


RESEARCH PAPER

 OPEN ACCESS



## Splice site proximity influences alternative exon definition

Francisco Carranza\*, Hossein Shenasa\*, and Klemens J. Hertel 

Department of Microbiology and Molecular Genetics, University of California Irvine, Irvine, California, USA

### ABSTRACT

Alternative splicing enables higher eukaryotes to expand mRNA diversity from a finite number of genes through highly combinatorial splice site selection mechanisms that are influenced by the sequence of competing splice sites, cis-regulatory elements binding trans-acting factors, the length of exons and introns harbouring alternative splice sites and RNA secondary structures at putative splice junctions. To test the hypothesis that the intron definition or exon definition modes of splice site recognition direct the selection of alternative splice patterns, we created a database of alternative splice site usage (ALTssDB). When alternative splice sites are embedded within short introns (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair, consistent with previous observations. However, when alternative splice sites are embedded within large flanking introns (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Thus, alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition. The results demonstrate that the spliceosome pairs splice sites that are closest in proximity within the unit of initial splice site selection.

### ARTICLE HISTORY

Received 10 February 2022  
Revised 07 May 2022  
Accepted 09 June 2022

### KEYWORDS

Alternative splicing; exon definition; Intron-exon architecture; splice site selection; bioinformatics; molecular biology

### Introduction



Pre-mRNA splicing is an essential step in eukaryotic gene expression that involves the excision of intronic sequences and the transesterification of exonic sequences by the spliceosome to generate protein coding mRNAs. Alternative exon inclusion is possible through a process known as alternative splicing. At least 95% of human genes undergo alternative splicing in response to cell cycle, developmental, tissue-specific or signalling cues. Alternative splicing increases proteomic diversity from a limited genome in a regulated fashion [1]. Thus, pre-mRNA splicing impacts gene expression [2].

The recognition of splice junctions by the spliceosome initiates the splicing reaction. The 5' splice site (5'ss) is defined by a nine-nucleotide consensus sequence that spans the exon/intron junction at the 5' end of each intron. The 3' splice site (3'ss) includes three sequence elements found within an approximately 40 nucleotides (nts) stretch, upstream of the 3' intron/exon junction. These include the intron/exon junction sequence, which contains the essential AG dinucleotide at the 3' end of the intronic sequence, the polypyrimidine tract (PPT), a region containing 15–20 pyrimidines located upstream of the intron/exon junction and the branch point sequence, a highly degenerate sequence that contains a conserved adenosine located upstream of the PPT.


Exon recognition is a highly combinatorial process that is known to be influenced by many cis- and trans-acting features. These include splicing enhancers, silencers, RNA secondary structure, the intron-exon architecture and the

sequence context of splice junctions [3–5]. The strength of splice sites is determined by how well they conform to consensus splice junction motifs that function in recruiting U1 snRNP to the 5'ss and U2AF to the 3'ss. Consensus similarity scores, derived from the modelling of short sequence motifs using the maximum-entropy principle (MaxEnt), define splice site strength numerically [6]. Splice sites are known to act synergistically and combined 5' and 3'ss scores are a much better predictor for exon inclusion than either splice site score alone [7]. Importantly, the ability of an exon to undergo various forms of alternative splicing is heavily influenced by the strength of its splice sites [8].

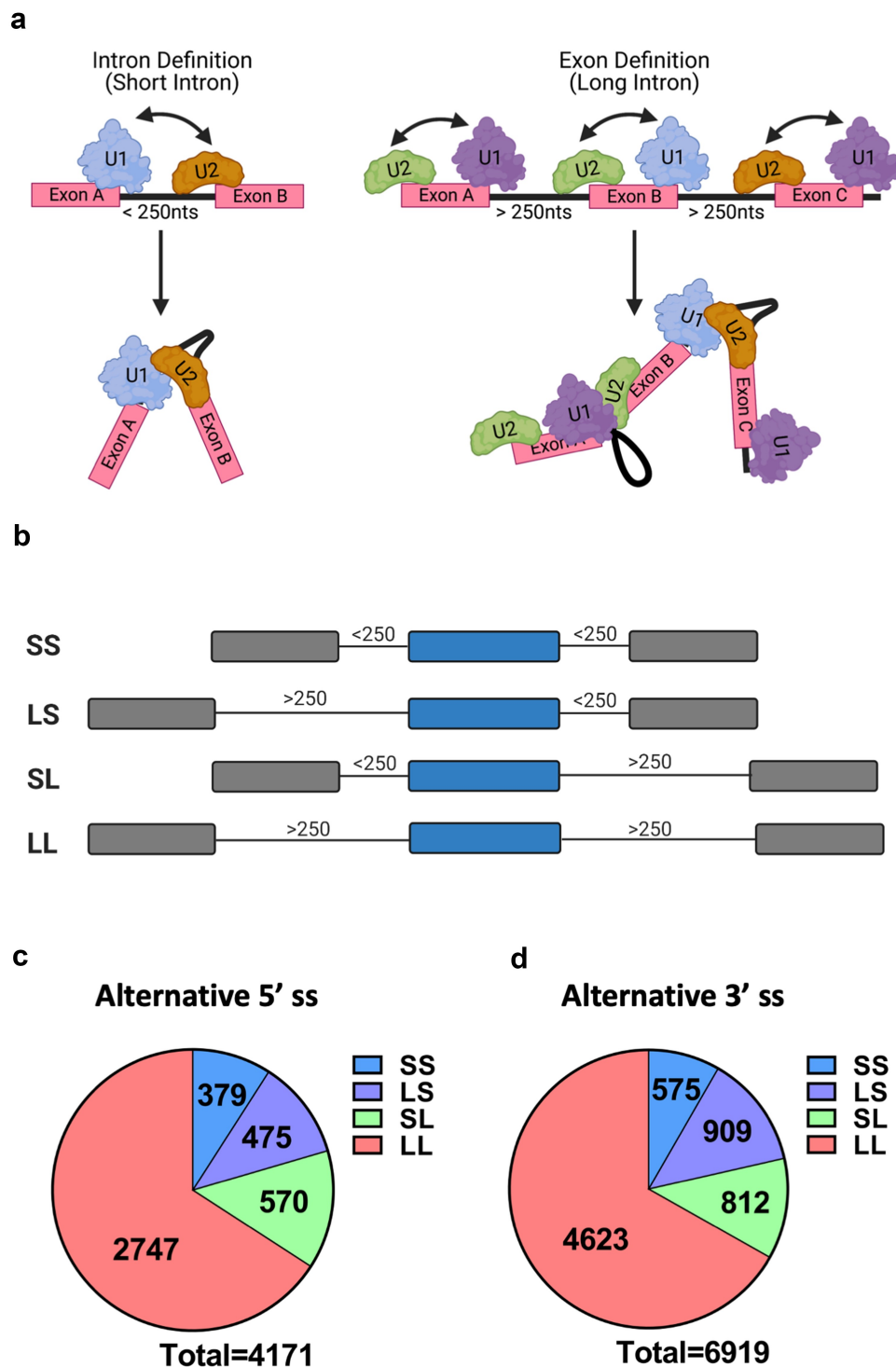
Another crucial factor in splice site selection is the genomic architecture [9–12]. The genomes in lower eukaryotes are characterized almost exclusively by the presence of short introns (<250 nts). By contrast, human genes harbour long introns, with >87% of introns longer than 250 nts [10]. This different genomic architecture has been shown to contribute significantly to the manner in which spliceosomal assembly occurs. The two proposed mechanisms through which splice sites are recognized are referred to as the exon or intron definition mode of splice site recognition (Figure 1(a)). During intron definition, the spliceosome assembles across the intron that will be excised. Under conditions that promote exon definition, initial splice site recognition is postulated to occur across the exon. This initial recognition is predicted to be followed by an additional splice site juxtapositioning step to induce intron excision. *In vitro* splicing

**CONTACT** Klemens J. Hertel  [khertel@uci.edu](mailto:khertel@uci.edu)  Department of Microbiology and Molecular Genetics, University of California Irvine, Irvine, California 92697, USA

\*These authors contributed equally.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15476286.2022.2089478>

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** Gene architecture and database (a) The two proposed modes of splice site recognition. During intron definition splice sites are recognized across the intron (left). Under exon definition (right) splice sites are initially recognized across the exon, followed by splice site juxtaposition. (b) ALTssDB categories of internal exons as defined by flanking intron size. S stands for short (less than 250 nts), L stands for long (greater than 250 nts). (c) and (d) Distribution of ALTssDB internal exon categories for alternative 5' (c) and alternative 3' (d) splice site events.

and transfection experiments of designer minigenes demonstrated that the transition between intron and exon definition occurs at an intron length of approximately 250 nts [10]. Thus, splice sites that are flanked by large introns (>250 nts) are recognized through exon definition, while intron-defined splice sites are associated with small flanking

introns (<250 nts). It is currently unknown how exon and intron definition influence alternative splice site selection.

Understanding the relationship between the splice site strength and intron-exon architecture splicing determinants has been a longstanding goal in deciphering the splicing code. The mechanisms utilized by the spliceosome to select the

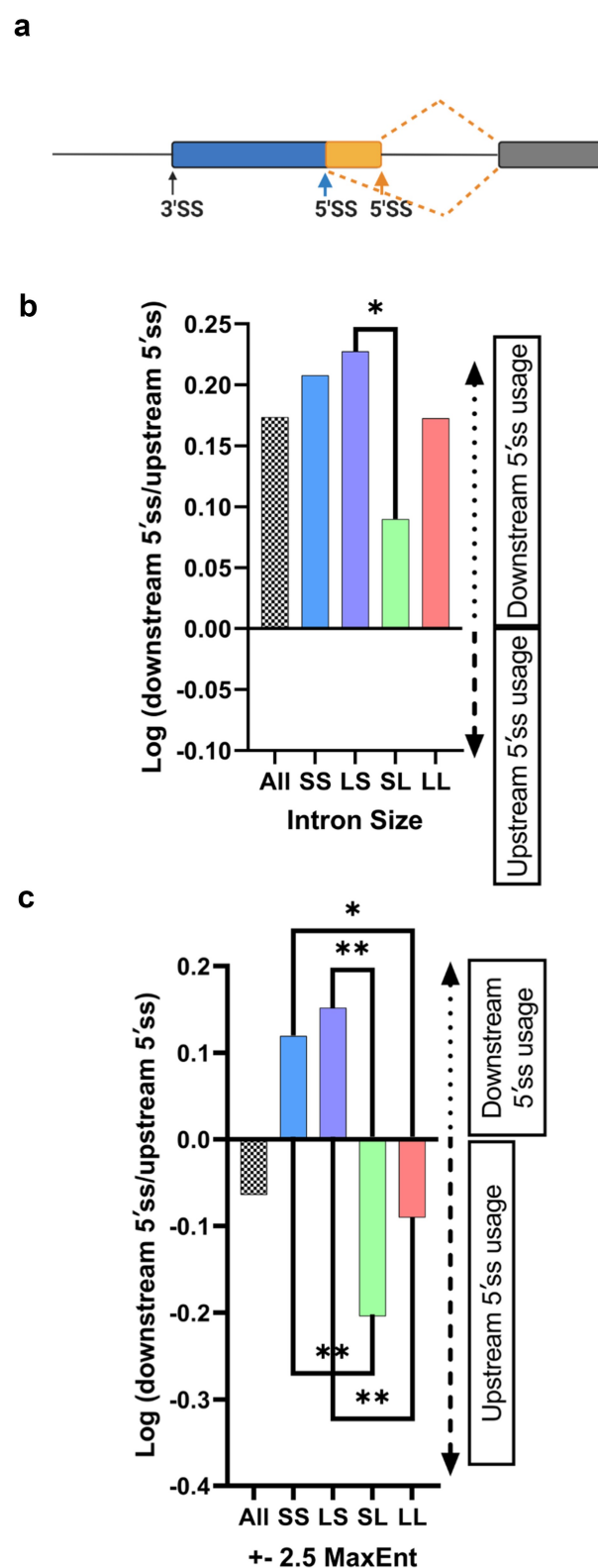
correct splice site in the presence of multiple nearby cryptic or alternative splice sites are still not completely understood. Differences in intron-exon architecture and splice site strength are known to be important in mediating alternative splice site selection [8]. A series of classical experiments demonstrated that the proximity between the 5' and 3' splice sites, across the intron, plays a crucial role in splice site preference [11]. Reed and Maniatis showed that the splice site closest to its intronic splicing partner was favoured over a distal competing splice site [11]. Thus, in the case of competing alternative 5' splice sites, the downstream 5'ss was preferred because it was more proximal to the pairing 3'ss. Similarly, between competing 3' splice sites, the upstream 3'ss was chosen. These observations suggest that in the absence of confounding factors, shorter distances between splice sites are favoured during intron-defined splicing. This may be because splice site pairing is more efficient across shorter distances. These experiments established a splice site selection proximity rule (for clarity referred to as the intron-centric proximity rule); however, it is unclear how dominant it is within the hierarchical nature of known splicing determinants.

In this study, we carried out computational analyses to assess the impact of the intron-centric proximity rule. We demonstrate that the intron-centric proximity rule is generally applicable for the intron definition mode of splice site definition. For the exon definition mode of splice site definition, we observe an exon-centric proximity rule that deviates from the classical intron-centric proximity rule. The 5' and 3' splice sites closest to each other across the intron are preferentially selected. Thus, when the unit of splice site definition is across the intron (intron definition), the 5' and 3' splice sites closest to each other across the intron preferentially pair. When the unit of splice site recognition is the exon (exon definition), the 5' and 3' splice sites closest to each other across the exon are preferentially selected. Our results provide evidence that alternative splicing decisions are influenced by the intron and exon definition modes of splice site recognition.

## Results

### *The influence of intron-exon architecture on 5' splice site selection*

To determine the impact of the intron-exon architecture and splice site strength on splice site selection, we created a database of alternative splice sites (ALTssDB) using the Human Exon Splicing Event Database HEXEvent [13], the Intron DB [14] and GeneBase [15]. MaxEntScan, a computational tool, was used to assign splice site scores [6]. To minimize variability, we focused on competing alternative 5' or 3' splice site pairs of internal exons with only one alternative splice pattern. Thus, ALTssDB catalogs pairs of alternative 5' splice sites competing for a common 3'ss or pairs of alternative 3' splice sites competing for a common 5'ss. ALTssDB reports the location of the major splice site and its competing alternative 3' or 5' splice site, corresponding exon sizes, usage levels, splice site scores and flanking intron lengths. Using these filters, ALTssDB captures 4,171 human 5'



**Figure 2.** 5' ss selection preference for different internal exon categories. (a) Model depicting alternative 5'ss patterns. (b) Bar graph depicting the preference for downstream or upstream 5' ss selection for different internal exon categories (sample size = 379 SS, 2,747 LL, 570 LS, 475 SL). A positive log ratio represents downstream 5' ss preference, a negative log ratio represents upstream 5' ss preference. (c) Splice site selection preference for alternative 5' splicing events with near equal splice-site strength scores ( $\Delta \pm 2.5$  MaxEnt, sample size = 88 SS, 725 LL, 104 LS, 143 SL), Fisher's exact test was performed. (b, c), \*\* $p < 0.01$ , \* $p < 0.05$ .

ss competition events and 6,919 human 3'ss competition events (Figure 1(b-d)).

We first tested whether the intron-centric proximity rule holds true when evaluating all alternative 5'ss events transcriptome-wide (Figure 2(a)). In agreement with the intron-centric proximity rule expectation that the downstream 5'ss should be selected over a competing upstream 5'ss, we observed a preference for downstream 5'ss selection in ~60% (2,497) of the alternative 5'ss splicing events (Figure 2(b), left bar).

To evaluate whether the 'intron definition' or 'exon definition' mode of splice site selection influence adherence to the intron-centric proximity rule, we parsed the 5'ss dataset into intron definition events (379 SS), exon definition events (2,747 LL), and hybrid events (570 LS, 475 SL) (Figure 1(b and c)). For the purpose of alternative 5'ss selection analysis, the hybrid architectural class LS was categorized as intron defined because the 5'ss is adjacent to a short intron and U1 snRNP binding to the 5'ss at the exon/intron junction initiates early spliceosome formation [16]. By analogy, the architectural class SL was considered exon defined because the 5'ss is contained within a long intron. Surprisingly, in all four intron architecture classes, the majority of events still displayed a preference for the downstream 5'ss, consistent with the intron-centric proximity rule, albeit to varying degrees (Figure 2(b)). For example, the downstream 5'ss is selected more frequently for intron definition events (represented by SS, LS) when compared to exon definition events (represented by LL, SL). These varying degrees of preference suggest that the intron definition mode of splice site selection adheres more stringently to the intron-centric proximity rule.

### The influence of intron-exon architecture on 5'ss selection in the absence of splice site strength differences

One important determinant that may mask the influence of splice site proximity is the difference in the splice site strength of competing splice sites. To determine the impact of splice site strength on alternative 5'ss selection, we compared the splice strength of the major 5'ss versus the alternative 5'ss. In 86% of the events evaluated the 5'ss with a higher predicted splice strength was the dominant 5'ss, irrespective of whether the exon was predicted to be recognized through exon definition (LL, SL) (85%) or intron definition (SS, LS) (90%) events. These results support the notion that splice site strength is a strong determinant in alternative 5'ss selection.

To determine how the exon and intron definition modes of splice site selection influence alternative splicing the impact of splice site strength differences was minimized computationally. This was achieved by isolating 5'ss competition events

with near equal splice site scores ( $\Delta\text{MaxEnt} = \pm 2.5$ ), resulting in 88 SS, 725 LL, 104 LS, and 143 SL events. Interestingly, when this splice site strength filter was applied, we observed that the upstream 5'ss is preferentially selected in 60% of competition events, inconsistent with the expectations of the intron-centric proximity rule (Figure 2(c), left bar). Strict intron definition events (SS category) display a downstream 5'ss selection preference, consistent with the intron-centric proximity rule, while strict exon definition events (LL) display a preference for the upstream 5' splice site (Figure 2(c)). The upstream preference under exon definition is inconsistent with the intron-centric proximity rule but consistent with an exon-centric proximity rule. These biases are heightened in the hybrid categories SL (upstream preference) and LS (downstream preference) (Figure 2(c)). These results suggest that for exon definition events the upstream 5'ss, which is proximal across the exon to the upstream 3'ss, is favoured. By contrast, for intron definition events, the 5'ss proximal across the intron to the downstream 3'ss is favoured.

### The influence of exon size on 5'ss selection

It is known that exon size can influence splice site selection [9,10]. To determine the influence of exon size on splice site selection, we compared splice patterns between three different exon size groups, exons smaller than 50 nts, exons between 50–250 nts in length, and exon longer than 250 nts. These cut-offs were chosen based on natural exon size distributions. We then calculated how frequently the major isoform contains the stronger 5'ss for the three different exon size classes (Table 1). When the major and the alternative exons are smaller than 50 nts, splice preference is driven almost exclusively by the stronger splice site score (Table 1). This preference weakens when the usage of the alternative 5'ss generates an exon greater than 50 nts. Thus, differences in exon size contribute to splice site selection, with a preference for generating shorter exons. A similar trend is observed for alternative patterns of major exons within the 50–250 nts range. The selection of alternative exons larger than 250 nts is much less likely to be driven by splice site differences. These data provide evidence that exon size contributes to splice site selection with a preference for defining smaller exons.

### Experimental verification of genome-wide computational analysis

To test whether the proposed exon-centric proximity rule can be confirmed experimentally, we tested five minigenes that contain an internal exon with two competing 5' splice sites of identical strength (MaxEnt 10.9, CAG/guaagu) and one 3'

**Table 1.** Alternative 5'ss selection and resulting exon length correlation. The table reports how frequently the major isoform contains the stronger splice site when the alternative splice site lies within one of three different exon size classes. <sup>a,b</sup>Within a column, means without a common superscript differ ( $p < 0.05$ ) between size categories in each column. <sup>1</sup>29% preference for the upstream 5'ss. <sup>2</sup>42% preference for the upstream 5'ss. <sup>3</sup>44% preference for the upstream 5'ss.

Exon size generated with minor splice site usage	Exon size generated with major splice site usage			
	Ex $\leq 50$ nts	Ex $\leq 50$ nts	50 < Ex $\leq 250$ nts	Ex > 250 nts
Ex $\leq 50$ nts	98% <sup>a1</sup>	86% <sup>a</sup>	100% <sup>a</sup>	
50 < Ex $\leq 250$ nts	73% <sup>b</sup>	87% <sup>a2</sup>	88% <sup>a</sup>	
Ex > 250 nts	75% <sup>b</sup>	58% <sup>b</sup>	77% <sup>b3</sup>	

splice site with a MaxEnt of 12.56 (uguccuuuuuuuccacag/CUG) (Figure 3(a)). All minigenes were designed to be recognized through exon definition (flanking intron size of 365 nts) and differ only in the resulting internal exon size. Cell transfection experiments demonstrated that for all constructs tested the upstream 5'ss was chosen exclusively (Figure 3(b)), consistent with the computational analysis demonstrating that upstream 5' splice sites are favoured under exon definition. To test if this splice site preference is altered when both competing splice sites are weakened, we mutated both 5' splice sites to have a MaxEnt score of  $-0.5$  (GAG/guguca). In the larger exon constructs (L and XL), this resulted in preferential internal exon skipping. In the M and S constructs, the upstream 5'ss maintained its preference (Figure 3(c)). These results demonstrate that in an isogenic exon definition context the 5'ss most proximal to the upstream 3'ss is favoured, supporting the computational analysis of an exon definition 'cross-exon proximity' preference.

### The influence of intron architecture on 3'ss selection

To investigate the impact of intron size and splice site strength on 3'ss selection we built a 3'ss dataset analogous to the 5'ss dataset described above (Figure 4(a)). For our analysis, we took into consideration that the 3'ss is recognized during the first and the second steps of splicing. Prior to the first step of splicing, the polypyrimidine tract is bound U2AF, which subsequently recruits U2 snRNP to the branch point. After the first step of splicing, the 3' splice junction YAG/N is selected before the exons are ligated via a transesterification reaction. It has been demonstrated that competing 3' splice sites in close proximity (up to 9 nts) are selected during the second step of splicing after identical first step definition [17]. Alternative 3' splice sites further apart (greater than 12–20 nts) are typically defined during initial splice site recognition using different polypyrimidine tract and branch points. Thus, we split the 3'ss dataset into 'first step recognition' ( $\geq 20$  nts apart from one another, 3839 events) and 'second step recognition' events ( $\leq 9$  nts apart from one another, 2317 events). Both 3'ss event groups show a preference for upstream 3'ss usage, consistent with the intron-centric proximity rule (Figure 4(b)). This preference is particularly strong for the second step alternative 3'ss events. Filtering to obtain competing 3'ss pairs with comparable strengths and categorizing these events into intron (S/S, 69 events) or exon definition (L/L, 664) events again demonstrated the influence of the intron architecture on 3'ss selection (Figure 4(c)). The strong upstream 3'ss preference observed for intron defined events (S/S) is significantly reduced when splice sites are selected in the exon definition mode (L/L). Consistent with our 5'ss analysis, the hybrid classes (SL, 88 events and LS, 147 events) display more extreme splice site preferences relative to the SS and LL classes, with SL mimicking intron definition and LS mimicking exon definition behaviour.

Together, our transcriptome-wide analyses demonstrate that the mode of splice site selection critically influences splice site choice. For intron definition, splice sites closest across the intron are preferentially selected. Under exon definition, the selection of splice sites closest across the internal exon are

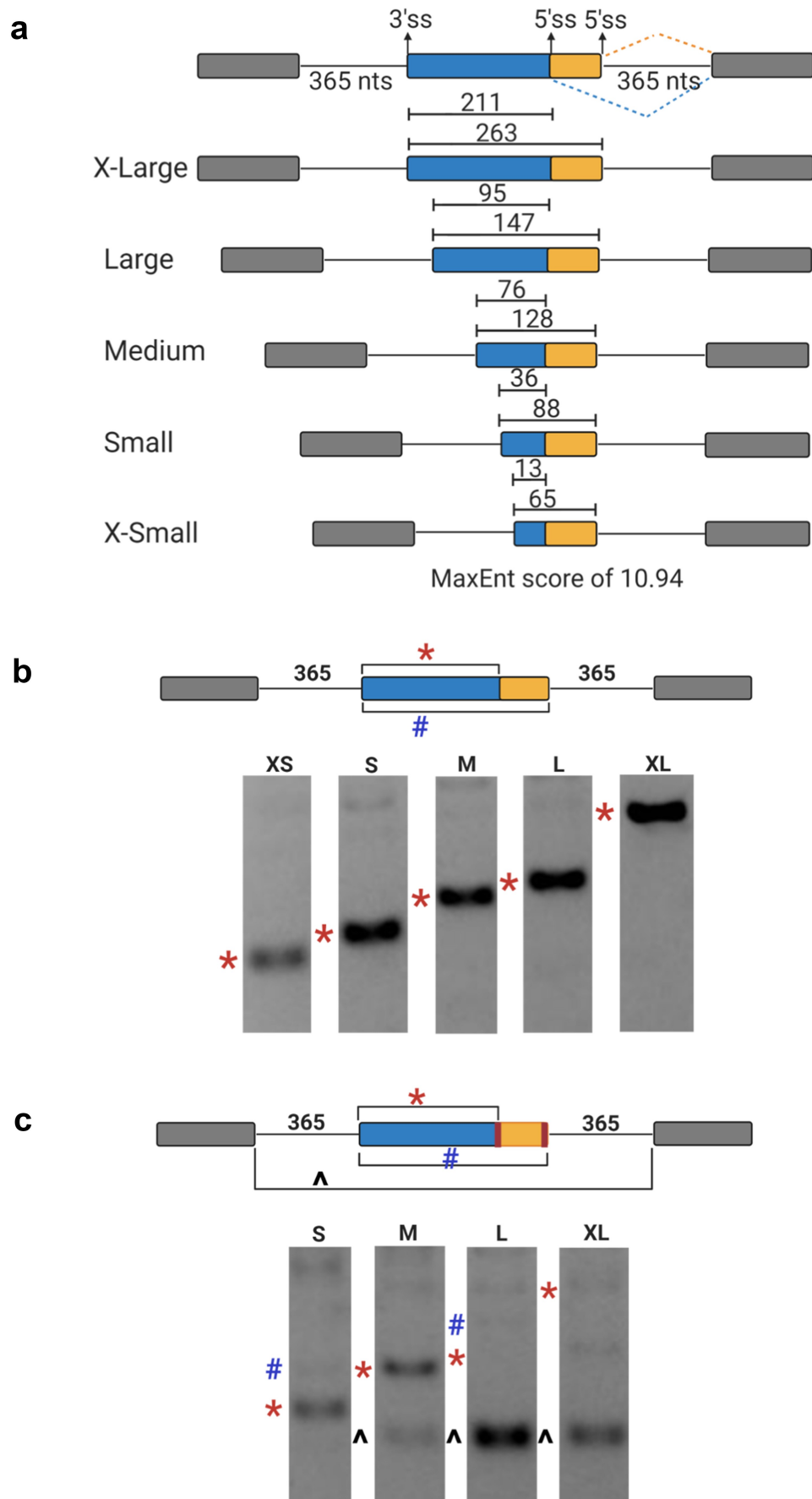
favoured. These results suggest that the gene architecture influences alternative splicing by promoting splice site recognition via the intron or exon definition pathway.

## Discussion

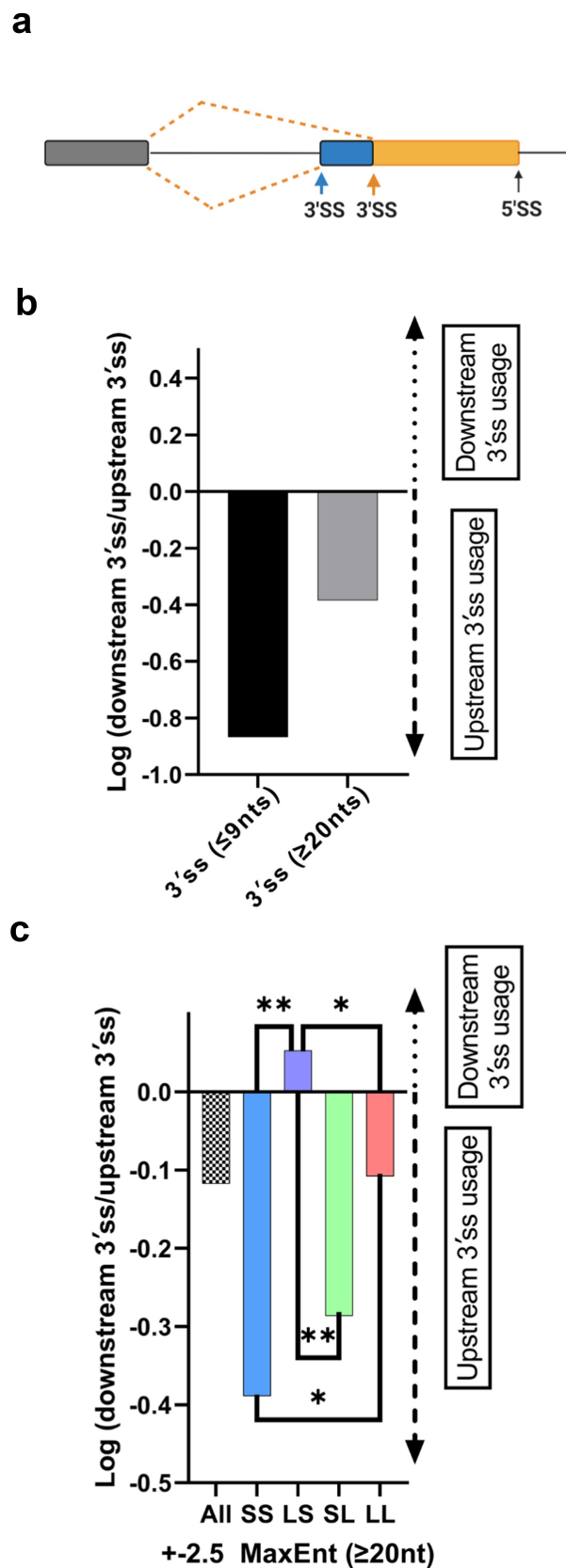
The regulation of pre-mRNA splicing is a combinatorial process that is controlled by splice site sequences, cis-regulatory elements binding trans-acting factors, the intron-exon architecture, and RNA secondary structure among other features [16]. Two mechanisms of splice site recognition have been proposed within the broader concept of intron-exon architecture. It has been postulated that under the intron definition splice sites are recognized across the intron, making the intron the initial unit recognized by the spliceosome. In an alternative mode of splice site recognition, splice sites are postulated to be initially recognized across the exon in a process called exon definition. Once the exon is defined as the initial unit of splice site recognition, subsequent structural rearrangements are predicted to recognize and pair the upstream and downstream splice sites across flanking introns [18]. The mechanisms of intron and exon definition have been studied in the field for almost 30 years [9,10,18–23].

Early evidence that the length of introns and exons is important came from size constraints on exon inclusion from minigenes that were transfected in cell culture. Large exons were efficiently spliced when flanked by short introns, consistent with an intron definition mechanism. However, when intron lengths were increased exons were only included efficiently if they were relatively short, less than  $\sim 500$  nts long. The latter observation suggests that the early spliceosome has a limited 'wing-span' when the exon is the unit of initial splice site recognition. Subsequently, biochemical studies demonstrated that intron definition is more efficient and that the rate of splicing for exon defined substrates is considerably slower. This study identified intron length as the primary determinant in the mode of splice site recognition employed by the early spliceosome and placed the transition from intron definition to exon-definition at the point when flanking introns become longer than 200–250 nts [10].

Another classical study used *in vitro* splicing assays to demonstrate that alternative splice site choice is influenced by the proximity between the pairing splice sites. When two splice sites are in competition, the splice site proximal to the intron is preferred. As a result, this proximity bias induces the preferential excision of the smaller intron [11]. This study and the pioneering study from Sterner and Berget when analysed together suggest that in the context of splice site competition, selection of proximal splice sites across an intron may allow the intron to be recognized through intron definition, while the selection of the distal splice site may lead to a larger unit of initial splice site recognition that may change the mode of splice site recognition all together [9,11]. In broader terms, the findings by Reed and Maniatis [11] indicated that perhaps proximity across the initial unit of splice site recognition would drive splice site selection and influence alternative splicing. We set out to determine whether the proximity of splice sites across the proposed initial unit of splice site recognition may provide genome-wide evidence for the two



**Figure 3.** Cross-exon selection of alternative 5' alternative splice sites. (a) Schematic of exon-defined mini-gene constructs with identical splice site strength (CAG/guaagu, MaxEnt = 10.9) used in transfection experiments. The size of the resulting internal exon is indicated for upstream (blue) and downstream 5' ss selection. (b) Representative image of ethidium bromide stained agarose gel splicing analysis. Bands denoting upstream (red symbol) or downstream (blue symbol) 5' ss usage are marked to the left of the image. (c) Splicing outcome of minigene constructs with identical but weakened competing 5' ss (GAG/guguca, MaxEnt = -0.5). Bands denoting upstream 5' ss usage (red symbol), downstream (blue symbol) 5' ss usage, or exon skipping (black symbol) are marked to the left of the image.



**Figure 4.** 3'ss selection preference for different internal exon categories. (a) Model depicting alternative 3'ss patterns. (b) Bar graph displaying the 3'ss preference for first (>20 nts distance between competing 3'ssplice sites, 2317 events) or second (<9 nts distance between competing splice sites, 3839 events) step selection. A positive log ratio represents downstream 3'ss preference, a negative log ratio represents upstream 3'ss preference. (c) Bar graph depicting the preference for downstream or upstream 3'ss selection with near equal splice-site strength scores for different internal exon categories ( $\Delta \pm 2.5$  MaxEnt/sample size = 69SS, 664 LL, 88 SL, 147 LS). Fisher's exact test was performed. (b, c), \*\*p < 0.01, \* p < 0.05.

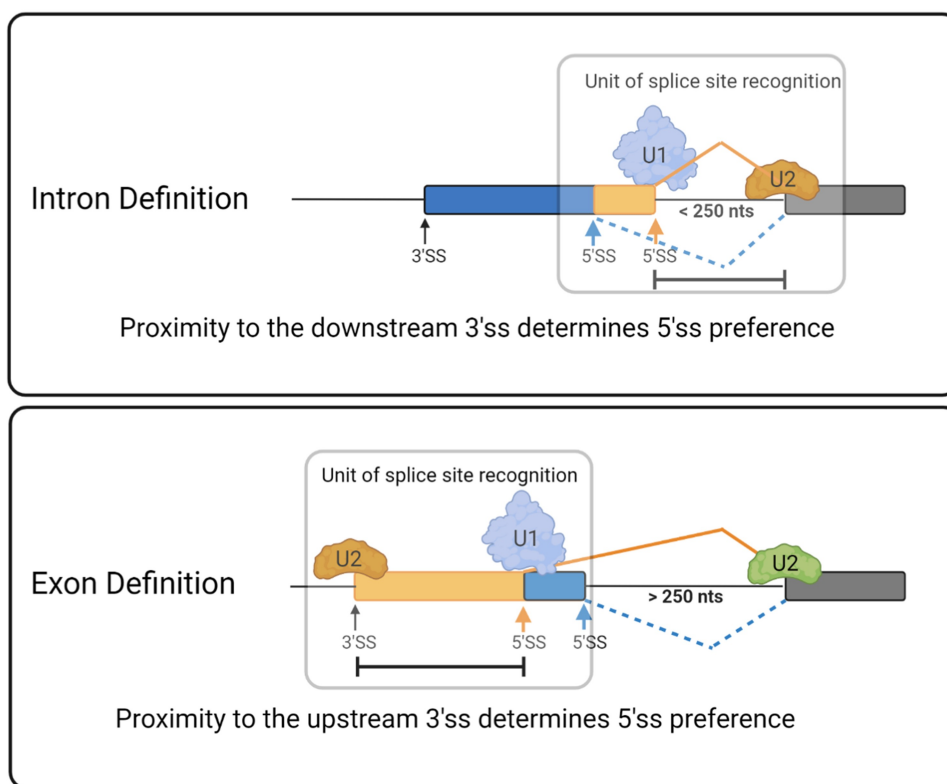
modes of splice site recognition and elucidate their roles in alternative splicing.

Our analysis of the alternative splicing events captured in ALTssDB permitted the derivation of several important conclusions. First, the intron-centric proximity rule observed by Reed and Maniatis is maintained within the context of the intron definition mode of splice site recognition [11]. In the context of exon definition, we observe an exon-centric proximity rule, where the proximity between 5' and 3' splice sites across the exon dictates splice site preference. Alternative exons subject to the intron-centric proximity rule undergo removal of the smaller intron and selection of the larger exon. Conversely, alternative exons subject to the exon-centric proximity rule undergo removal of the larger intron and selection of the smaller exon. Initially, these observations may appear inconsistent with each other, yet they highlight a commonality of spliceosomal assembly across the smallest unit of initial splice site recognition. For the intron definition mode of splice site recognition this unit is the intron, meaning the spliceosome assembles around the 5' and 3' splice sites that define the intron to be excised (Figure 5, top cartoon). For the exon definition mode of splice site recognition, the unit of recognition is the exon, meaning that initial splice site recognition by the spliceosome occurs across the exon (Figure 5, bottom cartoon). In both modes of splice site recognition, a preference for splice site selection that promotes the definition of the smaller initial recognition unit (as defined by the number of nucleotides) is observed. Thus, the proximity of 5'

and 3' splice sites within the unit of initial recognition determines preferential splice site selection (Figure 5). We therefore conclude that an additional mechanism of alternative splicing can be the proximity of splice sites across the initial unit of definition.

Since the initial concepts of intron and exon definition were introduced, generating supporting evidence for the existence of these two proposed modes of splice site recognition has been challenging. Initial studies were limited to select cases where insights were gained from transfected designer minigenes or *in vitro* transcribed RNAs spliced using the nuclear extract system [9,10,19,21–23]. These studies, while mechanistically enlightening, could not be extrapolated to the entire transcriptome.

Recent analyses of *in vivo* splicing kinetics offer more comprehensive insights into the mechanisms of exon recognition. These studies lend support to the notion that exon and intron definition events display different global splicing kinetics. They also raise questions about the generality of exon definition and intron definition [24–26]. A single molecule intron tracking technique was used to determine the amount of splicing as a function of RNA polymerase II position along the gene. This technique and an orthogonal nanopore-based variation found splicing rates to be strikingly fast in *Saccharomyces cerevisiae* [25] demonstrating that 50% of splicing can be completed 1.4 seconds after 3'ss synthesis for the genes studied. The onset of splicing for a subset of the analysed genes was detected only 26 nucleotides after



**Figure 5.** Unifying model for the influence of splice site proximity in alternative exon definition. Depending on the size of flanking introns the splice sites of internal exons are initially recognized across the intron (top – intron definition) or across the exon (bottom – exon definition). In both scenarios, the 5' and 3' splice sites closest to each other across the unit of initial splice site recognition are preferentially selected. Thus, in intron definition 5' and 3' splice sites across the intron are preferentially selected. In exon definition, 5' and 3' splice sites across the exon are preferentially selected.



transcription of the 3' ss. The observation that splicing can be completed before the entire exon is transcribed is consistent with an intron definition mechanism in *Saccharomyces cerevisiae*, but begs the question is exon definition possible in lower eukaryote? The average *Saccharomyces cerevisiae* exon is ~1400 bases suggesting that exon definition would be highly unlikely for those genes where splicing rates were calculated to occur on the order of several seconds [27]. However, a recently proposed unifying model provides evidence for exon definition in *Saccharomyces cerevisiae* [28]. Electron microscopy analyses suggest that the splicing factor Prp40 can bridge the 5' ss bound U1 snRNP and branch point sequence bound BBP/Mud2 (SF1/U2AF65 homologs) either across the intron or across the exon to define E-complex. Structural evidence for exon definition in *Saccharomyces cerevisiae* was supported by genetic and biochemical analysis, which included the circularization of single exon constructs in yeast splicing extracts. The latter study provides strong structural, biochemical, and *in vivo* evidence for exon definition, even in *Saccharomyces cerevisiae*, where most splice sites would be expected to be recognized through intron definition [28].

Regarding the intron-exon architecture of higher eukaryotes, ligation of 3' adapters and long read nanopore sequencing of nascent RNA were used to determine the splicing rates in *Drosophila* and human cells [26]. The nano-COP method determined that the majority of splicing in *Drosophila* occurs within 2 kilobases, once the 3' ss has been transcribed. This is in contrast to human cells where the majority of splicing is completed ~4 kilobases past the 3' ss [26]. The rate of splicing calculated from nano-COP is consistent with previous  $t_{1/2}$  measurements that are ~2 minutes for *Drosophila* and 7–14 minutes for mammalian cells [24,29–31]. Interestingly, nano-COP found that *Drosophila* introns less than 100 nts in length were spliced more quickly than introns greater than 300 nts, suggesting that intron definition is more efficient than exon definition [26]. These results are supported by an earlier study that used progressive metabolic labelling and also found a local maximum of splicing rates for introns that were 60–70 nts long [24]. However, the latter study also found that a subset of very long introns (>2944 nts) was spliced even more quickly with a  $t_{1/2}$  of ~1.5 minutes suggesting gene level and pathway-specific splicing programmes may have evolved to utilize the rapid splicing that very long exon-defined introns undergo. Taken together these kinetic measurements suggest that while exon definition is broadly less efficient and intron definition is broadly more efficient as was first shown by Fox-Walsh and Hertel [10], exceptions do exist.

Recent investigations provide further support that both intron definition and exon definition occur *in vivo* [32,33]. However, these studies present evidence that the mechanism by which splice sites are initially recognized is dictated by the difference in GC content, referred to as GC differential, between the exon and the flanking introns. Specifically, two architectures are described, referred to as the 'differential architecture' and the 'leveled architecture' [32,33]. 'Differential architecture' exons have a low GC content, and their flanking introns have an even lower GC content. 'The leveled architecture' exons are characterized by a high GC

content, less difference in the GC content of flanking introns and short introns. The former class of exons was demonstrated to be localized to the nuclear periphery and recognized through exon definition while the latter was shown to be localized to the nuclear centre and recognized through intron definition. Altering the GC content between exon and the downstream intron can be used to alter the mode splice site recognition, without changing the length of the intron. These observations suggest that intron length may not be the determining factor in the mode of splice site recognition for exon definition [32,33]. It will be important for future exon definition studies to consider the GC content across the exon and flanking introns. For example, a recent analysis of high-throughput mutagenesis data for an alternatively spliced exon in the proto-oncogene *RON* demonstrated that the alternatively spliced exon is recognized through exon definition, even though it is flanked by short introns on either side (87 and 80 nts) [34]. Thus, splice sites of short introns can be recognized through exon definition, perhaps because the unique GC content that typifies exon definition splice sites.

Finally, a recent transcriptome-wide study demonstrated that introns that undergo efficient co-transcriptional splicing have sharp structural transitions across the intron-exon boundary [35]. These introns display a peak of RNA structure downstream of the 5' ss and upstream of the 3' ss. Furthermore, some introns displayed enhanced co-transcriptional splicing under conditions where the elongation rate of RNA polymerase II was slowed down genome-wide, a process that promotes increased RNA folding. The latter group of introns had significantly steeper structural transitions when transcription was slow [35]. GC content is an indicator of the potential to form RNA secondary structures [36]. Thus, it may be the case that the differential architecture associated with exon definition is driven partially by the propensity for RNA secondary structure formation that can help delineate the intron-exon boundary.

We set out to determine the degree of agreement between the intron length-dependent definitions of 'intron-defined' and 'exon-defined' splice sites with the 'leveled' and 'differential' architecture. We calculated GC content differentials between the LL and SS architectural classes of alternatively spliced 5' and 3' ss exons. Remarkably, the intron length-defined LL and SS categories closely resemble the 'differential' and 'leveled' architectures respectively [32,33] (Supplemental Figure 2). Thus, the GC content, as defined the Amit et al. [32], of long introns (>250 nts) differs significantly from the GC content of short introns (<250 nts), suggesting that GC content or intron size definitions are comparable approaches to define exon and intron definition modes of splice site recognition. This notion is supported by evolutionary analyses that show the emergence of a distinct differential GC architecture as intron lengths increased through vertebrate evolution [37]. Thus, the emergence of the 'differential architecture' may be a co-evolutionary adaptation to define exons in the context of expanding introns. To evaluate whether the use of proximal or distal splice sites changes 'leveled' and 'differential' architecture designations, we calculated GC content for alternatively spliced exons captured by ALTsDB. Interestingly, the resulting GC differential does not change

significantly (Supplemental Figure 2), suggesting that alternative splice site selection is not dependent on differential GC content but contingent on defining the smallest unit of initial recognition.

Collectively, the results of our transcriptome-wide analysis of alternative splice site usage provide evidence that exon definition and intron definition do occur transcriptome-wide. When exons are flanked by long introns, the spliceosome tends to favour splice sites located internally within the exon being defined. By contrast, the spliceosome tends to move into the intron for splice site definition for exons flanked by short introns. These observations suggest that the spliceosome can define the exon and the intron independently.

Our computational analysis of alternative 3'ss events permitted an evaluation of alternative 3'ss selection in the context of first or second step recognition. Initial 3'ss selection is mainly driven by the strength of the polypyrimidine tract and the presence of a consensus branch point. Upon recruitment of U2 snRNP to the branch point and tri-snRNP incorporation, the first step of the splicing reaction is initiated without engaging the 3'ss junction. After spliceosome rearrangements, the 3'ss junction is selected during the second step of splicing as the spliceosome aligns the AG/N intron/exon junction into the active site. It is well established that competing 3'AGs in close proximity (less than 9 nts) use the same upstream polypyrimidine tract and branch point and that their selection is directed during the second step of splicing [17]. Interestingly, our analysis of alternative 3'ss selection in close proximity demonstrated that the upstream AG/N junction is almost exclusively chosen over the downstream AG/N. Thus, it appears that aligning the closest AG/N 3'ss junction is the default pathway of second step splice junction selection (Figure 4(b)).

The intron-exon architecture of genes is a major driver of splice site selection. Since the initial postulation of these two modes of splice site recognition, various forms of evidence have been presented, often in form of kinetic principles supporting intron or exon definition. However, measurements of splicing rates as a function of intron length do not constitute direct evidence of alternative spliceosomal assembly pathways. The ability of yeast E-complex to assemble across the intron or the exon is perhaps the strongest evidence yet for exon definition. Our study provides support for exon definition by demonstrating the spliceosome favours internal splice sites within exons when the splice site strengths of competing sites are comparable. This suggests that the exon is being defined and not the intron. This study provides a unifying model for splice site selection, whereby the spliceosome assembles across the smallest unit of initial splice site recognition. In the case of intron definition, this entails removal of smaller introns and inclusion of larger exons. Indeed, studying the evolutionary trends in intron-exon architecture, lower eukaryotes tend to have larger exons and smaller introns. Upon intron expansion and a gradual shift towards exon defined gene architecture, the initial unit of splice site recognition often tends to be the exons. This may be due to the increased number of decoy splice signals associated with larger genome sizes. It would therefore be expected that the

smaller exons would be favoured in higher eukaryotes. This trend is also broadly observed from yeast to humans. It is possible that the exon-centric proximity rule is an evolutionary adaptation to accurately recognize exons surrounded by long stretches of intronic sequence. Our results not only provide *in vivo* and transcriptome-wide evidence for exon definition, they also demonstrate that exon and intron definition influence alternative splicing in the context of alternative 5' or 3' splice site competition.

## Methods

### Construction of ALTssDB

ALTssDB was created using EST data from the Human Exon Splicing Events (HEXEvent) database [13]. HEXEvent contains information regarding the location of competing splice sites, the resulting exon sizes, alternative splice site usage levels and the gene associated with each mRNA. The HEXEvent data was filtered to obtain a dataset comprising of only pairs of competing 5' and 3' splice sites separately. This database was subsequently modified to include splice site junction information and splice site strength scores using MaxEntScan [6]. Although other approaches exist to evaluate the strength of 5 splice sites [38,39], MaxEntScan is the preferred tool as it also permits comparable splice site score derivation for 3 splice sites. Using an R script and IntronDB dataset, (a database detailing eukaryotic intron features) flanking intron lengths were added to the database [14]. Alternative splicing events were further filtered to include only events that have 10 or more EST counts. The data was filtered into four categories according to intron length and included: both flanking introns around the exon of interest being short (<250 nts, SS), both flanking introns being long (>250 nts, LL), the upstream intron being short and downstream intron being long (SL) or the upstream intron being long and the downstream intron being short (LS). ALTssDB does not differentiate between canonical U2 introns and U12-type introns. Given their rarity and limited involvement in alternative splicing beyond intron retention, it is anticipated that U12-type introns are not well represented in ALTssDB [40].

ALTssDB does not distinguish between isoforms that originate from a tissue specific splice switch or disease comparison. It lists all known splice patterns for a particular exon, independent of origin. EST data was used to build AltssDB to obtain high enough numbers of alternative splice site choices within the human genome to carry out all analyses. While datasets for tissue-specific splicing are available, the quantity of significant alternative splice site events is limiting.

### Plasmid design

Five minigene constructs were designed containing three exons and two introns. The plasmid design was based primarily on previously validated constructs used to study splice site strength [7]. The internal exon was designed to contain two functional competing 5' splice sites (CAG/guaagu), with equal MaxEnt scores MES of 10.9, separated by 52 nucleotides. The

sequence preceding the upstream 5' splice site was progressively shortened (Figure 1(b)). Additional constructs were created where the MES of both competing 5' splice sites were changed from 10.9 to -0.5 (GAG/guguca) for S, M, L, and XL plasmid. Lastly, the upstream 5' splice sites were changed from MaxEnt = 10.9 to MaxEnt = -5.2 (UCG/gucgau) for the S, M, and XL to show that the downstream 5'ss is viable (Supplementary Figure. 1).

### Cloning protocols to change splice site strength sequences

To linearize the plasmids, 10 nanograms (ng) of plasmid DNA obtained by midiprep was amplified using divergent primers. PCR reactions were carried out with NEB® Phusion® polymerase in 50 µL according to NEB protocols. DH5α E.coli midiprep derived plasmids in the PCR reaction were digested with 40 units of DpnI according to NEB protocols. Plasmids were purified with Zymo DNA clean and concentrator™ kit and DNA concentrations were obtained using a nanodrop 2000 instrument. For each construct, 0.03 picomoles of linearized plasmid DNA was mixed with a 10X molar ratio of phosphorylated double stranded DNA inserts, purchased from IDT, in 20 µL ligation reactions. Synthetic inserts were cloned into linearized vectors using T4 ligase according to the standard NEB protocol and 10 µL of the ligation reaction was transformed using in house DH5α E.coli cells. Colonies were screened using PCR to detect the correct size insert. Colonies with the correct size insert were grown from 3 mL cultures to 20 mL cultures and underwent midiprep DNA extraction. Plasmid DNA from each colony was sequenced to ensure the correct orientation of inserts.

### Cell transfections and RT-PCR Analysis

Transfection experiments were performed in triplicate using HeLa cells. 1 mL of  $0.1 \times 10^6$  cells/mL was plated into each well of 12 well plates. Cell confluency was checked 24 hours later and 1 µg of plasmid DNA was transfected according to Bioland Scientific's BioT protocol. Cells were harvested 48 hours post-transfection. Each well was washed two times with phosphate buffered saline (PBS) and subsequently RNA was extracted with the standard Trizol™ protocol. RNA pellets were resuspended in 50 µL water and put through ZYMO RNA Clean and Concentrator™ columns. Sample volumes were adjusted to 80 µL, yielding RNA concentrations of  $\leq 200$  ng/µL. DNase digestion was performed with Turbo™ DNase (Ambion®) according to Ambion's protocol in 100 µL reactions. RNA was subsequently extracted with 100 µL phenol: chloroform and the aqueous phase was put through ZYMO RNA Clean and Concentrator™ columns. DNase digested and purified RNA samples were resuspended in 25 µL. A nanodrop 2000 instrument was used to obtain RNA concentrations. Reverse transcription reactions were carried out in 20 µL using 250 ng of total RNA and 200 ng of OligodT18 primer according to SuperScript™ III protocol. PCR primers are as followed: first exon forward primer (5'cgctcgtcctcactctcttc3') and third exon reverse primer (5'agatcccaaggactcaaga3').

PCR primers were designed that bound the flanking exons and thus would detect upstream, proximal or downstream, distal 5' splice site usage. PCR reactions contained 5 µL cDNA (10% vol:vol), 0.2 mM dNTPs, 0.2 µM of each primer, 1.5 mM MgCl<sub>2</sub> and 0.25 units taq polymerase (Apex BioResearch). Semi-quantitative PCR using long extension times to limit PCR product size bias was carried out to demonstrate that the ratio of upstream and downstream splice site usage, or the alternative exon skipping pattern, remained constant throughout the dynamic linear range of the amplification reaction (data not shown). Based on these results 25 cycles of PCR were performed for each sample and 5 µL was subsequently loaded onto a 2% agarose gel and stained with ethidium bromide. Agarose gels were run at 150 V for 1 hour in 1X Tris-Borate EDTA (TBE).

### Calculating splice site selection preference

5'ss selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 5'ss over the upstream 5'ss. 3' splice site selection preference was determined by calculating the log ratio of the number of splice site events preferring the downstream 3'ss over the upstream 3'ss.

### Acknowledgments

This work was supported by grants from the NSF (DGE-1321846 to F.C.) and the NIH (R01 GM062287 to K.J.H).

### Data Availability Statement

The data that support the findings of this study are openly available in Dryad at ([https://datadryad.org/stash/share/kVSpUjVhjuHysLWLgG38JgYjy\\_UdcpODKw6DXs8](https://datadryad.org/stash/share/kVSpUjVhjuHysLWLgG38JgYjy_UdcpODKw6DXs8)).

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### Funding

This work was supported by the National Science Foundation [DGE-1321846]; National Institutes of Health [GM062287].

### ORCID

Klemens J. Hertel  <http://orcid.org/0000-0002-7560-9529>

### References

- [1] Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 2010 Jan;463(7280):457–463.
- [2] Li J, Wang Y, Rao X, et al. Roles of alternative splicing in modulating transcriptional regulation. *BMC Syst Biol*. 2017 Oct;11(5):89.
- [3] Fu XD, Ares M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet*. 2014 Oct;15(10):689–701.
- [4] Hertel KJ. Combinatorial control of exon recognition. *J Biol Chem*. 2008 Jan;283(3):1211–1215.

- [5] Erkelenz S, Mueller WF, Evans MS, et al. Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *RNA*. 2013 Jan;19(1):96–102.
- [6] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* 2004;11(2–3):377–394.
- [7] Shepard PJ, Choi E-A, Busch A, et al. Efficient internal exon recognition depends on near equal contributions from the 3' and 5' splice sites. *Nucleic Acids Res.* 2011 Nov;39(20):8928–8937.
- [8] Busch A, Hertel KJ. Splicing predictions reliably classify different types of alternative splicing. *RNA*. 2015 May;21(5):813–823.
- [9] Sterner DA, Carlo T, Berget SM, et al. Architectural limits on split genes. 1996; [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC26359/pdf/pq015081.pdf>
- [10] Fox-Walsh KL, Dou Y, Lam BJ, et al. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. 2005; [Online]. Available: [www.pnas.org/cgi/doi/10.1073/pnas.0508489102](http://www.pnas.org/cgi/doi/10.1073/pnas.0508489102)
- [11] Reed R, Maniatis T. A role for exon sequences and splice-site proximity in splice-site selection. *Cell*. 1986 Aug;46(5):681–690.
- [12] Movassat M, Forouzmand E, Reese F, et al. Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA*. 2019 Dec;25(12):1793–1805.
- [13] Busch A, Hertel KJ. HEXEvent: a database of human exon splicing events. *Nucleic Acids Res.* 2012 Oct;41(D1):D118–D124.
- [14] Wang D. IntronDB: a database for eukaryotic intron features. *Bioinformatics*. 2019 Nov;35(21):4400–4401.
- [15] Piovesan A, Caracausi M, Antonaros F. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database (Oxford)*. Vol. 2016, 2016. DOI:10.1093/database/baw153.
- [16] Shenasa H, Hertel KJ. Combinatorial regulation of alternative splicing. *Biochim Biophys Acta Gene Regul Mech.* 2019 Nov;1862(11–12):194392–194392. DOI:10.1016/J.BBAGRM.2019.06.003.
- [17] Dou Y, Fox-Walsh KL, Baldi PF, et al. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. *RNA*. 2006 Dec;12(12):2047–2056.
- [18] Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem.* 1995;270(6):2411–2414.
- [19] Robberson BL, Cote GJ, Berget SM, et al. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol.* 1990 Jan;10(1):84–94.
- [20] Schneider M, Will CL, Anokhina M, et al. Exon definition complexes contain the Tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. *Mol Cell.* 2010 Apr;38(2):223–235.
- [21] Talerico M, Berget SM. Intron definition in splicing of small *Drosophila* introns. *Mol Cell Biol.* 1994 May;14(5):3434–3445.
- [22] De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA.* 2013 Jan;4(1):49–60.
- [23] Lang KM, Spritz RA. RNA splice site selection: evidence for a 5' → 3' scanning model. *Science*. 1983 Jun;220(4604):1351–1355.
- [24] Pai AA, Henriques T, McCue K, et al. The kinetics of pre-mRNA splicing in the *Drosophila* genome and the influence of gene architecture. *eLife*. 2017 Dec;6: DOI:10.7554/eLife.32537
- [25] Carrillo Oesterreich F, Herzel L, Straube K, et al. Splicing of nascent RNA coincides with intron exit from RNA polymerase II. *Cell*. 2016 Apr;165(2):372–381.
- [26] Drexler HL, Choquet K, Churchman LS. Splicing kinetics and coordination revealed by direct nascent RNA sequencing through nanopores. *Mol Cell.* 2020 Mar;77(5):985–998.e8.
- [27] Koralewski TE, Krutovsky KV. Evolution of exon-intron structure and alternative splicing. *PLOS ONE*. 2011 Mar;6(3):e18055.
- [28] Li X, Liu S, Zhang L, et al. A unified mechanism for intron and exon definition and back-splicing. *Nature*. 2019 Sep;573(7774, Art. no. 7774):375–380.
- [29] Singh J, Padgett RA. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol.* 2009 Nov;16(11, Art. no. 11):1128–1133.
- [30] Rabani M, Raychowdhury R, Jovanovic M, et al. High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*. 2014 Dec;159(7):1698–1710.
- [31] Wachutka L, Caizzi L, Gagneur J, et al. Global donor and acceptor splicing site kinetics in human cells. *Elife*. 2019 Apr;8:e45056.
- [32] Amit M, Donyo M, Hollander D, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* 2012 May;1(5):543–556.
- [33] Tammer L, Hameiri O, Keydar I, et al. Gene architecture directs splicing outcome in separate nuclear spatial regions. *Mol Cell.* 2022 Mar;82(5):1021–1034.e8.
- [34] Enculescu M, Braun S, Thonta Setty S, et al. Exon definition facilitates reliable control of alternative splicing in the RON proto-oncogene. *Biophys J.* 2020 Apr;118(8):2027–2041.
- [35] Saldi T, Riemondy K, Erickson B, et al. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol Cell.* 2021 Apr;81(8):1789–1801.e5.
- [36] Shepard PJ, Hertel KJ. Conserved RNA secondary structures promote alternative splicing. *RNA*. 2008 Aug;14(8):1463–1469.
- [37] Gelfman S, Burstein D, Penn O, et al. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res.* 2012 Jan;22(1):35–50.
- [38] Wong MS, Kinney JB, Krainer AR. Quantitative activity profile and context dependence of all human 5' splice sites. *Mol Cell.* 2018 Sep;71(6):1012–1026.e3.
- [39] Freund M, et al. A novel approach to describe a U1 snRNA binding site. *Nucleic Acids Res.* 2003 Dec;31(23):6963–6975.
- [40] Olthof AM, Hyatt KC, Kanadia RN. Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics.* 2019 Aug;20(1):686.