# Intron Length and Recursive Sites Are Major Determinants of Splicing Efficiency in Flies

**February 12, 2017**

**Athma A. Pai[1], Telmo Henriques[2], Joseph Paggi[1], Adam Burkholder[3], Karen Adelman[2,4], Christopher B. Burge[1,*]**

1. Departments of Biology and Biological Engineering,
Massachusetts Institute of Technology
Cambridge, MA 02142
2. Epigenetics and Stem Cell Biology Laboratory and
3. Center for Integrative Bioinformatics
NIEHS, Research Triangle Park, NC 27709
4. Department of Biological Chemistry and Molecular Pharmacology,
Harvard Medical School, Boston, MA 02115

* Address correspondence to: cburge@mit.edu

## Abstract

**The dynamics of gene expression may impact regulation, and RNA processing can be rate limiting. To assess rates of pre-mRNA splicing, we used a short, progressive metabolic labeling/RNA sequencing strategy to estimate the intron half-lives of ~30,000 fly introns, revealing strong correlations with several gene features. Splicing rates varied with intron length and were fastest for modal intron lengths of 60-70 nt. We also identified hundreds of novel recursively spliced segments, which were associated with much faster and also more accurate splicing of the long introns in which they occur. Surprisingly, the introns in a gene tend to have similar splicing half-lives and longer first introns are associated with faster splicing of subsequent introns. Our results indicate that genes have different intrinsic rates of splicing, and suggest that these rates are influenced by molecular events at gene 5' ends, likely tuning the dynamics of developmental gene expression.**

To globally assess the kinetics of gene expression in *Drosophila melanogaster*, we applied a short time period metabolic labeling strategy involving 5, 10, or 20 min labeling with 4-thio-uracil (4sU) in S2 cells, followed by RNA sequencing[1]. These data were complemented by steady state RNA-seq data from 24 h 4sU-labeled RNA representing predominantly mature mRNA (Methods). As expected, read density in introns was highest at the 5 min time point and decreased rapidly to low levels at longer times (Fig. 1A). To assess the kinetics of splicing, we measured the proportion of intron-containing transcripts or percent spliced in (PSI or $\Psi$) value of each intron – representing the proportion of transcripts of a gene that containing the intron in an unspliced state – using the MISO software[2]. Intron PSI values decreased over time for > 98% of introns, reflecting the progress of splicing (Supplementary Fig. 1A).

The labeling design results in the isolation of transcripts that initiated transcription during the labeling period, as well as some transcripts that initiated earlier and continued their elongation during this period. Therefore, we inferred the mean time since synthesis for each intron captured under each given labeling regime, taking into account rate of transcription, and used these times in estimation of intron half-lives (Methods). Using these inferred synthesis times, we fit a first-order exponential decay model to PSI values (Fig. 1B) to obtain intron half-

lives with associated confidence intervals for 29,809 introns in 6,110 *Drosophila* genes with sufficient expression in S2 cells (Supplementary Fig. 1B-D; Supplementary Table 1). The median intron half-life ($t_{½}$) was 4.0 minutes. We obtained similar PSI values when using exclusively junction-spanning reads to assess splicing (Supplementary Fig. 1E), indicating minimal impact on estimated rates from lariat-derived reads, as expected (intron lariat half-lives < 15 seconds[3,4]). Therefore, the intron half-lives represent splicing half-times. In the remainder of this study we used MISO-derived splicing half-times because MISO PSI values fit better to exponential decay models than those based on junction reads alone (Supplementary Fig. 1C).

Our estimated rates of intron removal are consistent with previous reports of 0.5 - 10 minutes per intron in a handful of mammalian genes using qRT-PCR or imaging approaches[4-6], but slightly longer than recently estimated rates of < 2 min in yeast[7]. For typical *Drosophila* genes of 3-9 kb in length, the estimated time of transcription is about 2-6 minutes[8]. Thus, the intron half-lives we estimate (with median $t_{½}$ = 4 min) are consistent with common but not universal co-transcriptional splicing[9-11]. Based on simulations, our estimates are reasonably accurate and unbiased for half-lives between 3 and 300 min but have reduced accuracy for introns with splicing half-times of < 3 min or > 300 min (Supplementary Methods, Supplementary Fig. 1F). A half-life of 3 min would typically produce a PSI < 0.10 at the 10 min time point, and 9% of introns had PSI < 0.10 at this time point. This observation suggests that about 1 in 10 fly introns has a splicing half-time of less than 3 min; an alternative approach for half-life estimation was used for this subset to reduce bias (Methods). A similar analysis using data from the long time points suggests that >99.7% of analyzed fly introns have half-lives < 300 min, and our approach using MISO may somewhat overestimate the half-lives of longer introns for which the time required to transcribe the intron approaches or exceeds the duration of labeling (see Methods). For this reason, we used alternative analyses to study the splicing of very long introns by recursive mechanisms below.

The distribution of lengths of *Drosophila* introns has a sharp peak at 60-70 nt, with more than half of introns between 40-80 nt in length, and the remainder distributed over a broad range extending to tens of kilobases (kb) or more[12] (Supplementary Fig. 2A). This observation and evidence that natural selection favors short intron lengths in *Drosophila*[13-15] led us to hypothesize that introns with lengths close to the modal size of ~65 nt are spliced most efficiently[12]. We observed a strong relationship between intron length and splicing half-times (Spearman rho $P < 2.2 \times 10^{-16}$, Fig. 1C), with 30.7% of variance in splicing rates explained by intron length (Supplementary Methods). Specifically, introns with lengths in the range 60-70 nt were spliced most rapidly (median $t_{½}$ = 3.3 min, Fig. 1D), and median $t_{½}$ increased steadily to

20.3 min for introns > 10 kb. Genes containing introns with the shortest half-lives (< 3 min) had expression levels 2- to 5-fold higher than genes with slower-spliced introns of similar lengths, suggesting a relationship between expression and splicing rate.

Notably, for introns shorter than 60 nt, a negative relationship between intron length and splicing half-time was observed, and "ultra-short" introns with lengths between 40-50 nt had a median $t_{½}$ of 9.2 min, about 2.5-fold longer than that for the 60-70 nt class (Fig. 1D). Though ultra-short introns also have significantly weaker 5' and 3' splice sites[16] (Supplementary Fig. 2B), differences in splice site strength do not fully explain why introns < 50 nt are spliced so slowly (Supplementary Fig. 2C), suggesting that small size itself presents a barrier to efficient splicing[17]. Introns longer than 70 nt tend to have stronger splice sites (Supplementary Fig. 2D) and yet are spliced more slowly (whether controlling for splice site strength or not, Supplementary Fig. 2E), suggesting that bringing together intron ends for splicing may contribute materially to the time required for splicing[18]. However the time required to juxtapose intron ends cannot explain why introns with 40-50 nt lengths are spliced 2.5 times more slowly than those in the "optimal" 60-70 nt range, suggesting that some other factor – perhaps steric interference between snRNPs bound at the 5' splice site and branch site[19,20] – becomes suboptimal for the shortest *Drosophila* introns.

Metazoan introns are thought to be spliced in one of two modes – either by "intron definition", in which the U1 and U2 snRNPs that recognize the 5' splice site and branch point sequence first pair across the intron; or "exon definition", in which U1 snRNP initially pairs with the upstream U2 snRNP across the exon, followed by reassortment to form interactions with the downstream U2 snRNP across the intron[21]. Both modes occur in *Drosophila*, depending on exon-intron geometry, with short introns favoring intron definition and long introns/short exons favoring exon definition[21], so the overall slower splicing of long introns likely results, at least in part, from the extra time required for the additional molecular events involved in exon definition.

In mammals, it has been observed that PolII elongation rates are slower near regulated introns and that alternative introns are spliced more slowly[22-25]. Comparing to constitutive introns (CI), we observed that introns that are alternatively retained (RI) in other cells/tissues (but fully spliced out in S2 cells) have somewhat longer splicing half-times, independent of intron length, as do introns that flank exons that are alternatively skipped in other cells/tissues (SEflanking) (all Mann-Whitney $P < 0.01$ for comparisons between CI & RI, except in largest bin, and for all comparisons between CI & SEflanking; Fig. 1E). This observation suggests that the capacity for regulation may impose limits on splicing rate.

The long half-lives for the longest introns suggests that auxiliary splicing mechanisms might be critical for efficient splicing of very long introns in *Drosophila*. In particular, recursive splicing is a process associated with long introns in which a single intron is removed in multiple splicing events defined by recursive sites, which consist of juxtaposed 3' and 5' splice site motifs around a central AG/GT[26,27]. Some 130 recursively spliced introns have previously been identified in flies based on analysis of over 10 billion RNA-seq reads from the entire *Drosophila* ModENCODE project[28], suggesting that this phenomenon is relatively rare. However, the transient nature of recursive splicing intermediates makes it difficult to detect evidence for recursive splicing using standard RNA-seq data.

We hypothesized that our high-coverage nascent RNA data would more readily identify transient recursive events and better characterize the prevalence of recursive splicing. To do so, we used a computational pipeline to confidently identify recursive events using three key features (Fig. 2A, Supplementary Fig. 3). First, we searched for splice junction reads derived from putative recursive sites (RatchetJunctions), as previously described[28]. Second, we developed a new computational tool, RatchetPair, to identify read pairs that map to sites flanking putative recursive segments in manner where presence of splicing can be inferred from the size distribution of library fragments. Third, we developed the first automated software, RatchetScan, for inference of recursive sites from sawtooth patterns in read density. This type of pattern is an expected product of co-transcriptional recursive splicing and has been observed as a feature associated with many recursive introns[10,28].

Combining these three approaches, our analysis detected 541 candidate recursive sites in 379 fly introns (Supplementary Table 2). From this set, we curated a set of 243 “high confidence” recursive sites in 157 introns (with an FDR of 5%), and a “medium confidence” set of 298 sites (at an FDR of 20%; Methods). Overall, 98 introns contained multiple recursive sites, with up to seven such high-confidence sites observed in a single intron. For instance, intron 1 of the tenascin major (*Ten-m*) gene contains five recursive sites, two of which were previously unknown (Fig. 2A). Of the recursive sites previously reported by Duff *et al.*, 124 occurred in genes expressed in S2 cells. Our approach detected 119 (96%) of these known sites, as well as 126 novel high confidence sites and 296 novel medium confidence sites (Fig. 2B), thus increasing the number of recursive sites defined in this cell type by ~4-fold (Fig. 2B, Supplementary Fig. 3C). Both the high confidence and the medium confidence candidate recursive sites exhibited a strong juxtaposed 3'/5' splice site motif (Supplementary Fig. 4A). The greater numbers detected by our approach (2-4X more sites in this cell type), using less than

1/20$^{th}$ as many reads as used by Duff and colleagues, confirms the potential of nascent RNA analysis for analysis of recursive splicing.

Many very long introns (> 40 kb in length) have recursive sites, with 63% of such introns containing at least one high-confidence recursive site, and 70% when considering all identified recursive sites (Fig. 2C). This observation suggests that recursive splicing is the prevalent mechanism by which very large fly introns are excised. We assessed the sensitivity of our detection pipeline by running it on subsamples of reads ranging from 0.1% to 100% of the total (Fig. 2D). The shape of the resulting curve tapered off at higher coverage levels but never plateaued, indicating that new recursive sites were still being detected as read depth increased from 50% to 100% and therefore would likely increase further at higher read depths. A somewhat higher proportion of recursive sites were detected in high-expressed genes (TPM > 20) than low-expressed genes (TPM ≤ 20). However, subsampling of the reads mapping to high-expressed genes to levels comparable to those observed for low-expressed genes resulted in a substantially lower fraction of recursive sites at each depth, suggesting that very long introns in low-expressed genes are more likely to have recursive splicing than those in high-expressed genes (Fig. 2D). Together, these data suggest that the true fraction of very long introns that contain recursive sites may be substantially higher than our observed fraction of 63-70%, i.e. recursive splicing may be nearly universal in very long introns.

Recursive splice sites can be required for the processing of long introns[27]. The surprisingly widespread occurrence of recursive splice sites observed here raises the possibility that this mode of splicing has a substantial impact on processing of the fly transcriptome. Alternatively, it is possible that most recursive sites are functionally neutral, and that mRNA production is not impacted by their presence. The size of our dataset enabled us to examine four properties of recursive sites that could help to distinguish between these possibilities: sequence conservation; distribution in the fly genome; distribution within introns; and kinetics of splicing. In each case, the patterns observed suggest that most or all recursive sites have functional impact.

Both high and medium confidence recursive sites exhibited twice the level of evolutionary conservation observed in and around control AGGT motifs in long introns (Supplementary Fig. 4B), implying strong selection to maintain most or all of these sites. Recursively spliced introns were enriched in genes involved in functions related to development and morphogenesis (Supplementary Table 3). Both of these observations are consistent with results from a previous study based on a smaller sample of recursive introns[28].

It is still possible that longer introns contain more recursive sites purely by chance, rather than due to functional constraints on intron architecture. Indeed, while the majority of recursively spliced introns had just one recursive site, the number of sites increased roughly linearly with intron length (Supplementary Fig. 4C). However, the positioning of recursive sites within introns was significantly biased away from a random (uniform) distribution (Kolmogorov-Smirnov $P$ = 0.003; Fig. 2E). Instead, recursive sites in introns with only one such site tended to be located closer to the midpoint of the intron than expected by chance. Furthermore, the first recursive site in introns with two or three such sites tended to be located at approximately 33% and 25% of the way from the 5' end of the intron, respectively (Fig. 2E inset). The distribution of recursive sites within introns suggests that they are positioned so as to break the larger intron into “bite-sized” segments for the spliceosome (typically ~9-15 kb in length) rather than at random locations which would more often produce much longer and much shorter segments. Recursively spliced introns were also enriched in first introns relative to subsequent introns in fly genes (hypergeometric P < 0.05).

To ask whether recursive splicing contributes to the efficiency of processing of very long introns, we estimated the splicing half-times of individual recursive segments (Methods). Recursive segment half-times were the slowest for the first segment in the intron (Supplementary Fig. 5A), and the preponderance of junction reads spanning the 5' splice site and the recursive site indicate that recursive segments are most often spliced out in a 5' to 3' order (Supplementary Fig. 5B). Overall, recursive segments (mean length 9.1 kb, Supplementary Fig. 5C) had three-fold shorter half-times than non-recursive introns of the same length (Fig. 2F), supporting the conclusion that recursive splicing increases the speed of splicing for very long *Drosophila* introns. The overall rate of splicing of a recursive intron is likely comparable to that of its slowest recursive segment, whose splicing is expected to be rate limiting. Splicing half-times for the slowest segment of recursive introns were also at least two-fold shorter than those of non-recursive introns with lengths matching the entire recursive intron, again suggesting that recursive sites enhance efficiency of splicing (Supplementary Fig. 5D; Mann-Whitney $P < 2.2 \times 10^{-16}$). A more involved analysis, estimating the mean lifetime of a recursive intron as the maximum of a set of exponentials (corresponding to the waiting time for all recursive segments to be spliced) also yielded significantly shorter half-times for recursive than non-recursive introns of similar size (Mann-Whitney $P$ = 0.0013; Supplementary Fig. 5E).

Splicing accuracy is likely to be at least as important as speed, since splicing to an arbitrary (incorrect) splice junction will most often produce an mRNA that does not encode functional protein. As a simple measure of potential splicing errors, we tallied the fraction of

reads that spanned “non-canonical” splice junctions, involving pairs of intron terminal dinucleotides other than the three canonical pairs “GT-AG”, “GC-AG” and “AT-AC” that account for ~99.9% of all known fly introns. For the bulk of non-recursive introns (most of which are < 100 nt in length), the frequency of such non-canonical splicing was negligible (Fig. 2G, light grey dotted curve). However, for non-recursive introns with the much longer lengths typical of recursively spliced introns, potential splicing errors were much more frequent (Fig. 2G, gray curve), suggesting that the fly spliceosome loses accuracy as intron length increases. Notably, recursive introns had significantly lower frequencies of non-canonical junctions compared to similarly sized non-recursive introns (Fig. 2G, gold curve, Kolmogorov-Smirnov $P = 0.015$). Therefore, presence of recursive splice sites may increase the accuracy as well as the speed of splicing.

Together, the results shown in Figures 1 and 2 indicate that intron length explains about 30% of the variance in splicing rates and that presence of recursive splice sites is associated with a two-fold reduction in half-life. To explore what other factors might influence splicing efficiency of an intron, we used a multiple linear regression model to estimate the contribution of various candidate features to splicing half-time. We restricted our analysis to non-first introns within a narrow length range (60-70 nt) to exclude effects of length and focus on other variables (Supplementary Methods; Fig. 3A; Supplementary Figure 9). Overall, these other variables accounted for 29.5% of the variance in splicing rates. As expected, increased strength of both the 5' and 3' splice sites was associated with shorter intron half-life[29]. Length of the upstream exon and A+U content of the intron were also associated with shorter half-life. Longer upstream exons are expected to promote intron definition, which may speed splicing. Introns are U-rich in many metazoans and plants, and many proteins of the heterogeneous nuclear ribonucleoprotein (hnRNP) family bind A+U-rich motifs in introns and participate in splicing[30,31]. Surprisingly, however, two of the three most important features were properties of the gene rather than of the affected intron: gene expression and first intron length, both of which were associated with faster splicing of non-first introns in the gene. The relationship between gene expression and intron splicing rate is consistent with our observation that the fastest splicing introns tend to occur in higher-expressed genes.

The importance assigned to gene expression and the length of the first intron in a transcript, both gene-specific rather than intron-specific properties, raised the question of the relationship between splicing rates of different introns in the same gene. Since all of the introns in a transcript must be removed to produce a mature mRNA that can be exported and translated, natural selection (if present) may act primarily on the total elapsed time to splice all

introns rather than on the splicing rates of individual introns. We observed that splicing half-lives of introns drawn from the same gene tend to be more similar to each other than those of randomly sampled introns, e.g., comparing standard deviations (Mann-Whitney comparison of SDs, $P < 2.2 \times 10^{-16}$, Fig. 3B). This relationship remained when considering the coefficient of variation rather than the standard deviation (Supplementary Fig. 6B) and when controlling for variation in intron length (Supplementary Fig. 6C-D).

Genes with shorter median half-lives across introns (95$^{th}$ percentile, half-life < 2.0 min) are enriched for Gene Ontology categories involved in developmental processes (Supplementary Table 4), while those with the longest median half-lives across introns are enriched for metabolic processes (5$^{th}$ percentile, half-life > 10.8 min) (Supplementary Table 5). For instance, most of the introns in the metabolic gene *Cyp4d20* have half-lives in the 8-9 min range, while those in the developmentally-regulated gene *mthl4* are all between 3 and 4 min (Fig. 3C). These observations together with our results in Figure 3B indicate that properties of the gene impact the splicing rate of all introns, rather than each intron being independent in splicing rate of others in the gene. Such a relationship might result passively from differing levels of selection for efficient splicing experienced by different genes or classes of genes, or actively from gene-level features that impact the splicing efficiency of all introns in a gene.

It was intriguing that presence of a long first intron was associated with faster splicing of non-first introns within a gene. In *Drosophila*, first introns are twice as long on average as other introns (Supplementary Fig. 7A), and take 50% longer to splice than second, third, fourth or other intron positions. This trend persisted when controlling for intron length (Mann-Whitney $P < 2.2 \times 10^{-16}$, Fig. 3D, Supplementary Fig. 7B). A pattern of slower splicing of first introns has not been observed previously. In metazoans, transcriptional enhancers are often located in the first intron of genes[32,33], and we confirmed this trend in genes expressed in S2 cells (Supplementary Fig. 7C). Thus, the frequent presence of transcriptional enhancers may contribute to increased lengths of first introns. Presence of a transcriptional enhancer in an intron was associated with slower splicing of that intron (Supplementary Fig. 7D), but this association was not significant in the more complex model of splicing rate including other variables described above, so was not further explored here.

We observed a moderate but significant negative correlation between the length of the first intron in a gene and the median half-lives across downstream introns (Spearman rho $P < 4.4 \times 10^{-14}$, Fig. 3E, Supplementary Fig. 8A-B). This relationship remained when controlling for variability in the lengths of non-first introns by restricting the analysis to genes whose non-first introns were all between 60-70 nt (Supplementary Fig. 8C). For example, the gene *G$\alpha$i*,

encoding a G-protein alpha subunit, has a long (~5 kb) first intron that is slowly spliced ($t_{½}$ = 9 min), while the remaining introns are all spliced rapidly with half-lives between 1 and 3 min (Fig. 3C). These observations suggest that the architecture of the 5' end of the gene, notably the length of the 5'-most intron, impacts the splicing efficiency of the remaining introns. Longer first introns might allow additional time for the recruitment of factors that promote downstream splicing efficiency, consistent with the observation that transcription elongation in mammalian cells is slow before reaching the second exon[34].

Selection is likely to act on some genes to reduce the time to produce mature mRNA, particularly early in development when the time to produce proteins between cell cycles is short, and in stress response or signaling contexts where speed of response is important. Our findings suggest that a longer first intron requires more time to splice but may in some way enhance the efficiency of splicing of the remaining introns in the gene. Under this model, longer first introns should be favored in genes with larger numbers of introns, where the splicing time of the first intron is less likely to be rate limiting, because of the longer time required for the polymerase to reach the 3' end of the transcript and the greater likelihood that another intron's splicing is rate limiting. Consistent with this expectation, we observed a pronounced trend where genes with larger numbers of introns have longer first introns (Fig. 3F).

Taken together, our results present a picture in which the splicing rate of an intron is determined not only by features intrinsic to the intron itself, but also by properties of the gene in which the intron resides. Our analysis implicates an intron's length as a major contributor, explaining about 30% of the variance in intron half-life, with smaller contributions from splice site strength and intron composition. Gene expression level was the variable most strongly associated with shorter intron half-life in Figure 3A. When the rate of transcription initiation is higher, production of each transcript will occur closer to RNAs transcribed just prior. This proximity could lead to faster assembly of spliceosomes on new transcripts as a result of higher local concentration of spliceosome components recruited to and released from previously transcribed RNAs, particularly if splicing is generally co-transcriptional, as appears to be the case[10]. Curiously, increased length of the first intron – while requiring more time to process – was associated with faster splicing of the remaining introns of a gene, and with an increased number of introns per gene. Recursive sites, which appear to increase both the speed and accuracy of splicing, may be conserved in part to compensate for the increased splicing time and propensity for splicing error in long first introns.

# Methods

### Generation of 4sU RNA-seq

Newly transcribed RNA from 3 independent replicates of *Drosophila* S2 cells were labeled for 5, 10 and 20 minutes using 500 μM 4-thiouridine (Sigma, T4509). Additionally, two independent replicates of *Drosophila* S2 cells were labeled overnight with 4-thiouridine to approximate a steady-state RNA levels. To normalize samples and assess metabolic labeled RNA capture efficiency, several synthetic RNAs were spiked into the Trizol preparation at specific quantities per $10^6$ cells. Quantities were determined as described previously[35]. Total RNA was extracted with Trizol (Qiagen) and treated for 15 minutes with DNAseI amplification grade (Invitrogen) per manufacturer's instructions. To purify metabolic labeled RNA we used 300 μg total RNA for the biotinylation reaction. Separation of total RNA into newly transcribed and untagged pre-existing RNA was performed as previously described[1,36]. Specifically, 4sU-labeled RNA was biotinylated using EZ-Link Biotin-HPDP (Pierce), dissolved in dimethylformamide (DMF) at a concentration of 1 mg/ml. Biotinylation was done in labeling buffer (10 mM Tris pH 7.4, 1 mM EDTA) and 0.2 mg/ml Biotin-HPDP for 2 h at 25° C. Unbound Biotin-HPDP was removed by extraction with chloroform/isoamylalcohol (24:1) using MaXtract (high density) tubes (Qiagen). RNA was precipitated at 20,000*g* for 20 minutes with a 1:10 volume of 5 M NaCl and 2.5X volume of ethanol. The pellet was washed with ice-cold 75% ethanol and precipitated again at 20,000*g* for 5 minutes. The pellet was resuspended in 1 ml RPB buffer (300 mM NaCl, 10mM Tris pH 7.5, 1mM EDTA). Biotinylated RNA was captured using Streptavidin MagneSphere Paramagnetic particles (Promega). Before incubation with biotinylated RNA, streptavidin beads were washed 4 times with wash buffer (50 mM NaCl, 10 mM Tris pH 7.5, 1 mM EDTA) and blocked with 1% polyvinylpyrrolidone (Sigma) for 10 minutes with rotation. Biotinylated RNA was then incubated with 600 μl of beads with rotation for 30 min at 25° C. Beads were magnetically fixed and washed 5 times with 4TU wash buffer (1 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA, 0.1% Tween 20). Unlabeled RNA present in the supernatant was discarded. 4sU-RNA was eluted twice with 75 μl of freshly prepared 100 mM dithiothreitol (DTT). RNA was recovered from eluates by ethanol precipitation as described above. As per library preparation, RNA quality was assessed using a Bioanalyzer Nano ChIP (Agilent). Ribosomal RNA was removed prior to library construction by hybridizing to ribo-depletion beads that contain biotinylated capture probes (Ribo-Zero, Epicentre). RNA was then fragmented and libraries were prepared according to the TruSeq Stranded Total RNA Gold Kit (Illumina) using random hexamer priming. cDNA for the

two 'total' RNA samples were prepared using an equal mix of random hexamers and oligo-dT primers. Illumina spike-ins were also used for normalization among samples.

Libraries were sequenced on an Illumina HiSeq machine with paired-end 51 nt reads (100 nt reads for the 'total' RNA samples), generating an average of 126M read pairs per library. Reads for each sample were filtered, removing pairs where the mean quality score of one or both mates fell below 20. One million pairs were then extracted at random, and aligned to the dm3 reference assembly (4sU RNA-seq, due to enhanced coverage of introns) or an index composed of all FlyBase release 5.57 transcripts (total RNA-seq) using bowtie 0.12.8[37], allowing 2 mismatches, a maximum fragment length of 10kb, reporting uniquely mappable pairs only (`–m1 –v2 –X10000`). Mean fragment length and standard deviation was then assessed using CollectInsertSizeMetrics, a component of Picard Tools 1.62. All reads were subsequently aligned to dm3 with Tophat 2.0.4[38], utilizing bowtie1 as the underlying aligner, allowing up to 10 reported alignments, passing the fragment length and stdev calculated above, and setting the minimum and maximum intron size to those observed in the FlyBase 5.57 annotations (`--bowtie1 –g10 --min-intron-size 25 --max-intron-size 250000`). Strand-specific alignments were performed for the 4sU RNA-seq (`--library-type fr-firststrand`), while unstranded alignments were performed for the total RNA-seq (`--library-type unstranded`). Coverage tracks were generated for each using genomeCoverageBed, a component of bedtools v2.16.2[39].

Gene expression values (TPMs) in each replicate library were calculated using Kallisto[40] and the transcriptome annotations from FlyBase *Drosophila melanogaster* Release 5.57[41].

**Estimating splicing from 4sU-RNA-seq data**

To measure the efficiency at which each intron was spliced out in Drosophila S2 cells, we considered the percent of newly created transcripts that still had intronic reads in each timepoint. To do so, we used MISO[2] to calculate an intron-specific percent spliced in (PSI or $\Psi$) value, using reads within the intron body and junction reads spanning the 5' and 3' splice sites bordering the intron. We calculated a $\Psi$ value for each annotated intron in each replicate library for each time point using a custom MISO annotation built from FlyBase *Drosophila melanogaster* Release 5.57 annotations[41]. For all following analyses, we only considered the 29,809 introns that met the following conditions: (1) in a gene with TPM $\geq$ 2 in the total RNA libraries, (2) met minimum read coverage criteria for MISO estimation of $\Psi$ value in at least one

replicate for all time-points, (3) was completely or predominantly spliced out in the total RNA libraries ($\Psi < 0.2$), and (4) did not contain an annotated alternative splicing event such as a skipped exon. Introns flanking alternatively spliced exons were retained for further analysis.

**Estimating intron half-lives**

The labeling design used will result in the isolation of transcripts that initiated transcription at any time during the labeling period, of duration $t_{labeling}$, as well as transcripts that were elongated during this period but initiated prior to the labeling period. For a transcript to be labeled during the labeling period and include portions informative about the splicing of an intron, the polymerase must have either: (i) transcribed the 3'-most base of the intron during the interval of labeling – between time 0 and $t_{labeling}$ or (ii) have transcribed this base before time 0, but not terminated transcription prior to time 0. The time since synthesis of the intron in case (i) above will be distributed uniformly between 0 and $t_{labeling}$. And the time since synthesis of the intron in case (ii) will be distributed uniformly between $t_{labeling}$ and $t_{labeling} + \tau$, where $\tau$ is the time required to transcribe the portion of the transcript 3' of the intron. $\tau$ was estimated as the distance from the 3'-most base in the intron to the annotated transcription endpoint, in nt, divided by the typical *Drosophila* transcription rate of 1500 nt/min[8,42]. We therefore estimated the mean time since synthesis for each intron captured, $t_{inferred}$, as

$$t_{inferred} = \frac{t_{labeling} + \tau}{2}$$

where $t_{labeling}$ is defined by the experimental condition, and $\tau$ is calculated separately for each intron depending on its location in the transcript.

The approach outlined above rests on two main assumptions: 1) that all transcripts synthesized during the labeling period incorporate the 4sU label; and 2) that all transcripts containing label are equally likely to be captured. The first of these assumptions seem very plausible, while the second is undoubtedly an oversimplification. Additionally, our approach neglects any contribution from mRNA decay since mRNA half-lives are typically > 2 h in flies, much longer than typical splicing half-times estimated here[43].

We used the set of $\Psi$ values across inferred time points to estimate the rate at which each intron was excised by fitting a first-order exponential decay model:

$$\psi(t) = \psi_0 e^{-\lambda t_{inferred}}$$

where (1) t is the time since synthesis of the intron, (2) $\Psi(t)$ is the $\Psi$ value at time t (so that $\Psi(0) = \Psi_0 = 1$ by definition), and (3) $\lambda$ is the decay constant. This model was fit (using log-transformed $\Psi$ values for computational efficiency) and splicing half-lives were estimated as:

$$t_{\frac{1}{2}} = \frac{\ln(2)}{\lambda}$$

where $t_{1/2}$ represents the time at which half of the transcripts containing the intron have completed excision of the intron. To assess our power to measure the rates of intron loss over time and understand biases, we performed simulations across a range of splicing half-lives (described in Supplementary Methods). To estimate the error in our measurements, we propagated the uncertainly around our $\Psi$ estimates (using sampling-based confidence intervals from MISO) through to our estimates of splicing half-lives (described in Supplementary Methods).

**Estimates for introns with short half-lives and long introns**

The simulations described in the Supplementary Note indicated that the accuracy of our exponential fitting of half-lives was reduced, and had an upward bias, for introns with half-lives of ~3 min or less. Furthermore, splicing estimates are less accurate for very low $\Psi$ values substantially below 0.1. Therefore, to reduce bias and improve accuracy of estimation of short half-lives, used an alternative method to estimate splicing half-times for introns with a $\Psi < 0.1$ at the second time point. For this subset (representing 9% of all introns), we estimated the intron half-life using only the first time point (5 min labeling period). Our modeling based on MISO may over-estimate PSI values of long introns at short labeling periods because of the inclusion of reads deriving from transcripts where the polymerase is actively transcribing the intron at the end of the labeling period, resulting in over-estimation of half-lives for this class of introns.

**Calculating splice site scores**

We calculated the strength of splice sites using a maximum entropy model as implemented in maxEntScan[44], using 9 bp around the 5' splice site (-3:+6) and 23 bp around the 3' splice site (-20:+3). These models were optimized on mammalian splice site preferences, but seem to be reasonable also for *Drosophila* and have been used in gene prediction in fly genomes.

**Identifying enhancer regions**

We used *Drosophila* transcriptional enhancers defined by the STARR-seq enhancer testing assay in Arnold *et al.*[33]. Significant STARR-seq peaks were overlapped with annotated *Drosophila* introns to identify transcriptional enhancers located within introns.

**Identifying sites of recursive splicing**

We used three features of recursive sites found in our nascent sequencing data to identify recursive sites: (1) splice junction reads derived from putative recursive sites ("RachetJunctions"), (2) recursive-site spanning pairs, specifically read pairs that map to sites flanking putative recursive segments such that the fragment length can only be accounted for by the presence recursive intermediate ("Rachet Pair"), and (3) a sawtooth pattern in intronic read density ("RachetScan"). Details of the statistical methods and our computational pipeline for recursive site detection are described in the Supplementary Material. Conservation of recursive sites was estimated using per nucleotide phastCons scores[45] from a 15-way *Drosophila* alignment downloaded from UCSC Genome Browser.

**Splicing rates in recursively spliced introns**

We quantified splicing rates for each recursive segment independently by mapping the original alignments onto new gene models with the upstream exon, recursive segment and downstream segment contiguous and then running the pipeline described above on these transformed alignments. Specifically, the new gene models consist of a single intron with the same length as the recursive segment flanked by exons with the same lengths as the original up and downstream exons.

Reads were mapped onto these gene models such that whether any other recursive splicing event has yet been completed was irrelevant. These reads were split into four classes:

(1) *Reads entirely in the recursive segment or either exon.* Reads entirely inside any region were directly mapped onto their corresponding region in the new gene model.
(2) *Reads overlapping the upstream end of the recursive segment.* Reads can overlap the upstream end of the recursive segment either by being a junction read from the upstream exon to the beginning of the recursive segment or being an unspliced read overlapping the beginning of the recursive segment. In either case, a read was placed at the upstream exon-intron boundary in the new gene model.
(3) *Reads overlapping the downstream end of the recursive segment.* Reads overlapping

the downstream end of the recursive segment were placed on the intron-downstream exon boundary in the new gene model.

(4) *Exon-exon junction reads.* In order to be ambivalent to whether all other recursive splicing events have yet been completed all junction reads spanning from the upstream exon to a recursive splice site downstream of the recursive segment or the downstream exons were treated equivalently. All were added as splice junction reads between the upstream and downstream exons in the new gene model.

Any reads not fitting into one of these classes were not included in the new alignments. Notably, this includes reads overlapping either exon-intron junction in the original alignments, reads in recursive segments not being considered, and splice junction reads aligning with an upstream end not at the upstream exon.

**Estimating splicing accuracy**

We estimated the accuracy of splicing in *Drosophila* introns by identifying non-annotated junction reads with non-canonical splice site sequences within annotated introns. To do so, we first re-mapped the raw 4sU-seq reads with the STAR v2.5 software[46], with the mapping parameter `--outSAMattribute NH HI AS nM jM` to mark the intron motif category for each junction read in the final mapped file.

The `jM` attribute adds a `jM:B:c` SAM attribute to split reads arising from exon-exon junctions. All junction reads were first isolated and separated based on the value assigned to the `jM:B:c` tag. Junction reads spanning splice sites in the following categories were considered to be annotated or canonical: (1) any annotated splice site based on FlyBase *D. melanogaster* Release 5.57 gene structures [`jM:B:c,[20-26]`], (2) intron motifs containing “GT-AG” (or the reverse complement) [`jM:B:c,1` or `jM:B:c,2`], (3) intron motifs containing “GC-AG” (or the reverse complement) [`jM:B:c,3` or `jM:B:c,4`], and (4) intron motifs containing “AT-AC” (or the reverse complement) [`jM:B:c,5` or `jM:B:c,6`]. Junction reads with `jM:B:c,0` were considered to arise from non-canonical non-annotated splice sites. We calculated the frequency of inaccurate splice junctions for each intron as a ratio of the density of reads arising from non-canonical non-annotated splice sites to the density of all junction reads from the intron.

**Gene Ontology analyses**

All Gene Ontology analyses were performed using the DAVID v6.7 database and biological processes gene ontology categories. For gene ontology analyses of recursively spliced introns, all genes with introns greater than 10,000 kb were used as a background set. For gene ontology analyses of fast or slowly spliced genes, all genes for which we were able to calculate a splicing half-life for at least 3 introns were used as a background set.

**Data availability**

Sequencing data have been deposited at the Gene Expression Omnibus (GEO) database under accession GSE93763.

**Code availability**

Code is available upon request and will be made available in a public repository prior to publication.

**Acknowledgements**

We thank Marvin Jens and other members of the Burge lab, Brent Graveley and Phil Sharp for helpful comments on this manuscript.

**Funding**

This work was supported by a Jane Coffin Childs Postdoctoral Fellowship (A.A.P.), by NIH grant R01-GM085319 (C.B.B.), and by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences to K.A. (Z01 ES101987).

**Author contributions**

K.A. and C.B.B. conceived and designed the study. T.H. performed experiments. A.B. mapped the data and performed bioinformatics analyses. A.A.P. analyzed the data with help from J.P. A.A.P. and J.P. performed modeling analyses and wrote supplemental methods. A.A.P. and C.B.B. wrote the manuscript. All authors read and approved the final submission.

## Figure legends

**Figure 1. Splicing efficiency is tightly linked to intron length in Drosophila.**
**(A)** Nascent RNA coverage across second intron of *tango9*. Colors represent time points, with 5 min after 4sU pulsing (*darkest shade*), through 10 min, 20 min, and total RNA sample (*lightest shade*), where the inferred timepoints (accounting for the mean lifetime of the transcripts) are 2.9 min, 5.4 min, and 10.4 min. **(B)** Decrease in proportional intronic reads ($\Psi$) across time with exponential fit (*dotted curve*). Initial time is drawn at 0 minutes assuming that 100% of transcripts contain intron immediately after transcription. **(C)** Running median of local splicing half lives across distribution of intron lengths. Median is computed in sliding bins of 50 introns. **(D)** Splicing half lives across bins of 10 nt intron lengths. Background bars display number of *Drosophila* introns in each bin. **(E)** Splicing half lives in different categories of intron regulation, where constitutive introns (*blue*) are spliced out faster than either regulated retained introns (*orange*) or introns flanking alternative exons (*yellow*).

**Figure 2. Recursive splicing is common mechanism to efficiently splice very long introns. (A)** Nascent RNA coverage across the first intron of *Ten-m*, which is recursively spliced. Vertical lines indicate location of detected recursive sites and curved lines indicate split junction reads between splice sites. The distribution of junction reads detected at each recursive site per timepoint is depicted in the inset. **(B)** Number of recursive sites identified in this study, across sites previously identified, novel high-confidence sites, and novel candidate sites, with 5 sites that were previously identified but not detected in this study. **(C)** Distribution of intron lengths for all introns over 1kb (*grey*), with medium-confidence recursive introns in light tan and high-confidence recursive introns in dark tan. **(D)** Percentage of introns greater than 40 kb with at least one detected recursive site across various sub-samples of read coverage, where 100% indicates the percentage of recursive introns detected in the full dataset. Introns are subdivided into all introns detected (*yellow*), introns from lowly expressed genes (TPM ≤ 20; *light blue*), and introns from highly expressed genes (TPM > 20; *dark blue*). **(E)** Relative positions of recursive sites within introns for random sites chosen from a uniform distribution (*grey*) and single recursive sites in an intron (*dark tan*). Inset shows the mean and standard error of the fractional distances of the first recursive segment for introns with varying numbers of recursive sites. **(F)** Splicing half-lives for individual recursive segments (*dark tan*) and full non-recursive introns chosen to match recursive segment lengths (*grey*). (G) Splicing accuracy measured by

percentage of non-canonical unannotated reads for recursive introns (*gold*), non-recursive introns matched for intron length (*grey*), and all non-recursive introns (*light grey, dotted*).

**Figure 3. Splicing efficiency across introns within a gene. (A)** Relative importance for variables influencing variance in splicing half-lives, when using a multiple linear-regression to account for variance in half-lives for non-first introns between 60-70 nt. **(B)** Mean variance of splicing half-lives across introns within a gene relative to randomly sampled introns (chosen to match the distribution of lengths within actual genes). **(C)** Half-life distributions across introns within three representative genes, with the gene schematics below. **(D)** Splicing half-lives for first introns (*red*), relative to second, third, fourth, and fifth introns (*dark grey*) and non-first introns chosen to match the distribution of first intron length (*light grey*). **(E)** Median splicing half-lives of non-first introns (mean within a bin, *y-axis*) within a gene within bins of first intron lengths (*x-axis*), with standard errors across the mean. **(F)** Cumulative distributions of first-intron lengths (*x-axis*) for groups of genes classified by number of annotated introns (*colors*).

## Supplementary Figure Legends

**Supplementary Figure 1. Calculating rates of splicing in Drosophila. (A)** Proportion of intronic reads in a transcript (Ψ values, *y-axis*) for nascent RNA collected 5min, 10min, and 20min after 4sU labeling, with a time point labeled overnight representing steady-state or total RNA levels. The overall decreases in Ψ values over time indicate increased completed splicing over time. **(B)** The distribution of coefficients from an exponential fit to the Ψ values across time. **(C)** The distribution of $R^2$ values obtained from fitting an exponential model to Ψ values across time, with Ψ values estimated using MISO (*blue*) or only junction reads (*grey*). The preponderance of positive coefficients and high $R^2$ values indicates that an exponential decay model is appropriate. **(D)** Standard error estimates on half-life propagated from Ψ confidence intervals (see Methods) across a range of splicing half-life times. **(E)** A high concordance between the distribution of Ψ values calculated using only exon-exon and exon-intron spanning junction reads (x-axis) vs. the Ψ values calculated using the MISO software, all at 5 minutes after 4sU labeling. **(F)** Splicing half-lives estimated from simulated data (*y-axis*) relative to the simulated half-life (*x-axis*), with splicing half-life range from 2 to 1000 minutes and 0.1 to 20 minutes (*inset*).

**Supplementary Figure 2. Properties of splicing efficiency across varying intron lengths.** **(A)** Distribution of intron lengths in the *Drosophila melanogaster* genome. **(B)** Distribution of splice site strengths (MaxEnt score, *y-axis*) across both 3' splice sites (*orange*) and 5' splice sites (*blue*) for introns between 40-100 nt (*x-axis*). On average, 40-50 nt introns have have weaker splice site scores. **(C)** The distribution of splicing efficiency (half-lives, *y-axis*) for very short 40-50 nt introns (*dark blue*), relative to the distributions for 60-70 nt introns matching for the distributions of 40-50nt 3' splice site strength (*light blue*, t-test $P = 0.0001$), 5' splice site strength (*light blue*, t-test $P = 0.0033$), and both 5' and 3' splice site strengths (*light blue*, t-test $P = 0.0004$). 40-50 nt introns are consistently spliced out slower than other introns, independent of their weaker splice site strengths. **(D)** Distribution of splice site strengths (MaxEnt score, *y-axis*) across both 3' and 5' splice sites for introns binned into quantiles of intron length (*x-axis*). **(E)** The distribution of splicing half-lives (*y-axis*) for very long introns greater than 10 kb (dark blue), relative to the distributions of 60-70 nt introns matching for the distributions of 10 kb + 3' splice site strength (light blue, t-test $P = 7.093 \times 10^{-14}$), 5' splice site strength (light blue, $P < 10^{-16}$), and both 5' and 3' splice site strengths (light blue, $P < 10^{-16}$). Introns > 10 kb in length are

consistently spliced more slowly than other introns, independent of their stronger splice site strength.

**Supplementary Figure 3. Identifying sites of recursive splicing. (A)** Schematic indicating two computational approaches used to detect recursive sites: junction split and spanning reads (*top*) and automatic detection of sawtooth patterns (*bottom*). **(B)** Number of recursive sites identified by one of multiple identification pipelines, with the majority of recursive sites identified by both junction reads and sawtooth scores, as well as present in the Duff *et al.* dataset. **(C)** The gene expression levels of genes with recursive introns (TPM, *y-axis*) relative to the junction spanning read support for each recursive intron (read count, *x-axis*), showing the varying power to identify recursive sites with the sawtooth recursive method (*orange*), junction-spanning reads alone (*blue*), or both methods (*black*). **(D)** The probability derived from the sawtooth MCMC model of a site being a recursive site for the final set of recursive sites (*light orange*), all sites with minimal support from any method (*dark orange*), and random sites placed down in the same introns (*grey*). **(E)** The sawtooth score (see Methods) for the final set of recursive sites (*light orange*), all sites with minimal support from any method (*dark orange*), and random sites place down in the same introns (*grey*). **(F)** The cumulative distribution of distances between the recursive site identified with the sawtooth recursive method and the best matching recursive motif (orange) and random sites placed down in the same introns (grey) are significantly different.

**Supplementary Figure 4. Properties of recursively spliced introns. (A)** Sequence logo for all intronic AG|GT sites (*top*), medium-confidence recursive sites (*middle*) and high-confidence recursive sites (*bottom*). **(B)** Conservation of sequences around all detected recursive sites, with average phastCons scores for medium-confidence recursive sites (*yellow*), high-confidence recursive sites (*gold*), and random AG|GT sites in introns increasingly larger than 1kb (*grey*). **(C)** Full intron length distributions for introns (*y-axis*) with varying numbers of recursive sites (*x-axis*).

**Supplementary Figure 5. Rates of recursive splicing. (A)** Splicing half-lives (*y-axis*) for recursive segments with varying positions across the intron (*x-axis*), where on average, all segments in an intron tend to be spliced out at similar rates. **(B)** The number of splice junction reads (*y-axis*) spanning a 5' splice site and recursive site (*blue*), two recursive sites (*gold*), and a recursive site and 3' splice site (*yellow*) across the time-points (*x-axis*). **(C)** Distribution of

lengths of recursive segments (nucleotides, *x-axis*) for medium-confidence recursive segments (*yellow*) and high-confidence recursive segments (*gold*). **(D)** The distribution of splicing half-lives (*y-axis*) for the longest recursive segments in introns (*gold*) relative to non-recursive introns chosen to match the length of the recursive segments (*grey*). **(E)** The distribution of mean life-times (y-axis) for recursively spliced introns (estimated by the maximum of exponentials from constituent recursive segment splicing rates, *gold*) relative to non-reucursive introns chosen to match the length of the recursive introns (*grey*).

**Supplementary Figure 6. Variance in splicing half-lives across introns in a gene. (A)** The proportion of annotated introns detected for each gene (*x*-axis), with 76% of genes having sufficient coverage in 100% of annotated introns. **(B)** The cumulative distribution of variance in splicing half-lives (coefficient of variation, *x-axis*) across introns within a gene (*dark blue*) and introns randomly sampled to match the distribution of lengths of introns within actual genes (*dark grey*). This trend is consistent when excluding the first intron of each gene (*light blue*) and doing a similar sampling strategy excluding the length of the first introns (*light grey*). **(C)** Splicing half-lives (*y-axis*) before (*dark blue*) and after (*light blue*) correcting for the non-linear correlation between splicing half-lives and intron length. Length-correction was done by subtracting a running median of local splicing half-lives, where medians were computed in 50 intron sliding bins. **(D)** The cumulative distribution of variance in length-corrected relative splicing half-lives (coefficient of variation, *x-axis*), within categories of observed and sampled introns as in (B).

**Supplementary Figure 7. Correcting for the effects of intron length.** (**A)** The distribution of splicing efficiency (median half-lives, *y-axis*) vs. intron length (mean nucleotides, *x-axis*) for introns in different positions across a gene. First introns are longer and more slowly spliced than non-first introns. **(B)** Correcting for length does not account for the slower splicing of first introns, where splicing half-lives (*y-axis*) for first introns are still slower than for non-first introns in both the measured half-lives (*dark blue*) and the length-corrected relative half-lives (*light blue*). (**C**) The percentage of enhancers within an intron (*y-axis*) for first introns (*blue*) and non-first introns (*grey*) binned by quintiles of intron length (*x-axis*). (**D**) The distribution of intron half-lives (*y-axis*) for introns containing an enhancer (*blue*) and without an enhancer (*grey*) binned by quintiles of intron length (*x-axis*).

**Supplementary Figure 8. First-intron length and splicing efficiency.** Running median (*red*) of local gene-specific median splicing half-lives across the distribution of first intron lengths, for

raw splicing half-lives (**A**) and splicing half-lives corrected for intron length (**B**). Running median is computed in sliding bins of 50 genes. **(C)** Selected genes with four or five total introns, of which all of the non-first introns are 60-70 nt in length and have low variance across their splicing half-lives. Varying first intron lengths across these genes (nucleotides, *x-axis*) shows a correlation between first intron length and the median half-lives for these genes (*y*-axis).

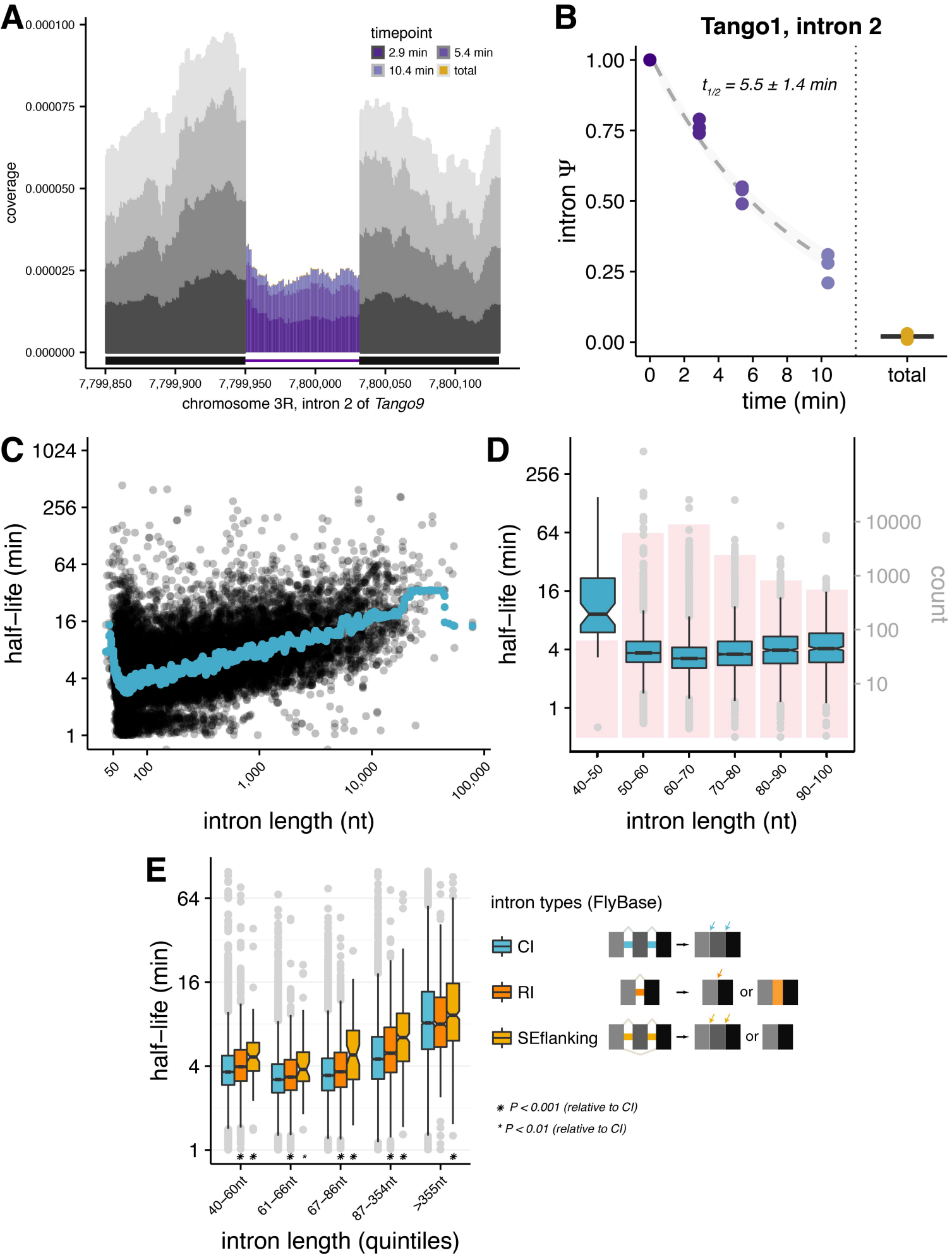
**Supplementary Figure 9.** Coefficients from a multiple linear-regression with several parameters (y-axis), where the coefficient represents the % change in half-life concordant with a 1% change in each parameter. Bars indicate the standard error and the size of the mean dot indicates the –log10 p-value for the significance of the individual parameter.

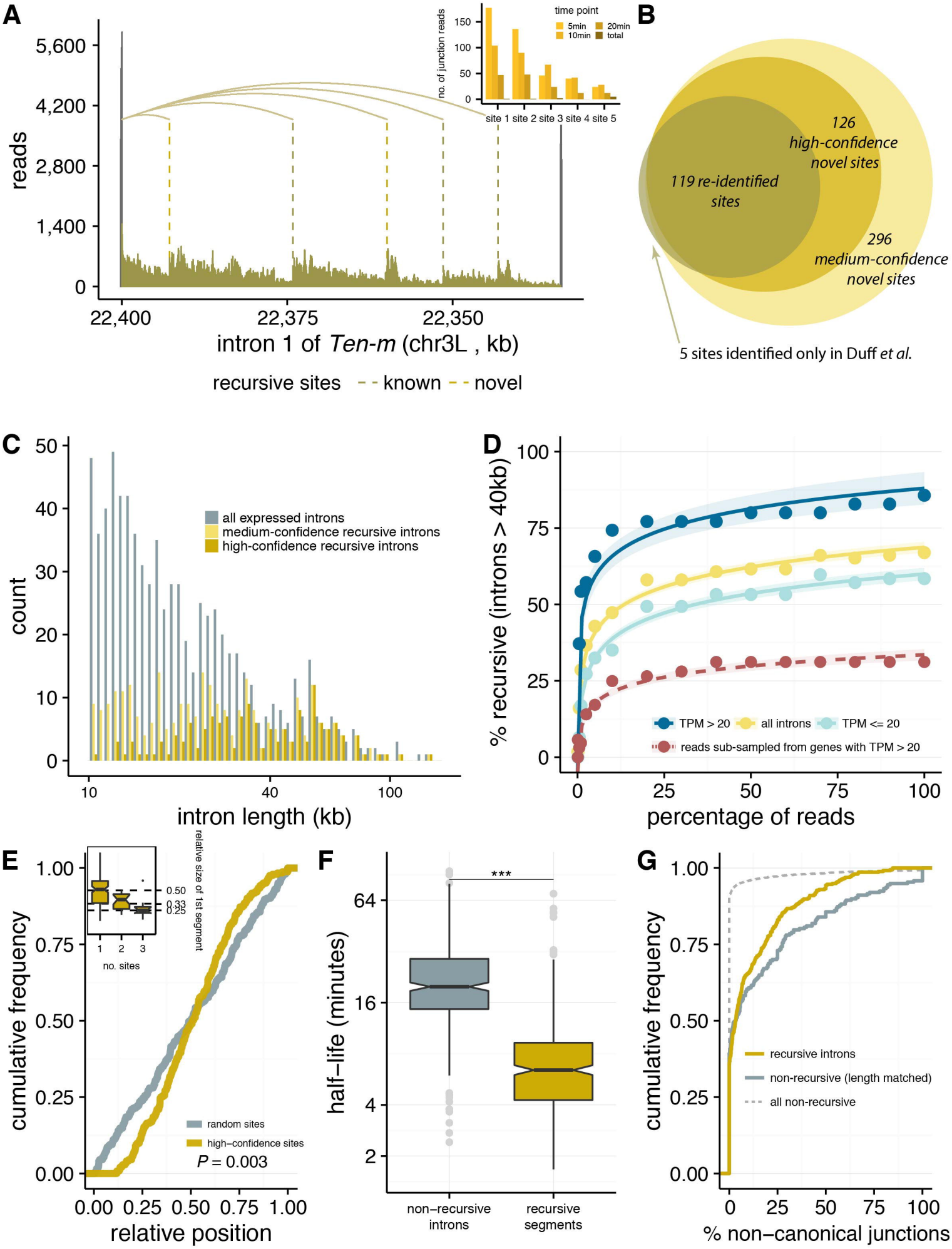
## References

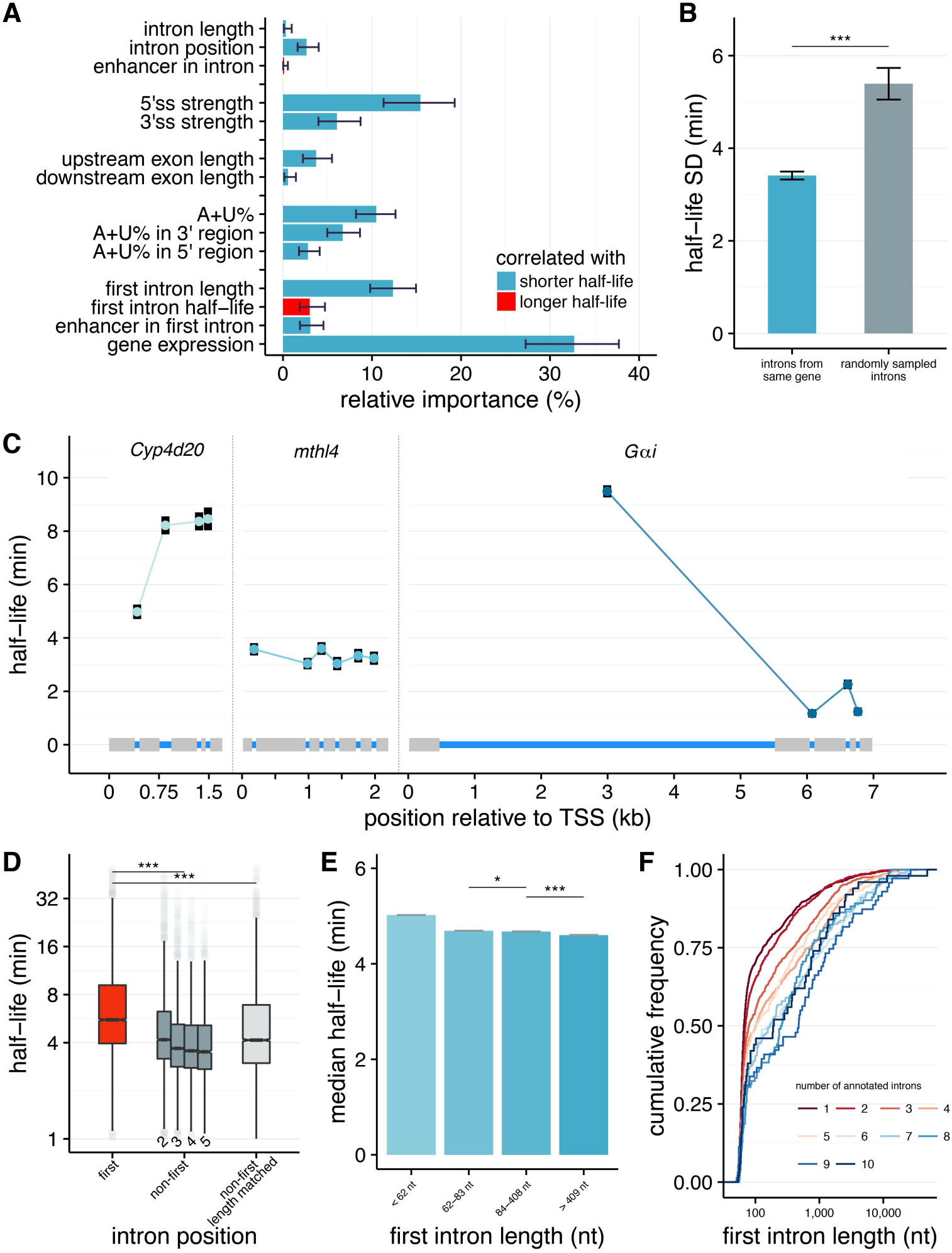

1. Windhager, L. *et al.* Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research* **22,** 2031–2042 (2012).
2. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Meth* **7,** 1009–1015 (2010).
3. Sharp, P. A. *et al.* Splicing of Messenger RNA Precursors. *Cold Spring Harb Symp Quant Biol* **52,** 277–285 (1987).
4. Coulon, A. *et al.* Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife Sciences* **3,** e03939 (2014).
5. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology* **16,** 1128–1133 (2009).
6. Martin, R. M., Rino, J., Carvalho, C., Kirchhausen, T. & Carmo-Fonseca, M. Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. *Cell Reports* **4,** 1144–1155 (2013).
7. Carrillo Oesterreich, F. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165,** 372–381 (2016).
8. Ardehali, M. B. & Lis, J. T. Tracking rates of transcription and splicing in vivo. *Nature Structural & Molecular Biology* **16,** 1123–1124 (2009).
9. Khodor, Y. L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development* **25,** 2502–2512 (2011).
10. Brugiolo, M., Herzel, L. & Neugebauer, K. M. Counting of co-transcriptional splicing. *F1000Prime Rep* (2013).
11. Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R. & Blencowe, B. J. Dynamic Integration of Splicing within Gene Regulatory Pathways. *Cell* **152,** 1252–1269 (2013).
12. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci USA* **98,** 11193–11198 (2001).
13. Carvalho, A. B. & Clark, A. G. Genetic recombination: Intron size and natural selection. *Nature* **401,** 344–344 (1999).
14. Parsch, J. Selective Constraints on Intron Evolution in Drosophila. *Genetics* **165,** 1843–1851 (2003).
15. Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M. & Andolfatto, P. 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,** 1226–1234 (2010).
16. Farlow, A., Dolezal, M., Hua, L. & Schlötterer, C. The Genomic Signature of Splicing-Coupled Selection Differs between Long and Short Introns. *Molecular Biology and Evolution* **29,** 21–24 (2012).
17. Wieringa, B., Hofer, E. & Weissmann, C. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell* **37,** 915–925 (1984).
18. Crawford, D. J., Hoskins, A. A., Friedman, L. J., Gelles, J. & Moore, M. J. Single-molecule colocalization FRET evidence that spliceosome activation precedes stable approach of 5′ splice site and branch site. *Proc Natl Acad Sci USA* **110,** 6783–6788 (2013).
19. Black, D. L. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes & Development* **5,** 389–402 (1991).
20. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *WIREs RNA* **4,** 49–60 (2013).
21. Berget, S. M. Exon Recognition in Vertebrate Splicing. *J Biol Chem* **270,** 2411–2414

(1995).
22. la Mata, de, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Molecular Cell* **12,** 525–532 (2003).
23. Jonkers, I., Kwak, H., Lis, J. T. & Struhl, K. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife Sciences* **3,** e02407 (2014).
24. Mayer, A. *et al.* Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161,** 541–554 (2015).
25. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
26. Hatton, A. R., Subramaniam, V. & Lopez, A. J. Generation of Alternative Ultrabithorax Isoforms and Stepwise Removal of a Large Intron by Resplicing at Exon–Exon Junctions. *Molecular Cell* **2,** 787–796 (1998).
27. Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J. & Lopez, A. J. Subdivision of Large Introns in Drosophila by Recursive Splicing at Nonexonic Elements. *Genetics* **170,** 661–674 (2005).
28. Duff, M. O. *et al.* Genome-wide identification of zero nucleotide recursive splicing in Drosophila. *Nature* **521,** 376–379 (2015).
29. Hicks, M. J., Mueller, W. F., Shepard, P. J. & Hertel, K. J. Competing Upstream 5′ Splice Sites Enhance the Rate of Proximal Splicing. *Molecular and Cellular Biology* **30,** 1878–1886 (2010).
30. Goodall, G. J. & Filipowicz, W. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* **58,** 473–483 (1989).
31. Lorković, Z. J., Wieczorek Kirk, D. A., Lambermon, M. H. L. & Filipowicz, W. Pre-mRNA splicing in higher plants. *Trends in Plant Science* **5,** 160–167 (2000).
32. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471,** 480–485 (2011).
33. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339,** 1074–1077 (2013).
34. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **16,** 167–177 (2015).
35. Henriques, T. *et al.* Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Molecular Cell* **52,** 517–528 (2013).
36. Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R. & Boothroyd, J. C. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature Biotechnology* **23,** 232–237 (2005).
37. Ben Langmead, Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).
38. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
39. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
40. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* (2016). doi:10.1038/nbt.3519
41. St Pierre, S. E., Ponting, L., Stefancsik, R., McQuilton, P.FlyBase Consortium. FlyBase 102—advanced approaches to interrogating FlyBase. *Nucleic Acids Research* **42,** D780–D788 (2014).
42. Garcia, H. G., Tikhonov, M., Lin, A. & Gregor, T. Quantitative Imaging of Transcription in Living Drosophila Embryos Links Polymerase Activity to Patterning. *Current Biology* **23,** 2140–2145 (2013).

43. Herold, A., Teixeira, L. & Izaurralde, E. Genome-wide analysis of nuclear mRNA export pathways in Drosophila. *The EMBO Journal* **22,** 2472–2483 (2003).
44. Yeo, G. & Burge, C. B. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *http://www.liebertpub.com/cmb* (2004). doi:10.1089/1066527041410418
45. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15,** 1034–1050 (2005).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).


A
0.000100
0.000075
0.000050
0.000025
0.000000
coverage
timepoint
2.9 min
5.4 min
10.4 min
total
7,799,850
7,799,900
7,799,950
7,800,000
7,800,050
7,800,100
chromosome 3R, intron 2 of Tango9
B
Tango1, intron 2
$t_{1/2}$ = 5.5 ± 1.4 min
1.00
0.75
0.50
0.25
0.00
intron Ψ
0
2
4
6
8
10
total
time (min)
C
1024
256
64
16
4
1
half−life (min)
50
100
1,000
10,000
100,000
intron length (nt)
D
256
64
16
4
1
half−life (min)
10000
1000
100
10
count
40–50
50–60
60–70
70–80
80–90
90–100
intron length (nt)
E
64
16
4
1
half−life (min)
40–60nt
61–66nt
67–86nt
87–354nt
>355nt
intron length (quintiles)
intron types (FlyBase)
CI
RI
SEflanking
or
or
* P < 0.001 (relative to CI)
* P < 0.01 (relative to CI)


A
reads
5,600
4,200
2,800
1,400
0
22,400
22,375
22,350
intron 1 of Ten-m (chr3L , kb)
recursive sites
known
novel
no. of junction reads
time point
5min
10min
20min
total
site 1 site 2 site 3 site 4 site 5
B
126
high-confidence
novel sites
119 re-identified
sites
296
medium-confidence
novel sites
5 sites identified only in Duff et al.
C
count
all expressed introns
medium-confidence recursive introns
high-confidence recursive introns
intron length (kb)
D
% recursive (introns > 40kb)
TPM > 20
all introns
TPM <= 20
reads sub-sampled from genes with TPM > 20
percentage of reads
E
cumulative frequency
relative size of 1st segment
no. sites
random sites
high–confidence sites
P = 0.003
relative position
F
half–life (minutes)
***
non–recursive
introns
recursive
segments
G
cumulative frequency
recursive introns
non-recursive (length matched)
all non-recursive
% non–canonical junctions


A
intron length
intron position
enhancer in intron
5'ss strength
3'ss strength
upstream exon length
downstream exon length
A+U%
A+U% in 3' region
A+U% in 5' region
first intron length
first intron half–life
enhancer in first intron
gene expression
correlated with
shorter half-life
longer half-life
0
10
20
30
40
relative importance (%)
B
***
half–life SD (min)
0
2
4
6
introns from same gene
randomly sampled introns
C
Cyp4d20
mthl4
Gαi
half–life (min)
0
2
4
6
8
10
0
0.75
1.5
0
1
2
0
1
2
3
4
5
6
7
position relative to TSS (kb)
D
***
***
half–life (min)
1
4
8
16
32
2
3
4
5
first
non-first
non-first length matched
intron position
E
*
***
median half–life (min)
0
2
4
6
<62 nt
62–83 nt
84–408 nt
>409 nt
first intron length (nt)
F
cumulative frequency
0.00
0.25
0.50
0.75
1.00
number of annotated introns
1
2
3
4
5
6
7
8
9
10
100
1,000
10,000
first intron length (nt)