

UC Irvine

UC Irvine Previously Published Works

Title

The architecture of pre-mRNAs affects mechanisms of splice-site pairing

Permalink

<https://escholarship.org/uc/item/3b12w5ff>

Journal

PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 102(45)

ISSN

0027-8424

Authors

Fox-Walsh, KL
Dou, YM
Lam, BJ
et al.

Publication Date

2005

DOI

10.1073/pnas.0508489102

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

The architecture of pre-mRNAs affects mechanisms of splice-site pairing

Kristi L. Fox-Walsh*, Yimeng Dou†, Bianca J. Lam*, She-pin Hung*, Pierre F. Baldi†, and Klemens J. Hertel**

*Departments of Microbiology and Molecular Genetics and †Information and Computer Sciences, University of California, Irvine, CA 92697-4025

Communicated by Thomas Maniatis, Harvard University, Cambridge, MA, September 28, 2005 (received for review March 30, 2005)

The exon/intron architecture of genes determines whether components of the spliceosome recognize splice sites across the intron or across the exon. Using *in vitro* splicing assays, we demonstrate that splice-site recognition across introns ceases when intron size is between 200 and 250 nucleotides. Beyond this threshold, splice sites are recognized across the exon. Splice-site recognition across the intron is significantly more efficient than splice-site recognition across the exon, resulting in enhanced inclusion of exons with weak splice sites. Thus, intron size can profoundly influence the likelihood that an exon is constitutively or alternatively spliced. An EST-based alternative-splicing database was used to determine whether the exon/intron architecture influences the probability of alternative splicing in the *Drosophila* and human genomes. *Drosophila* exons flanked by long introns display an up to 90-fold-higher probability of being alternatively spliced compared with exons flanked by two short introns, demonstrating that the exon/intron architecture in *Drosophila* is a major determinant in governing the frequency of alternative splicing. Exon skipping is also more likely to occur when exons are flanked by long introns in the human genome. Interestingly, experimental and computational analyses show that the length of the upstream intron is more influential in inducing alternative splicing than is the length of the downstream intron. We conclude that the size and location of the flanking introns control the mechanism of splice-site recognition and influence the frequency and the type of alternative splicing that a pre-mRNA transcript undergoes.

alternative splicing | bioinformatics | EST database | intron length

Pre-mRNA splicing is an essential process that accounts for many aspects of regulated gene expression. Of the $\approx 25,000$ genes encoded by the human genome (1), $>60\%$ are believed to produce transcripts that are alternatively spliced. Thus, alternative splicing of pre-mRNAs can lead to the production of multiple protein isoforms from a single pre-mRNA, exponentially enriching the proteomic diversity of higher eukaryotic organisms (2, 3). Because regulation of this process can determine when and where a particular protein isoform is produced, changes in alternative-splicing patterns modulate many cellular activities.

The spliceosome assembles onto the pre-mRNA in a coordinated manner by binding to sequences located at the 5' and 3' ends of introns. Spliceosome assembly is initiated by the stable associations of the U1 small nuclear ribonucleoprotein particle with the 5' splice site, branch-point-binding protein/SF1 with the branch point, and U2 snRNP auxiliary factor with the pyrimidine tract (4). ATP hydrolysis then leads to the stable association of U2 snRNP at the branch-point and functional splice-site pairing (5).

Intron size has been correlated with rates of evolution (6) and the regulation of genome size (7, 8). The exon/intron architecture has also been shown to influence splice-site recognition (9–11). For example, increasing the size of mammalian exons results in exon skipping. However, the same enlarged exons were included when the flanking introns were small (11). Thus, splice-site recognition is more efficient when introns or exons are small. Because, in the human genome, the majority of exons are

short and introns are long (12), it is expected that the vast majority of splice sites in the human genome are recognized across the exon. Lower eukaryotes have a genomic architecture that is typified by small introns and flanking exons with variable length, suggesting that splice-site recognition occurs across the intron (10, 13, 14). Consistent with this model, expansion of small introns in yeast or *Drosophila* causes loss of splicing, cryptic splicing, or intron retention (9, 15). Taken together, these observations suggest that splice sites are recognized across an optimal nucleotide length.

It is unknown whether splice-site recognition across the intron or across the exon results in similar efficiencies of spliceosomal assembly and/or splice-site pairing. Here, we demonstrate that splice-site recognition across the intron ceases when the intron reaches a length between 200 and 250 nt. Because splice-site recognition is more efficient across the intron, alternative splicing is less likely for exons flanked by short introns. This influence is supported experimentally and by computational analyses of *Drosophila* and human alternative-splicing databases. We conclude that the size and location of the flanking introns control the mechanism of splice-site recognition and influence the frequency and the type of alternative pre-mRNA splicing.

Methods

RNA and Splicing Reactions. A detailed description of the construction of the two-exon (A–D) and the three-exon (L/L–S/S) pre-mRNA substrates is summarized in *Supporting Text*, which is published as supporting information on the PNAS web site. All plasmids were linearized with XhoI, *in vitro* transcribed with SP6 RNA polymerase (Promega), uniformly labeled with ^{32}P , and gel purified on 7 M urea polyacrylamide gel. *In vitro* splicing reactions were performed in 30% HeLa nuclear extract as described in ref. 16. Bands were visualized and quantitated by using PhosphorImager analysis and QUANTITY ONE software (Bio-Rad). Percent spliced is defined as spliced products/(unspliced RNA + spliced products). To derive kinetic rate constants, time points were fit to a first-order rate description for product appearance. Transfection experiments with Lipofectamine (Life Technologies) were performed in HeLa cells grown in MEM supplemented with 2 mM glutamine and 10% FBS according to manufacturer protocols. Each splicing experiment was repeated at least three times.

Computational Analysis. The computational analysis was based on the Alternative Splicing Database (ASD) (17). ASD is a computer-generated data set of transcript-confirmed splice patterns, alternative-splice events, and the associated annotations. ASD is downloadable and provides statistics and coverage similar to other databases analyzing alternative splicing (18, 19). We used a large EST database to determine the frequency of alternative splicing, because it included alternative-splicing information for

Conflict of interest statement: No conflicts declared.

Abbreviations: ESEs, exonic splicing enhancers; ASD, Alternative Splicing Database.

†To whom correspondence should be addressed. E-mail: khertel@uci.edu.

© 2005 by The National Academy of Sciences of the USA

Table 1. Rate constants and fold activation for test substrates at various intron lengths

| Intron length | Substrate | | | |
|-----------------|-----------|---------|-----------|---------|
| | A | B | C | D |
| 120-nt | 12 ± 7 | 42 ± 15 | 44 ± 23 | 65 ± 13 |
| Fold activation | 1 | 3.5 | 3.7 | 5.4 |
| Normalized | 18.5 | 65 | 68 | 100 |
| 200-nt | 1.4 ± 0.9 | 45 ± 3 | 62 ± 11 | 94 ± 19 |
| Fold activation | 1 | 32 | 44 | 67 |
| Normalized | 1.5 | 48 | 65 | 100 |
| 250-nt | 0.9 ± 0.4 | 6.4 ± 4 | 12 ± 2 | 45 ± 6 |
| Fold activation | 1 | 7 | 13 | 50 |
| Normalized | 2 | 14 | 26 | 100 |
| 340-nt | 0.5 ± 0.1 | 14 ± 3 | 12 ± 6 | 60 ± 17 |
| Fold activation | 1 | 28 | 24 | 120 |
| Normalized | 0.8 | 23 | 20 | 100 |
| 425-nt | 0.3 ± 0.1 | 5 ± 1 | 1.9 ± 0.2 | 50 ± 8 |
| Fold activation | 1 | 17 | 6.3 | 165 |
| Normalized | 0.6 | 10 | 4 | 100 |

All rate constants were determined from time points taken over a 3-h splicing reaction and are expressed in units of (10^{-2}-hr^{-1}). Fold activations were normalized to construct D. Experimental error for each rate determination was within 20%. Rates determined from different experiments varied <50%.

of both splice sites are recruited to the intron simultaneously. However, constructs with introns >200 nt demonstrated synergistic kinetics. We conclude that the change from splice-site recognition across the intron to splice-site recognition across the exon occurs when the intronic length is between 200 and 250 nt.

Mechanisms of Splice-Site Recognition and Alternative Splicing. The kinetic analysis summarized in Fig. 1 demonstrates that the upstream 5' splice site and the downstream 3' splice site are recognized simultaneously across introns <200 nt. Significantly, in the absence of ESEs, splice-site recognition across the intron is a much more efficient process than splice-site recognition across the exon (Fig. 2). Thus, splice-site recognition across the intron may be able to rescue the inclusion of internal exons harboring weak splice sites. To test this hypothesis, we designed a series of pre-mRNA substrates containing three exons for *in vitro* splicing analysis in which the internal exon contains splice sites that are insufficiently recognized in the absence of ESEs. The four substrates generated differed only in their ability to be recognized across each intron by changing the length of the intron from <200 to >250 nt (Fig. 3A), thus permitting or discouraging splice-site recognition across the intron. As expected, the internal exon is predominantly excluded when flanked by two long introns (Fig. 3B, substrate L/L, lane 2). However, we observed significant inclusion of the internal exon if one of the flanking introns is short enough to support splice-site recognition across the intron (Fig. 3B, substrates S/L, S/S, and L/S, lanes 4, 6, and 8). In fact, two short introns increase exon inclusion ≈ 30 times greater than two long introns (Fig. 3B, substrate S/S, lane 6).

To further analyze the effect of intron size on alternative splicing, we examined splicing of the same pre-mRNA substrates in cell culture. Qualitatively, results similar to those seen in Fig. 3B were observed in transfection experiments (Fig. 3C). When the weak internal exon is flanked by long introns, it is predominantly excluded from the splicing pattern (Fig. 3C, lane 1). However, the presence of small neighboring introns rescued internal exon inclusion (Fig. 3C, lanes 2–4). We conclude that splice-site recognition across the intron can

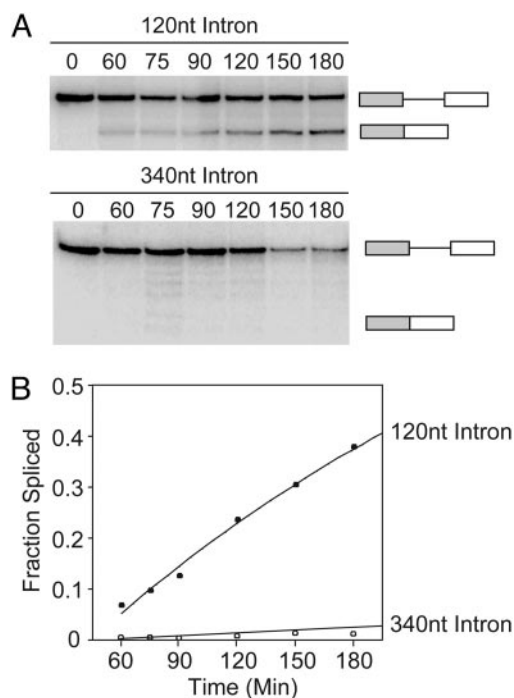


Fig. 2. Splice-site recognition across the intron is more efficient than splice-site recognition across the exon. (A) Representative autoradiogram depicting the efficiency of splicing for substrates containing weak 5' and 3' splice sites that differ only in the length of their intron, as indicated above each panel. The diagrams (Right) represent pre-mRNA and spliced products. (B) Quantitation of the data in A.

promote inclusion of exons flanked by weakly defined splice sites *in vitro* and *in vivo*.

Exon/Intron Architecture and Alternative Splicing in the *Drosophila* and Human Genomes. To estimate the fractions of splice sites that may be recognized through cross-intron interactions, we recorded the flanking-intron lengths for every internal exon within the human and *Drosophila* genomes. Genome information was obtained from the ASD, which contains information about the exon/intron structure and EST-verified alternative-splicing events of several thousand genes (17). Within the human genome, many exons are flanked by at least one short intron, creating two separate populations, separated roughly by the intron length that we propose to represent the transition of splice-site recognition from across the intron to across the exon (Fig. 4A). As expected from previous intron-length analyses (7, 24), a very different distribution is seen in the *Drosophila* genome, where $\approx 85\%$ of exons are flanked by at least one short intron (Fig. 4B). An overlay of the *Drosophila* and human genomes demonstrates that the minimum intron length in the human genome is at the same location that demarcates the maximum intron length of the major *Drosophila* exon population (see Fig. 6, which is published as supporting information on the PNAS web site). This difference in genome constraint may reflect specific compositional variations between the *Drosophila* and human spliceosomes.

Because splice-site recognition across the intron rescued exon inclusion (Fig. 3), we investigated whether intron length influences alternative splicing within the *Drosophila* and human genomes. To do so, the flanking-intron information of each exon was correlated with exon-skipping and alternative-splice-site-activation events reported in the ASD to compute the probability that an exon is involved in alternative splicing, without taking into consideration the contributions made by

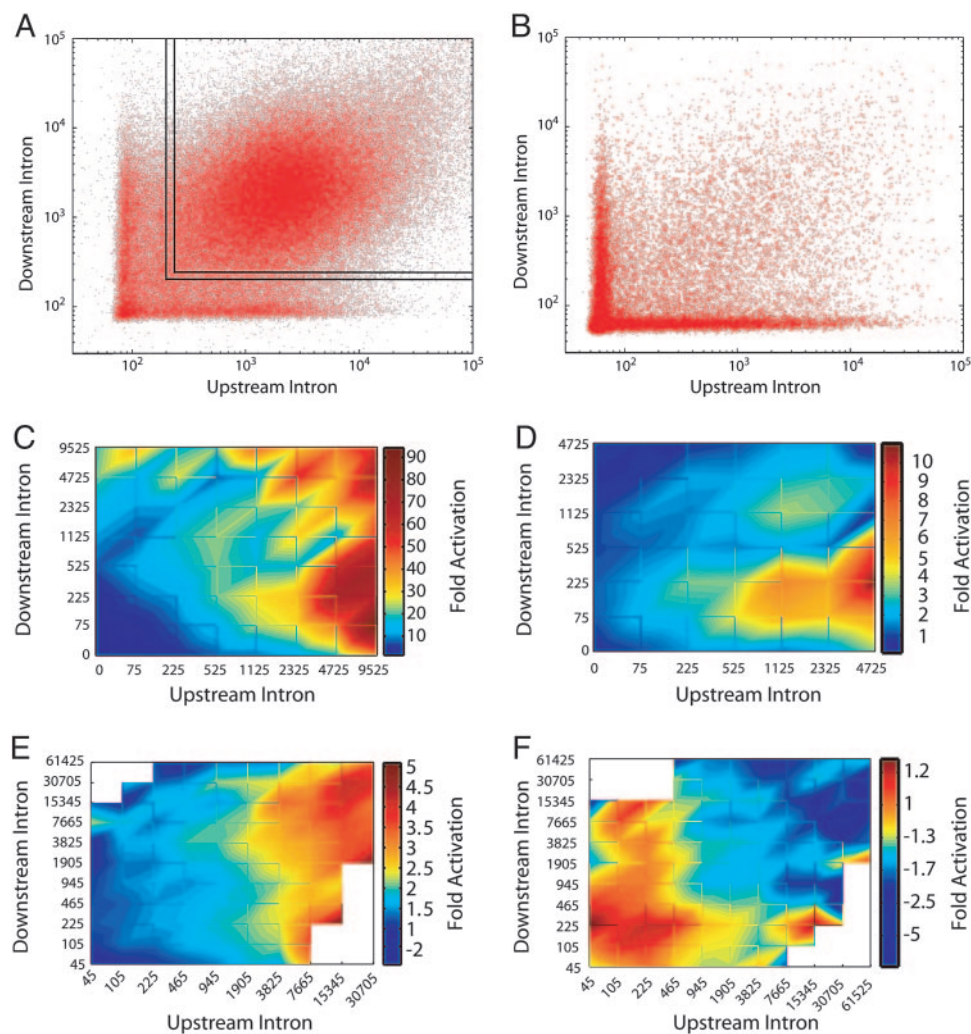


Fig. 4. Computational analysis of intron size and alternative splicing in the *Drosophila* and human genomes. (A and B) Every internal exon within the human or the *Drosophila* genomes is displayed as a function of the nucleotide length of its upstream (x axis) and downstream (y axis) introns. Each point within the scatter plots represents a unique exon. The x and y axes are shown in log scale. (A) Exon profile of the human genome. The majority of introns are long; however, $\approx 25\%$ of exons are flanked by at least one short intron. The vertical and horizontal lines demarcate the experimentally determined 200- to 250-nt transition from cross-intron to cross-exon recognition. (B) Exon profile of the *Drosophila* genome. The majority ($>85\%$) of exons are flanked by at least one short intron. (C–F) Color diagrams displaying the probability of an exon's undergoing alternative splicing as a function of the length of its flanking introns. The upstream and downstream intron length increases exponentially along the x and y axes. The color scale (Right) represents the fold increase in the probability of an exon's being alternatively spliced relative to the probability calculated for exons that are flanked by introns <225 nt. (C) Probability diagram for exon skipping within the *Drosophila* genome. (D) *Drosophila* alternative 5' or 3' splice-site usage. (E) Human exon skipping. (F) Human alternative 5' or 3' splice-site usage.

bp. Importantly, exon length is tightly distributed when compared with intron length (12). These results demonstrate that maintaining exon size in the human genome is more important to the architecture and evolution of a gene than is maintaining intron size. In contrast to the human genome, exon size varies much more than intron size in yeast (7). The maximum intron length of 182 nt lies well within the size limitations of splice-site recognition across the intron. Taken together, these considerations support the notion that the majority of splice sites in higher eukaryotes are recognized across the exon, whereas lower eukaryotes employ splice-site recognition across the intron (10).

It is well established that several types of exon and intron elements influence splice-site choice. The most prominent include the exon/intron junction signals and splicing enhancers and silencers (4). Our results show that the exon/intron architecture is an additional parameter that affects the effi-

ciency of splice-site recognition and alternative pre-mRNA splicing. When compared in otherwise isogenic test substrates, splice-site recognition across the intron could rescue the inclusion of a weak internal exon by >10 -fold (Fig. 3). Even though our computational analysis ignored the contributions made by variable splice sites, enhancers, and silencers, a striking increase in the probability of alternative splicing was observed for *Drosophila* exons, whose splice sites are recognized across the exon. Thus, the exon/intron architecture in *Drosophila* is a major determinant in governing the probability of alternative splicing. Within the human genome, we observed a qualitatively similar trend for exon-skipping events but with a reduced magnitude (Fig. 4). One major difference between the *Drosophila* and human gene architecture is intron length. Human genes are dominated by long introns (87% of introns are >250 nt), whereas short introns are much more common in *Drosophila* (66% are <250 nt). One possible explanation for

