoring GC is either explained by context-dependent mutational pattern (Clemente & Vogl, 2012) or by gBGC (Jackson et al., 2017). If gBGC is a meiotic process, we would expect the effects of gBGC to be similar between autosomes and the X chromosome, since the lack of recombination in male meiosis in *Drosophila* and the reduced effective population size of the X chromosome cancel out. This leads to the same expected intensity of gBGC for autosomes and the X chromosome, unlike with directional selection.

Bearing this in mind, we quantified the effect of gBGC for autosomes and X from the SFS of GC-changing mutations constructed based on their segregating GC frequency. We found a weak force, significantly different from $B = 0$, favoring GC both in autosomes and X (table 1). On the other hand, the values inferred from the GC-conservative mutations were not significantly different from 0, as expected under gBGC. Additionally, to understand whether the strength significantly differs between autosomes and X, we compared the pooled data likelihood to the sum of likelihoods. B values inferred separately do not fit the data significantly better than the pooled data (LRT $\chi^2_{df=1} = 2.238, p = 0.135$). This indicates the same efficiency of the GC favoring force in autosomes and the X chromosome, further suggesting gBGC as the cause of the deviation in the SFS, rather than the selection.

Table 1: B values inferred from the SFS of GC-changing and conservative mutations, for autosomes, X chromosome and pooled data (autosomes+X). Significance of values are from the LRT, comparing $B = 0$ model.

	B (GC-changing)	B (GC-conservative)
Autosomes	0.499^{***}	-0.041
X	0.596^{***}	-0.026
Pool	0.51^{***}	-0.04

*** $p < 0.001$

Strand asymmetry patterns

Next, we studied strand asymmetry patterns in both regions to examine the selective pressure exerted by the splicing process and observe the differences/similarities between the patterns created by neutral versus selective forces. Initially, we concentrated on the mononucleotide asymmetry. The proportion of complementary nucleotides both in the 5LR and 3PT of autosomal and X-linked introns significantly differ from the 1:1 ratio (table 2, S1). In the 5LR, there is a slight excess of T over A ($\chi^2_{df=1} = 18.464$, $p < 0.001$; $\chi^2_{df=1} = 27.928$, $p < 0.001$ for autosomes and X, respectively) and a bigger excess of C over G ($\chi^2_{df=1} = 449.45$, $p < 0.001$; $\chi^2_{df=1} = 56.976$, $p < 0.001$ for autosomes and X, respectively). The proportion of nucleotides are calculated from the consensus sequences, and are similar to previously reported values in Bergman et al. (2017), calculated from monomorphic sites. The same qualitative pattern is observed for the 3PT, where the bias towards T and C is much more pronounced (between A and T $\chi^2_{df=1} = 14474$, $p < 0.001$; $\chi^2_{df=1} = 1940$,

$p < 0.001$ for autosomes and X, respectively and between G and C $\chi^2_{df=1} = 9877.3$, $p < 0.001$; $\chi^2_{df=1} = 1517$, $p < 0.001$ for autosomes and X, respectively). Accordingly, the resulting asymmetry scores are higher for the 3PT and S_{CG} (table 2).

Table 2: Proportion of nucleotides (A, T, G, C) and mononucleotide asymmetry scores (S_{CG}, S_{TA}) in 5LR and 3PT of autosomal introns

	A	T	G	C	S_{CG}	S_{TA}
5LR	32.01%	32.68%	16.45%	18.86%	6.81%	1.02%
3PT	21.12%	46.95%	8.66%	23.27%	45.78%	37.95%

The observed nucleotide bias, an excess of the pyrimidines T and C, conforms to that previously noted for the 3PT, which has been labeled "polypyrimidine" tract. The trend is more subtle but similar in the 5LR. As the 5LR of short introns is considered to be least selectively constrained, the trend there should reflect non-selective processes, *i.e.,* transcription-associated mutation bias. As such processes should be similar over the whole region, we investigated how the trend is changing depending on the position. The 5LR shows similar scores along its length without a significant correlation with position, whereas the biases in the 3PT increase slowly but significantly towards the 3' junction (fig. 6). It thus seems that the selection pressure is increasing with increasing proximity to the splicing signal.

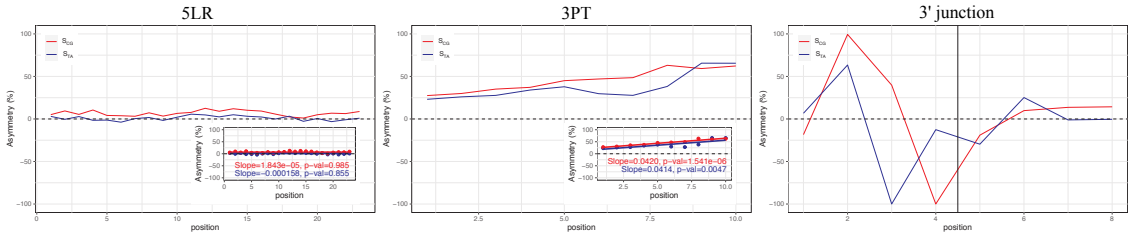


Figure 6: Per position mononucleotide asymmetry scores in 5LR, 3PT and the 3' junction. TA and CG asymmetries are shown in blue and red, respectively. Dashed horizontal lines corresponds to symmetry at 0. Vertical line on the far right plot corresponds to the intron-exon boundary. The inset plots for 5LR and 3PT show the regressions between position and asymmetry scores.

If over-representation of pyrimidines in the 3PT is caused by selection against premature splicing, it should not act on a single base; rather the 3' splice site consensus sequence is a tetramer, $YAG|G$, where especially the central AG is conserved. Selection should thus suppress the occurrence of this particular motif near the splice junction, such that it should be depleted on the coding strand. Hence, oligonucleotides associated with the 3' splice signal should be strongly under-represented in the 3PT compared to the 5LR. The most conserved sequence in the tetrameric splicing signal is the dimer AG and it is unclear how much the preceding Y and following G contribute, *i.e.,* if a dimeric, trimeric, or tetrameric motif is most strongly selected against. We thus explored asymmetries of all possible dimeric, trimeric, or tetrameric motifs. Trimers provide more information compared to dimers, such that the relative importance of bases following or preceding the AG motif can be gauged, while the shear number of tetramers makes them difficult to present. We thus show the results for trimers in the text. Generally, results from dimers and tetramers (which can be found in the supplementary information) confirm those from trimers.

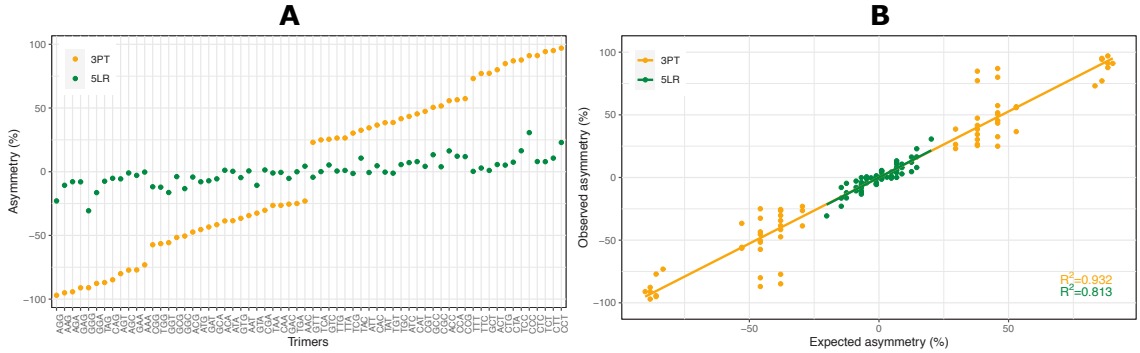


Figure 7: (A) Asymmetry scores per trimer, per region. (B) Correlation between the observed asymmetry of trimers and that expected from the base composition for each region. Orange dots corresponds to 3PT, while green dots are 5LR.

Asymmetry scores are symmetrically distributed around 0, since each motif and its reverse complement have the same value with opposite signs. In the 5LR, mononucleotide proportions of complementary bases differ only slightly, such that counts of oligonucleotides and their complement should be approximately equal. As expected, pairs of asymmetry scores from the 5LR are close to symmetry with a relatively low variance of 0.01 for both autosomes and the X chromosome. The variance of pairs of asymmetry scores from the 3PT is much higher with 0.39 for the autosomes and 0.42 for the X chromosome (fig. 7A, S3A). This indicates selection in the 3PT. Tellingly, the most under-represented trimers in the 3PT all contain the *AG* dimer, *i.e.,* the most conserved part of the consensus 3' splice site. Indeed, the systematic biases in these motifs, compared to the 5LR and other random motifs in the 3PT, cannot be explained by neutral evolution. Rather they seem to be the result of selection depleting splicing signals from the polypyrimidine tract.

Next, we asked whether there is a correlation between the asymmetries of oligo- and mononucleotides. Because of the limited information of mononucleotides, their asymmetry patterns hardly allow inference of splicing-coupled selection. However, under-representation of particular motifs in the 3PT, together with the position-dependent skew, can be related to avoidance of splicing signal in the 3' extreme of introns. To understand the connection between the two levels of asymmetries, we examined the relationship between the observed asymmetries of trimers and that expected from the base composition of the region under consideration (fig. 7B). They were highly and significantly correlated (Pearson $R^2 = 0.815$, $P < 0.0001$; $R^2 = 0.932$, $P < 0.0001$, for 5LR and 3PT, respectively). In the case of the 3PT, this high correlation suggests that the extreme skew in the base composition (high pyrimidine content) can be indeed explained by selection against the most conserved part *AG* of the 3' splice signal.

Selection coefficients

Deviations from strand symmetry in the 5LR could not be associated with selection-driven processes, either at the mono- or oligonucleotide level. The observed slight, constant deviations along the region possibly reflect the effect of transcription-coupled repair leading to mutational asymmetries (transcription-associated mutation bias-TAMB), in line with the recent studies (Bergman et al., 2017; Deger et al., 2019; Törmä et al., 2020). Furthermore, deviations in the biallelic frequency spectra of the 5LR were more compatible with the neutral process, *GC*-biased

gene conversion, rather than directional selection. The inferred strength of directional force causing the deviation was not significantly different from 0 for GC-conservative mutations, as opposed to those from GC-changing mutations. This supports the presence of a directional force differing between these two classes of mutations, likely gBGC. Additionally, due to the lack of recombination in *Drosophila* males, we expected that the intensity of gBGC should be identical for autosomes and X, and the expectations have been met. Thus, even though strand symmetric evolution can not be assumed for the 5LR, it can still be used as a neutral reference for each motif in the 3PT, as the non-adaptive forces affecting these two regions should be similar.

Selection was assessed as the force leading to depletion or excess of motifs in the 3PT compared to the 5LR, which takes into account the presence of non-adaptive forces. As with the asymmetry scores, the strongest negative selection coefficients belong to the trimer motifs containing AG (fig. 8, S3B). Strikingly, the three lowest selection coefficients belong to the motifs AGG, TAG, and CAG, perfectly corresponding to the consensus 3' splice signal, $YAG|G$. While selection against these trimers is relatively strong with absolute values of nearly three, the highest inferred positive selection coefficient is around 1, much lower than in the reverse direction. These slightly over-represented sequences are rich in pyrimidines, *i.e.,* C and T. Higher selection coefficients against trimers involved in splicing than those for pyrimidine-rich trimers support further that over-representation of pyrimidines is a result of selection against splicing motifs.

The results from the dimer and tetramer analyses confirm this pattern: motifs containing AG and YAG have the most extreme skew in the asymmetry patterns and higher selection coefficients (fig. S4). Interestingly, there is a slight increase in the inferred selection strengths from dimer to tetramer. Stronger purifying selection might be expected as the motif gets longer and therefore more specific. Strong selection on long motifs may lead to apparent selection on shorter motifs, even affecting the base composition leading to the over-representation of pyrimidines. Moreover, we do not observe systematic biases in asymmetry or high selection strengths for the parts of the hexameric 5' splice signal at the dimer (*e.g.,* GT), trimer (*e.g.,* GTR, GGT) or tetramer (*e.g.,* $GTRA$, $GGTR$) level, but for 3' splice signal starting from dimers (fig. S4). As we predicted from the sequential nature of splicing, selection against the 3' splice signal in the pyrimidine tract is stronger than that against the 5' splice signal.

Indeed, inferred selection strengths against motifs of the 3' splice signal are always stronger compared to those of the 5' splice signal (fig. S5). We are aware that the 5' splice signal also contains the AG dimer and has relatively high purine content. However, the most conserved part in the 5' splice signal is the GT dimer, which is not depleted at the dimer level (fig. S5, S4). Therefore, we posit that strong selection against the 3' splice signal affects the base composition to create the polypyrimidine tract.

Even though we excluded regions overlapping with coding sequences, there are studies showing an excess of phase 0 introns in alternatively spliced genes (Long et al., 1995) and higher sequence conservation of splice signals in exons flanking phase 0 introns (Long & Deutsch, 1999). Thus, to assess whether intron phases influence our estimates, we created additional datasets including (i) only non-phase 0 introns and (ii) only phase 0 introns in *D. melanogaster*. Since common phase 0 introns between *D. melanogaster* and *D. simulans* might be also associated with a higher conservation, we created another dataset (iii) excluding all common phase 0 introns. Selection coefficients were similar for each trimer in all three datasets (fig. S6), showing that the magnitude of selection varies little within the genome.

Since the splicing machinery is highly conserved, we would expect selection against the splicing signal in the polypyrimidine tract to be similar among species. To investigate this expectation,

we also analyzed the introns of *D. simulans* (fig. S7). The magnitude of the selection strength was similar and consistent for each trimer motif in both species.

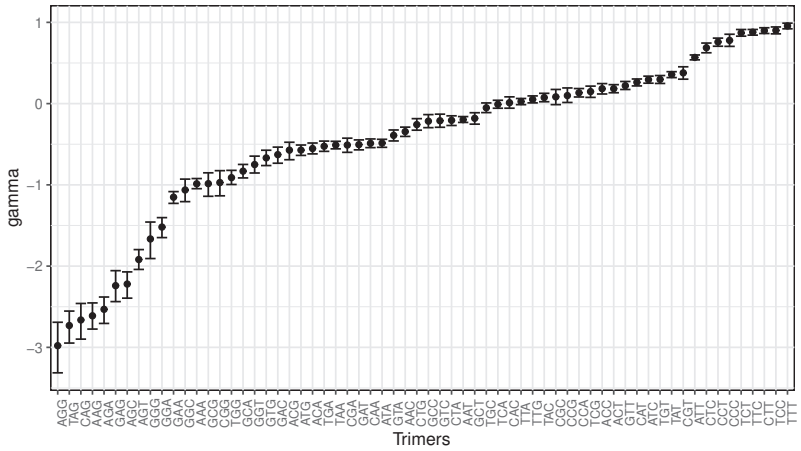


Figure 8: Selection coefficients of each trimer in autosomal 3PT, calculated by equation 4. Error bars represent 95% CIs from 1000 bootstraps of the datasets.

Asymmetry patterns from other eukaryotic species

We used the 5LR of *Drosophila* short introns as a neutral reference to compare to DNA-asymmetry patterns in the 3PT, validate the effect of selection, and quantify its strength. But we are not aware of a similar neutrally evolving region in other eukaryotic species. Therefore, our study focused on *Drosophila* short introns. Given our results from *Drosophila*, we checked the universality of this process qualitatively in other eukaryotes, where no neutral control region is available. If 3' splice signal associated motifs are similarly under-represented, we suggest that the same evolutionary process is responsible for the polypyrimidine tract in other eukaryotes as in *Drosophila*.

To test this, we analyzed the asymmetry patterns in polypyrimidine tracts of human, sea urchin, rice, mouse-ear cress (*Arabidopsis thaliana*), moss, and worm (*Caenorhabditis elegans*). Comparative analyses suggested that the strength of the polypyrimidine tract (with respect to *e.g.,* length, pyrimidine content) differs between plants and metazoa, and shows a gradual increase from *C. elegans* to human within metazoa (Schwartz et al., 2008). The species we selected cover this diversity. Among them, *C. elegans* displays the most unusual properties regarding splicing (Riddle et al., 1997). Their introns are exceptionally short, most under 60 nucleotides, and seem to lack an obvious polypyrimidine tract. Additionally, they use splicing signals not containing *AG* more often than other species (Aroian et al., 1993; Riddle et al., 1997).

We again focused on the short intron class of each species to minimize selection not related to splicing (fig S8). Even though the relative nucleotide composition varied between species, pyrimidines were over-represented close to the 3' ends of short introns. In human and sea urchin, both C and T increased in abundance, while for other species only T increased in relative frequency (fig. S9). The distribution of asymmetry scores of dimeric and trimeric motifs in these pyrimidine

enriched regions resembles that in *Drosophila* (fig. S10, S11). At the dimer level, the most under-represented motif was *AG* for all species, except *C. elegans*. In *C. elegans AG* is not a requirement for maximum splicing efficiency, but *AA* or *AC* dinucleotides are also used as a 3' splice signal (Aroian et al., 1993; Riddle et al., 1997). Thus, observing *AA* as the most under-represented dimer in asymmetry scores still supports that the nucleotide composition of the 3' region of short introns evolved to avoid premature splicing. At the trimer level, *AG* containing motifs had again the lowest asymmetry scores in human, sea urchin, rice and *Arabidopsis*. In moss and *C. elegans*, they were also under-represented (negative asymmetry scores), yet did not always have the lowest values. As in *C. elegans*, this possibly reflects the variation in splicing signal usage in moss.

Discussion

Many *in vivo* and *in vitro* experiments have provided information about the functional importance of the 3' polypyrimidine tract (3PT) in splicing (Green, 1986; Spellman et al., 2005). It is clear that the significance of the polypyrimidine tract is not just to fulfill a length requirement, but is functionally important for splicing. Indeed, it serves as the binding site for the required splicing factors (Spellman et al., 2005) and mutations in it result in a decreased splicing efficiency. Increasing pyrimidine content just before the 3' splice site of a short intron enhanced the removal of the intron, while insertions between the 5' splice site and the branch point had the opposite effect (Guo et al., 1993). Although the 3PT's role in splicing is well established, it has not been characterized how its nucleotide composition evolved. Here we present evidence that the high pyrimidine content is a result of purifying selection against the occurrence of a 3' splice signal, acting on the region between the branch point and 3' end, rather than selection for pyrimidine bases. This finding might also explain the increased efficiency of splicing with higher pyrimidine content: as purifying selection depletes purine containing motifs that could behave as cryptic 3' splice sites, the accuracy of splicing enhances.

Although the main splicing signals are necessary for accurate splicing, the surrounding sequence context is also important. Our study shows how this sequence context evolved. It is conceivable that once the pyrimidine content in the 3PT increased due to avoidance of the 3' splice signal, splicing factors co-evolved with this pattern to develop higher binding specificity to pyrimidines. In a similar manner, it has also recently been shown that mutation bias together with purifying selection created precursor exonic splicing enhancers (pre-ESEs), to which trans-acting factors adapted later (Rong et al., 2020).

We did not only characterize the evolution of the nucleotide composition due to splicing signal avoidance qualitatively, but also quantified the current selection strengths acting on the motifs associated with splicing. For this, we used the 5' side of the *Drosophila* short introns. Many population genetic studies have already used this region as a neutral reference to detect selection in *Drosophila* (Parsch et al., 2010; Lawrie et al., 2013; Machado et al., 2020). We also found no association of the nucleotide composition in this region with selective processes both in the population genetic analysis of site frequency spectra and DNA-asymmetry patterns. The nature of the splicing process likely explains the lack of constraint in this region: the 5' splice-site cleavage and loop formation between branch point are coupled reactions and base composition in the 5LR is irrelevant as long as the splice site is recognized and appropriate length for loop formation is conserved.

We only utilized short introns in this study, since in longer introns other processes may interfere (Parsch, 2003; Haddrill et al., 2005). By using only *Drosophila* short introns we ensured

that the 5LR is close to neutrality, and we minimized other constraints in the polypyrimidine tract. Additionally, experiments have shown that intron length is important for intron/exon recognition and splicing. Splice sites are recognized differently in long and short introns, depending on the distance between donor and acceptor sites. With longer introns, exons define the recognition unit ("exon definition"): the 5' end of the exon (which corresponds to the 3' end of the previous intron) and the 3' end of the exon (which corresponds to the 5' end of the next intron) are recognized as a pair (Berget, 1995). With short introns, introns define the recognition unit ("intron definition"): 5' and 3' splice sites of the same intron are recognized as a pair (Talerico & Berget, 1994). This implies that intron length will also influence the selection pressure on splicing motifs: with intron definition, purifying selection to purge cryptic splice sites between the branch point and 3' end should be stronger in short introns. Stronger splicing-coupled selection in short introns of several eukaryotic species, including *Drosophila melanogaster* compared to long introns has also been suggested by Farlow et al. (2012).

In *Drosophila*, the presence of an established neutral reference, the 5LR, enabled the quantification of selection strength in the 3PT, in addition to detecting presumable selection in the 3PT by comparing asymmetry patterns. In other eukaryotes, the preferred splice signal and the nucleotide composition varies, as does the nucleotide composition and length of the 3PT (Nguyen & Xie, 2019; Coolidge et al., 1997), although pyrimidines are generally over-represented (Coolidge et al., 1997). Due to the lack of a neutral reference in other eukaryotes, quantification of the selection strength is not possible. Nevertheless the asymmetry pattern within the 3PT also suggests the same mechanism as in *Drosophila:* oligomers containing the 3' splice signal are under-represented in 3PT of short introns, *i.e.,* the region between the branch point and the 3' splice signal. Thus selection against premature splicing in the 3PT seems to be universal among eukaryotes.

Acknowledgements

The authors thank all members of the Vienna Graduate School of Population Genetics for support and discussion. The work was funded by the Austrian Science Fund (FWF; W1225-B20).

Competing interests

The authors declare no competing interests.

References

Afreixo, V., Bastos, C. A., Garcia, S. P., Rodrigues, J. M., Pinho, A. J., & Ferreira, P. J. (2013). The breakdown of the word symmetry in the human genome. *Journal of Theoretical Biology*, *335*, 153 – 159. doi:10.1016/j.jtbi.2013.06.032.

Aroian, R. V., Levy, A. D., Koga, M., Ohshima, Y., Kramer, J. M., & Sternberg, P. W. (1993). Splicing in *Caenorhabditis elegans* does not require an AG at the 3' splice acceptor site. *Molecular and Cellular Biology*, *13*, 626–637. doi:10.1128/mcb.13.1.626-637.1993.

Berget, M. S. (1995). Exon recognition in vertebrate splicing. *The Journal of Biological Chemistry*, *270*, 2411–2414. doi:10.1074/jbc.270.6.2411.

Bergman, C. M., & Kreitman, M. (2001). Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Research*, *11*, 1335–1345. doi:10.1101/gr.178701.

Bergman, J., Betancourt, A. J., & Vogl, C. (2017). Transcription-associated compositional skews in *Drosophila* genes. *Genome Biology and Evolution*, *10*, 269–275. doi:10.1093/gbe/evx200.

Borges, R., Szöllősi, G. J., & Kosiol, C. (2019). Quantifying GC-biased gene conversion in great ape genomes using Polymorphism-Aware models. *Genetics*, *212*, 1321–1336. doi:10.1534/genetics.119.302074.

Breathnach, R., & Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. *Annual Review of Biochemistry*, *50*, 349–383. doi:10.1146/annurev.bi.50.070181.002025.

Cartegni, L., Chew, S., & Krainer, A. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Review Genetics*, *3*, 285–298. doi:10.1038/nrg775.

Clemente, F., & Vogl, C. (2012). Unconstrained evolution in short introns? - an analysis of genome-wide polymorphism and divergence data from *Drosophila*. *Journal of Evolutionary Biology*, *25*, 1975–1990. doi:10.1111/j.1420-9101.2012.02580.x.

Coolidge, C. J., Seely, R. J., & Patton, J. G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Research*, *25*, 888–896. doi:10.1093/nar/25.4.888.

Deger, N., Yang, Y., Lindsey-Boltz, L., Sancar, A., & Selby, C. (2019). *Drosophila*, which lacks canonical transcription-coupled repair proteins, performs transcription-coupled repair. *Journal of Biological Chemistry*, *294*, 8092–18098. doi:10.1074/jbc.AC119.011448.

Farlow, A., Dolezal, M., Hua, L., & Schlötterer, C. (2012). The genomic signature of splicing-coupled selection differs between long and short introns. *Molecular Biology and Evolution*, *29*, 21–24. doi:10.1093/molbev/msr201.

Grabowski, P. J., Padgett, R. A. ., & Sharp, P. A. (1984). Messenger RNA splicing in vitro: an excised intervening sequence and a potential intermediate. *Cell*, *37*, 415–427. doi:10.1016/0092-8674(84)90372-6.

Green, M. R. (1986). Pre-mRNA splicing. *Annual Review of Genetics*, *20*, 671–708. doi:10.1146/annurev.ge.20.120186.003323.

Green, P., Ewing, B., Miller, W., Thomas, J., Pamela, Program, N. C. S., & Green, E. D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, *33*, 514–517. doi:10.1038/ng1103.

Guo, M., Lo, C. P., & Mount, M. S. (1993). Species-specific signals for the splicing of a short *Drosophila* intron in vitro. *Molecular and Cellular Biology*, *13*, 1104–1118. doi:10.1128/mcb.13.2.1104-1118.1993.

Haddrill, P. R., Charlesworth, B., Halligan, D. L., & Andolfatto, P. (2005). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology*, *6*, R67. doi:10.1186/gb-2005-6-8-r67.

Halligan, D. L., Eyre-Walker, A., Peter, A., & Keightley, P. D. (2004). Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Research*, *14*, 273–279. doi:10.1101/gr.1329204.

Halligan, D. L., & Keightley, P. D. (2006). Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, *16*, 875–884. doi:10.1101/gr.5022906.

Hu, T. T., Eisen, M. B., Thornton, K. R., & Andolfatto, P. (2013). A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, *23*, 89–98. doi:10.1101/gr.141689.112.

Jackson, B. C., Campos, J. L., Haddrill, P. R., Charlesworth, B., & Zeng, K. (2017). Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in Drosophila. *Genome Biology and Evolution*, *9*, 102–123. doi:10.1093/gbe/evw291.

Jaillon, O., Bouhouche, K., Gout, J., Aury, J.-M., Noel, B., Saudemont, B., Nowacki, M., Serrano, V., Porcel, B. M., Ségurens, B., Le Mouël, A., Lepère, G., Schächter, V., Bétermier, M., Cohen, J., Wincker, P., Sperling, L., Duret, L., & Meyer, E. (2008). Translational control of intron splicing in eukaryotes. *Nature*, *451*, 359–362. doi:10.1038/nature06495.

Katz, D., Baptista, J., Azen, P., S, & Pike, C., M (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, *34*, 469–474. doi:10.2307/2530610.

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., Langley, C. H., & Pool, J. E. (2015). The *Drosophila* genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics*, *199*, 1229–1241. doi:10.1534/genetics.115.174664.

Lawrie, D. S., Messer, P. W., Hershberg, R., & Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics*, *9*, 1–18. doi:10.1371/journal.pgen.1003527.

Lawrie, D. S., & Petrov, D. A. (2014). Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, *30*, 133–139. doi:10.1016/j.tig.2014.02.002.

Long, M., & Deutsch, M. (1999). Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. *Molecular Biology and Evolution*, *16*, 1528–1534. doi:10.1093/oxfordjournals.molbev.a026065.

Long, M., Rosenberg, C., & Gilbert, W. (1995). Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences, USA*, *92*, 12495–12499. doi:10.1073/pnas.92.26.12495.

Ludwig, M. Z. (2002). Functional evolution of noncoding DNA. *Current Opinion in Genetics and Development*, *12*, 634–639. doi:10.1016/S0959-437X(02)00355-6.

Machado, H. E., Lawrie, D. S., & Petrov, D. A. (2020). Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. *Genetics*, *214*, 511–528. doi:10.1534/genetics.119.302542.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, *19*, 330–338. doi:10.1016/S0168-9525(03)00116-1.

Mitchell, D., & Bridge, R. (2006). A test of Chargaff's second rule. *Biochemical and Biophysical Research Communications*, *340*, 90–94. doi:10.1016/j.bbrc.2005.11.160.

Mount, M. S. (1982). A catalogue of splice junction sequences. *Nucleic Acids Research*, *10*, 459–472. doi:10.1093/nar/10.2.459.

Mount, M. S., Burks, C., Herds, G., Stormo, D. G., White, O., & Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Research*, *20*, 4255–4262. doi:10.1093/nar/20.16.4255.

Nguyen, H., & Xie, J. (2019). Widespread separation of the polypyrimidine tract from 3' AG by G tracts in association with alternative exons in metazoa and plants. *Frontiers in Genetics*, *9*. doi:10.3389/fgene.2018.00741.

Padgett, R., Grabowski, P., Konarska, M., Seiler, S., & Sharp, P. (1986). Splicing of messenger RNA precursors. *Annual Review of Biochemistry*, *55*, 1119–1150. doi:10.1146/annurev.bi.55.070186.005351.

Padgett, R., Konarska, M., Grabowski, P., Hardy, S., & Sharp, P. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science*, *225*, 898–903. doi:10.1126/science.6206566.

Parsch, J. (2003). Selective constraints on intron evolution in *Drosophila*. *Genetics*, *165*, 1843–1851. doi:10.1093/genetics/165.4.1843.

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., & Andolfatto, P. (2010). On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution*, *27*, 1226–34. doi:10.1093/molbev/msq046.

Pennacchio, L. A., & Rubin, E. M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nature Review Genetics*, *2*, 100–109. doi:10.1038/35052548.

Riddle, D., Blumenthal, T., Meyer, B., & Priess, J. (1997). *C. elegans II*. (2nd ed.). N.Y.: Cold Spring Harbor.

Rogers, R. L., Cridland, J. M., Shao, L., Hu, T. T., Andolfatto, P., & Thornton, K. R. (2014). Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molecular Biology and Evolution*, *31*, 1750–1766. doi:10.1093/molbev/msu124.

Rong, S., Buerer, L., Rhine, C. L., Wang, J., Cygan, K. J., & Fairbrother, W. G. (2020). Mutational bias and the protein code shape the evolution of splicing enhancers. *Nature Communications*, *11*, 2845. doi:10.1038/s41467-020-16673-z.

Ruskin, B., & Green, M. R. (1985). Role of the 3' splice site consensus sequence in mammalian pre-mRNA splicing. *Nature*, *317*, 732–734. doi:10.1038/317732a0.

Ruskin, B., Greene, J. M., & Green, M. R. (1985). Cryptic branch point activation allows accurate in vitro splicing of human β-globin intron mutants. *Cell*, *41*, 833–844. doi:10.1016/S0092-8674(85)80064-7.

Ruskin, B., Krainer, A. R., Maniatis, T., & Green, M. R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell*, *38*, 317–331. doi:10.1016/0092-8674(84)90553-1.

Schwartz, S. H., Silva, J., Burstein, D., Pupko, T., Eyras, E., & Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Research*, *18*, 88–103. doi:10.1101/gr.6818908.

Shepelev, V., & Fedorov, A. (2006). Advances in the exon-intron database (EID). *Briefings in Bioinformatics*, *7*, 178–185. doi:10.1093/bib/bbl003.

Spellman, R., Rideau, A., Matlin, C., A.and Gooding, Robinson, F., McGlincy, N., Grellscheid, S. N., Southby, J., Wollerton, M., & Smith, C. W. J. (2005). Regulation of alternative splicing by PTB and associated factors. *Biochemical Society Transactions*, *33*, 457–460. doi:10.1042/BST0330457.

Stephan, W., Rodriguez, V. S., Zhou, B., & Parsch, J. (1994). Molecular evolution of the metallothionein gene *Mtn* in the *melanogaster* species group: Results from *Drosophila Ananassae*. *Genetics*, *138*, 135–143. doi:10.1093/genetics/138.1.135.

Talerico, M., & Berget, M. S. (1994). Intron definition in splicing of small *Drosophila* introns. *Molecular and Cellular Biology*, *14*, 3434–3445. doi:10.1128/mcb.14.5.3434-3445.1994.

Thanassoulis, G., & Vasan, R. (2010). Genetic cardiovascular risk prediction: will we get there? *Circulation*, *122*, 2323–2334. doi:10.1161/CIRCULATIONAHA.109.909309.

Touchon, M., Arneodo, A., d'Aubenton-Carafa, Y., & Thermes, C. (2004). Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Research*, *32*, 4969–4978. doi:10.1093/nar/gkh823.

Touchon, M., Nikolay, S., Arneodo, A., d'Aubenton-Carafa, Y., & Thermes, C. (2003). Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Letters*, *555*, 579–582. doi:10.1016/S0014-5793(03)01306-1.

Törmä, L., Burny, C., Volte, V., Senti, K., & Schlötterer, C. (2020). Transcription-coupled repair in *D. melanogaster* is independent of the mismatch repair pathway. doi:10.1101/2020.04.07.029033 preprint at https://www.biorxiv.org/content/10.1101/2020.04.07.029033v1.

Vogl, C., & Bergman, J. (2015). Inference of directional selection and mutation parameters assuming equilibrium. *Theoretical Population Biology*, *106*, 71–82. doi:10.1016/j.tpb.2015.10.003.

Zhang, C., Li, W., Krainer, A. R., & Zhang, M. Q. (2008). RNA landscape of evolution for optimal exon and intron discrimination. *Proceedings of the National Academy of Sciences*, *105*, 5797–5802. doi:10.1073/pnas.0801692105.

Supplement

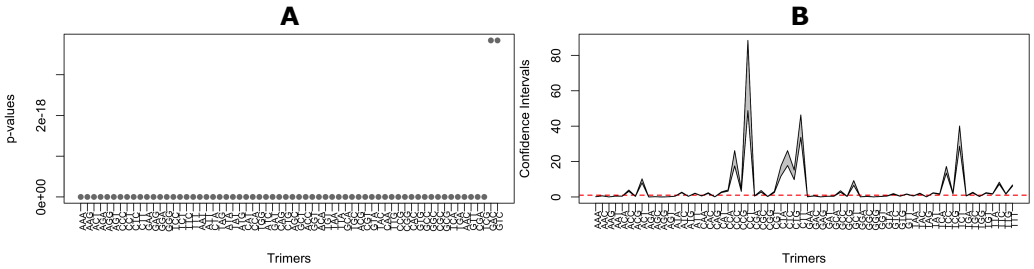


Figure S1: Tests of strand symmetric evolution in the 3PT. (A) Ordered p-values from chi-square tests, for the equality of forward and reverse complement trimers. (B) Confidence intervals for the ratios of forward and reverse complement trimers. Red dashed lines correspond to the tolerance range to assume equivalence.

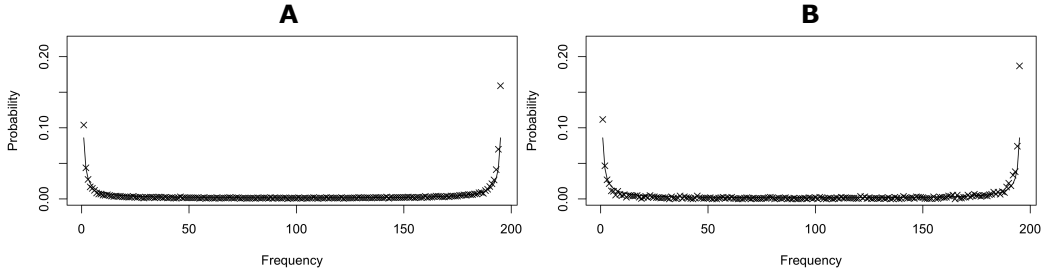


Figure S2: Biallelic spectra of GC-changing mutations (A/C, A/G, T/C, T/G) constructed based on their segregating GC frequency for (A) Autosomal short introns (B) X-linked short introns. For this analysis, only introns shorter than 66 bp were used. The lines represent expected site frequencies under neutral equilibrium; crosses represent observed site frequencies. The deviations are significant ($\chi^2_{df=194} = 1875.3$, $p < 0.001$ and $\chi^2_{df=194} = 518.81$, $p < 0.001$ for Autosome and X, respectively)

Table S1: Proportion of nucleotides (A, T, G, C) and mononucleotide asymmetry scores (S_{CG}, S_{TA}) in 5LR and 3PT of X-linked introns

	A	T	G	C	S_{CG}	S_{TA}
5LR	29.27%	31.45%	18.39%	20.89%	6.37%	3.58%
3PT	19.89%	44.70%	9.58%	25.83%	45.87%	38.41%

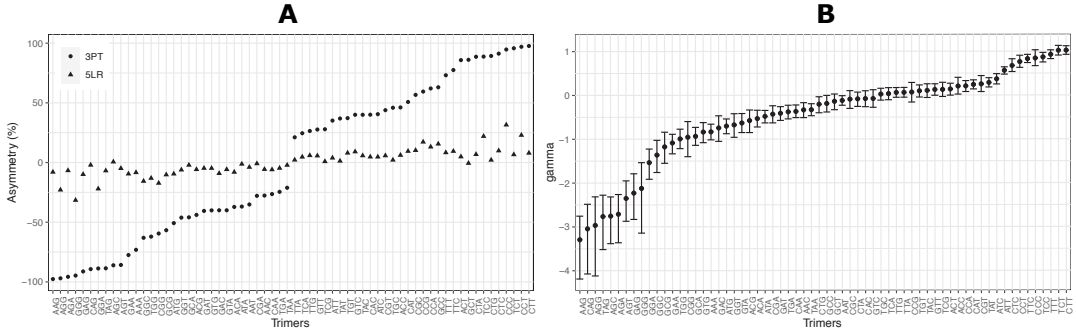


Figure S3: Results for X-linked introns. (A) Asymmetry scores per trimer, per region. Circles represent 3PT and triangles represent 5LR (B) Selection coefficients of each trimer in 3PT. Error bars represent the 95% CIs from 1000 bootstraps of the datasets.

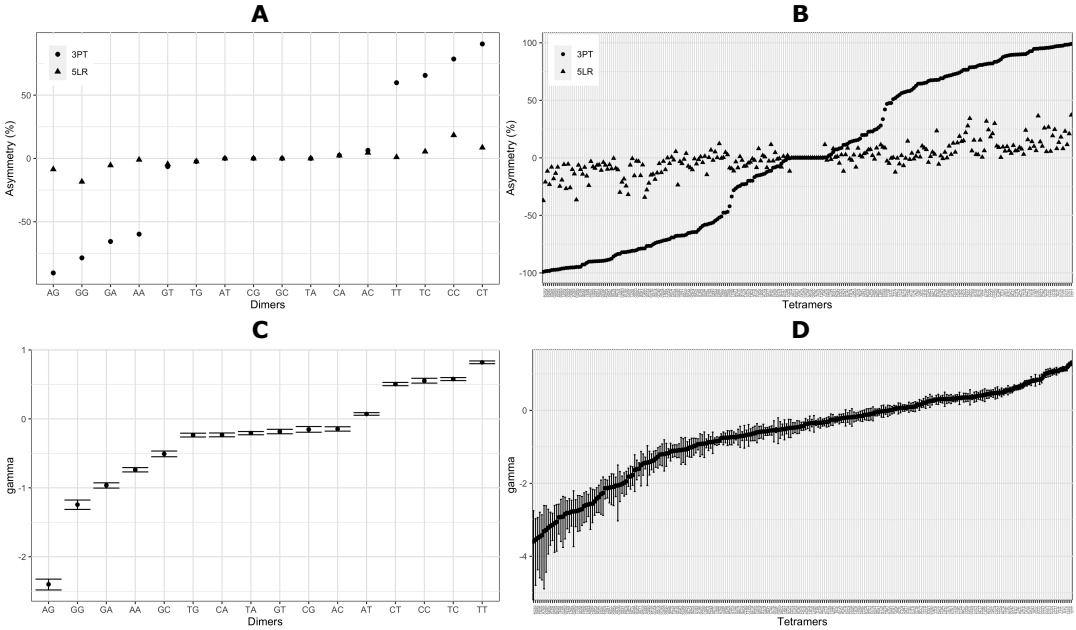
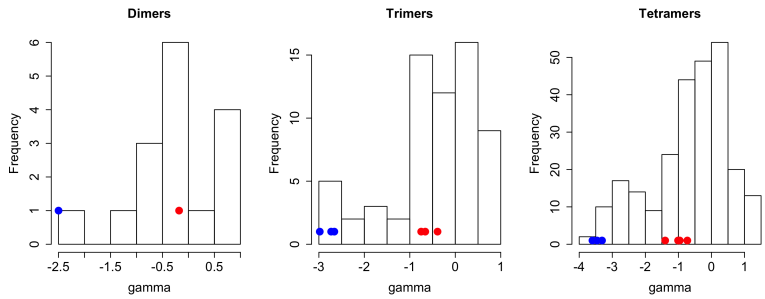


Figure S4: Results for dimers (A-C) and tetramers (B-D). (A-B) Asymmetry scores per region, per dimer and per tetramer, respectively. Circles represent 3PT and triangles represent 5LR (C-D) Selection coefficients of each dimer and tetramer in the 3PT, respectively. Error bars represent the 95% CIs from 1000 bootstraps of the original datasets.



3' splice signal Y**AG**|G	**AG**	**AGG, TAG, CAG**	**TAGG, CAGG, AGGG**
5' splice signal G|**GT**RAG	**GT**	**GGT, GTG, GTA**	**GTAA, GTGA, GGTA, GGTG**

Figure S5: Selection strength distributions for dimer (16), trimer (64) and tetramer (256) motifs. Dots in the graphs correspond to the gamma values for specific motifs from 3' splice signal (blue) or 5' splice signal (red). The lower panel shows the pinned motifs in histograms.

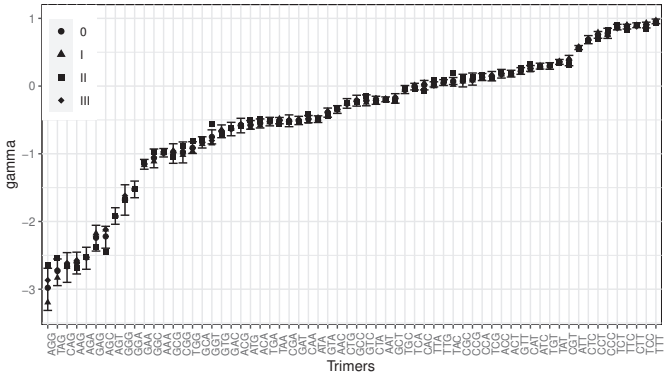


Figure S6: Selection coefficients of each trimer in the 3PT for different datasets. 0 (circle): original dataset with all short introns. I (triangle): dataset including only non-phase 0 (3n+1, 3n+2) introns. II (square): dataset including only phase 0 (3n) introns. III (diamonds): dataset excluding the common phase 0 introns between *D. melanogaster* and *D. simuans*. Error bars represent the 95% CIs from 1000 bootstraps of the original datasets.

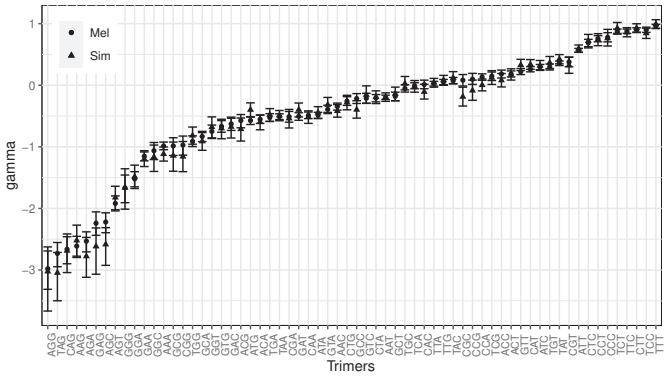


Figure S7: Selection coefficients of each trimer in the 3PT for different species. Circles are estimates from *D. melanogaster*, while triangles are estimates from *D. simuans*. Error bars represent the 95% CIs from 1000 bootstraps of the datasets.

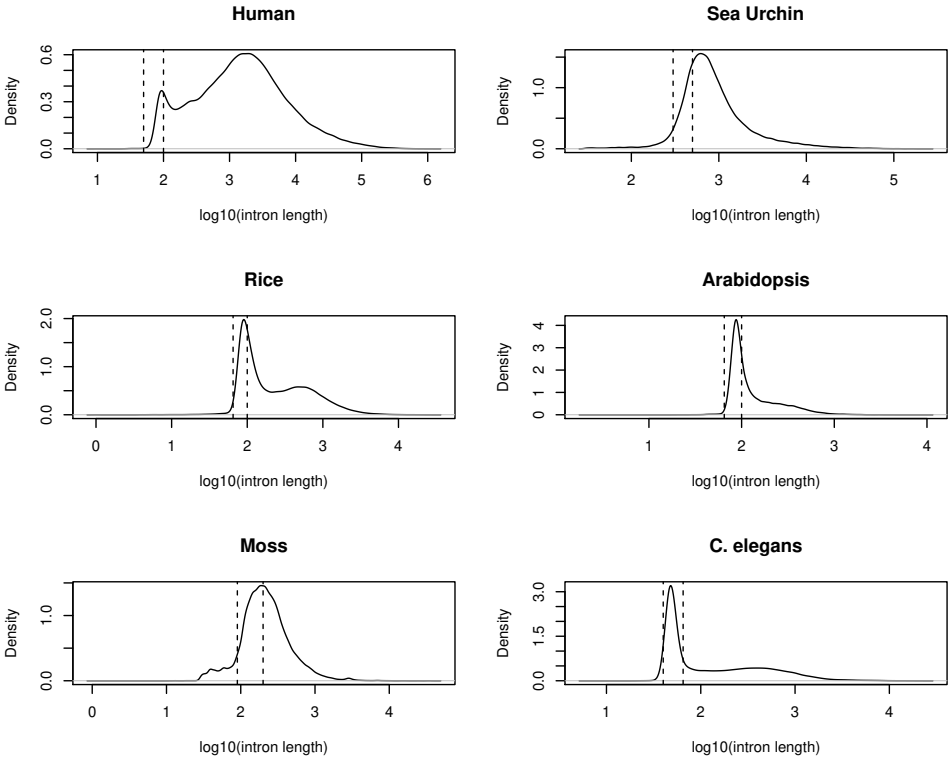


Figure S8: The length distribution of introns obtained from exon-intron database, EID (Shepelev & Fedorov, 2006). For each species, vertical dashed lines represent the length range of extracted and analyzed short intron class. Human: 50-100 bp, Sea Urchin: 300-500 bp, Rice: 65-100 bp, *Arabidopsis*: 65-100bp, Moss: 90-200 bp, *C. elegans*: 40-65 bp.

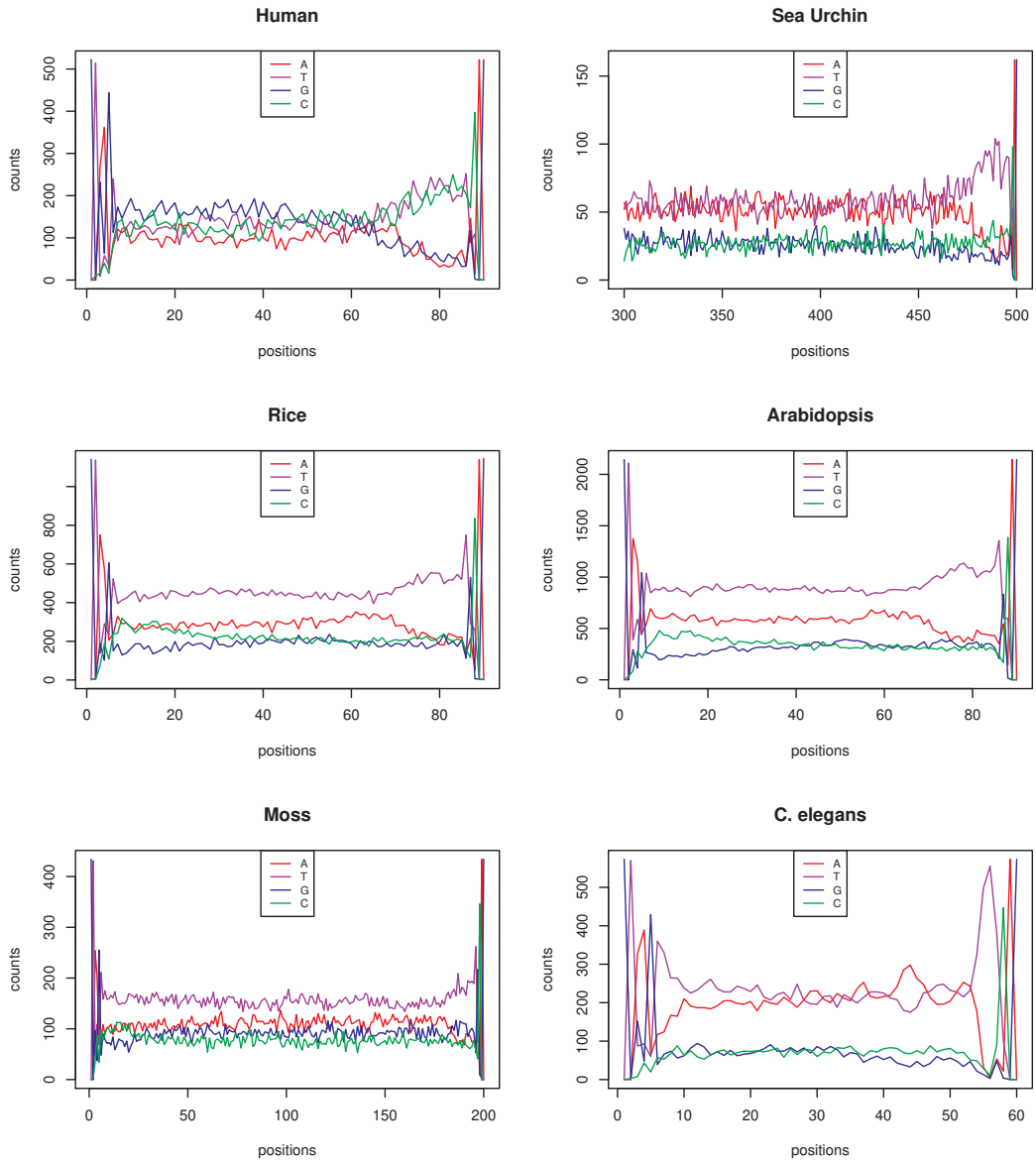


Figure S9: Nucleotide composition of introns with length 90 for human, 500 for sea urchin, 90 for rice and *Arabidopsis*, 200 for moss, and 60 for *C. elegans*. Only one candidate length class were chosen to visualize pattern. Total numbers of introns used for each species are 523, 162, 1152, 2145, 434 and 573, respectively.

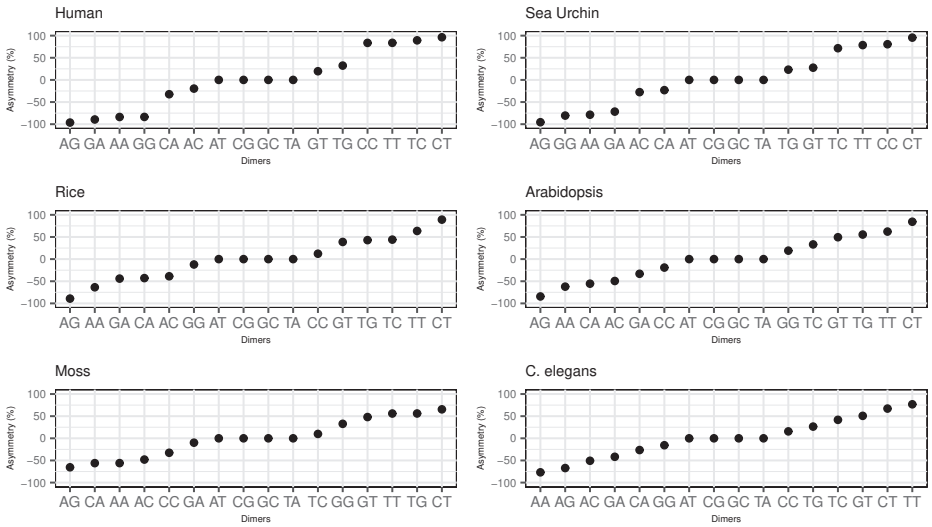


Figure S10: Asymmetry scores of dimers from the pyrimidine enriched regions in the short introns of 6 eukaryotic species.

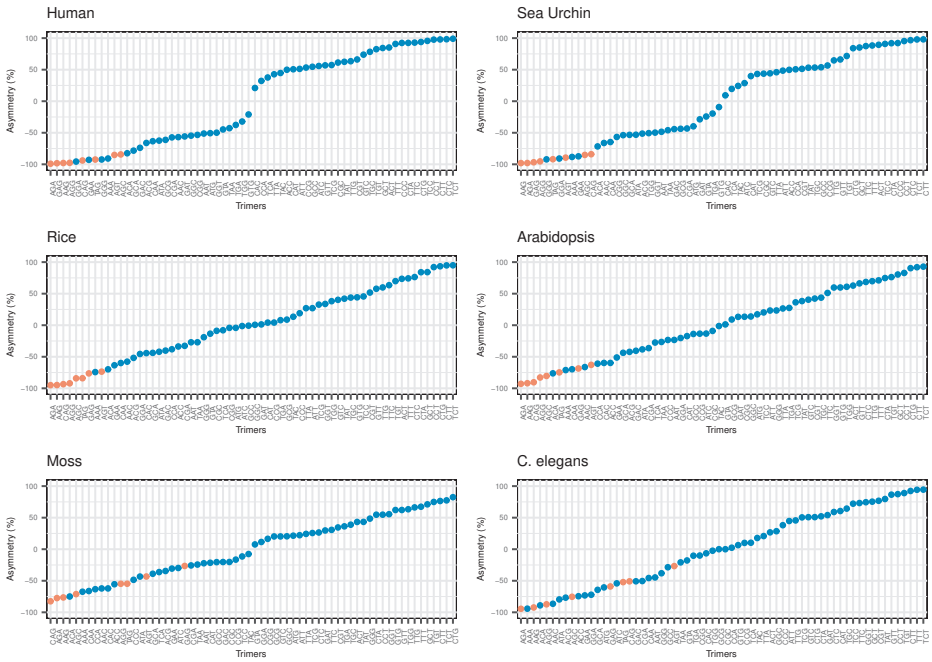


Figure S11: Asymmetry scores of trimers from the pyrimidine enriched regions in the short introns of 6 eukaryotic species. Motifs containing *AG* in it are depicted with red points, while non-*AG* motifs are represented with blue.