MOLECULAR BIOLOGY

An interpretable model of pre-mRNA splicing for animal and plant genes

Kayla McCue[1,2] and Christopher B. Burge[1,2]*

Pre-mRNA splicing is a fundamental step in gene expression, conserved across eukaryotes, in which the spliceosome recognizes motifs at the 3′ and 5′ splice sites (SSs), excises introns, and ligates exons. SS recognition and pairing is often influenced by protein splicing factors (SFs) that bind to splicing regulatory elements (SREs). Here, we describe SMsplice, a fully interpretable model of pre-mRNA splicing that combines models of core SS motifs, SREs, and exonic and intronic length preferences. We learn models that predict SS locations with 83 to 86% accuracy in fish, insects, and plants and about 70% in mammals. Learned SRE motifs include both known SF binding motifs and unfamiliar motifs, and both motif classes are supported by genetic analyses. Our comparisons across species highlight similarities between non-mammals, increased reliance on intronic SREs in plant splicing, and a greater reliance on SREs in mammalian splicing.

INTRODUCTION

The removal of intronic sequences from pre-mRNA transcripts, splicing, is a key step in transcript maturation. Catalyzed by the spliceosome, splicing is widespread in eukaryotic organisms and essential for expression of many genes (*1*). The 5′ and 3′ splice site (SS) motifs and the branch point sequence (BPS) form the core sequence elements required for splicing. These motifs are recognized by components of the spliceosome in a process that pairs the 5′ and 3′SS to define the intron between them (*2*). However, these motifs do not, by themselves, contain sufficient information to fully explain the splicing patterns observed in many organisms (*3*). Instead, splicing is additionally affected by diverse splicing regulatory elements (SREs), which are recognized by a wide array of protein splicing factors (SFs), many deeply conserved in evolution (*4*).

Large-scale cell-based screening has been used to identify sequences that affect exon inclusion, intron inclusion, or splice site usage (*5–8*), with most of these studies defining specific sets of SREs. Other studies have used computational approaches followed by minigene validation experiments to identify exonic SREs (*9, 10*), and a great deal of mutational analysis has identified SREs active in specific exons or introns (*11–13*). Likewise, dozens of SFs have been extensively studied biochemically and genetically (*14*). In vitro binding preferences of RNA-binding proteins (RBPs), including many SFs, have been mapped (*15, 16*), and tools for modeling binding have been developed (*17, 18*). Recently, in vivo binding and splicing activity following RNA interference knockdown have been assessed for dozens of SFs (*19*). SS motifs have also been explored experimentally as well, including a recent large-scale screen that measured the activity of all possible 5′SS sequences in three different exonic contexts (*20*). Various computational models focused on determining SS strength and classifying short sequences as 5′ or 3′SSs were among the earliest models related to splicing (*21–23*).

Over the past two decades, more comprehensive models of splicing have been developed, predominantly focused on mammalian systems. An early approach showed that adding known exonic splicing enhancer (ESE) and especially exonic splicing silencer (ESS) elements to SS motifs improved the accuracy of SS prediction (*5*). Other models of splicing have focused more on the prediction of splice-altering mutations or the percent spliced in (PSI) for cassette (alternatively spliced) exons. These models use a variety of features to make their predictions, for instance, the hexamer additive linear (HAL) model predicts the change in PSI for a cassette exon following mutation based on the hexamer compositions of the wild type and variant sequences (*8*). There have also been efforts to define a comprehensive "splicing code" of relevant cis-acting elements, with more than 1000 features, including exon lengths and binding motifs for known SFs (*24–26*). These features have then been used in Bayesian neural network models to predict splicing disruption caused by genetic variants and relative PSI values for cassette alternative exons in different tissues. Models have also been developed that more explicitly seek to determine the likelihood of a variant disrupting splicing (*27, 28*).

More recently, extremely high predictive performance for splicing has been achieved via the application of deep neural network methods such as SpliceAI (*29*), which predicts splice sites from large segments of flanking human genomic sequence (up to 10 kb). These methods produce "black box" models, which learn parameters that are essentially uninterpretable. In silico mutagenesis experiments can be used to interrogate which sequence regions are important for specific predictions (*29*). A mutational scan of hundreds of regions identified the expected reliance on core SS motifs but failed to detect enrichment for known SRE motifs, suggesting that the genomic features learned by the model overlap only partially with those used in SS recognition by the spliceosome (*30*).

"White box" models, in contrast to black box models, are designed to be readily interpretable, allowing users to understand how the inputs and parameters are used to reach the conclusions of the model. Therefore, we developed a white box model where the structure and parameters are directly inspired by different aspects of the splicing process. The resulting interpretability allowed us to assess the relative importance of "structural" features such as exon and intron lengths and to derive scores for short sequences as SREs that capture many known regulatory motifs and identify other motifs that are predictive of function.

To this end, we assumed that recognition by the spliceosome primarily involves three key aspects: motifs of variable strength at the 5′

[1]Computational and Systems Biology PhD Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.
*Corresponding author. Email: cburge@mit.edu

and 3′SS; SREs located near the SS; and SS pairing, including distance constraints and preferences. New models of SS strength incorporating triplet preferences in a maximum entropy framework were found to improve on previous non-neural models (*22*). Here, SREs were assumed to act locally (e.g., within 80 or 100 nt of the SS) and additively to enhance or suppress usage of nearby SS. Combining these first two features yielded a simple model we call context-aware SS (CASS), which may have applications as a measure of SS strength in local context. Additionally, we assumed that pairing of SSs by the spliceosome involves not only steric considerations that enforce a minimum intron length but also preferences for specific exon and intron lengths that reflect processes of exon and intron definition (*31*). These three features were modeled together in our SMsplice model, whose structure is closely related to a hidden semi-Markov model (HSMM), enabling use of a classic HSMM algorithm to identify the most likely splicing pattern in a transcript. To maintain tractability, other features known to play a role in the splicing of some or all introns were omitted, including the BPS motif, recursive splicing, RNA secondary structure of the pre-mRNA, and impacts of SREs at longer distances (*32*). Despite these omissions, our model yielded moderately to highly accurate predictions in a variety of animals and a model plant and yielded putative SRE motifs in each organism and insights into the relative importance of different features across evolution.

RESULTS

CASS scores more than double the performance of MaxEnt scores alone

We sought to develop scores representing the potential of individual sequence positions to function as a 5′ or 3′SS, considering first the core SS motifs, and then the impact of nearby SREs (Fig. 1A). For SS motif scores, we developed an updated version of the MaxEntScan method (*22*), taking advantage of much larger training sets of high-quality splice sites and improved computational resources to develop more complex and accurate models. Our new version is a "third-order" model as it captures dependencies between triplets of positions rather than just the pairs of positions in the core motifs considered previously (Methods). We defined core motif regions as in the original model, consisting of 9 nt at the 5′SS (−3 to +6), and 23 nt at the 3′SS (−20 to +3, including the polypyrimidine tract). The models return a log-odds score of a given sequence of length 9 or 23 based on the ratio of the probability of the sequence as a 5′SS or 3′SS divided by the probability of the sequence under a background model. This new version of the method yielded improved discrimination over the original, with a test set area under the receiver operator characteristic curve (AUC) of 0.9982 compared to the original's 0.9971 for 5′SS and 0.9965 versus the original's 0.9960 for the 3′SS model (Fig. 1B), and larger improvements in other organisms (below). To use these models as SS predictors, we simply set a cutoff on the SS scores and predicted as a SS any position with score exceeding a threshold that maximized the F_1 score, defined as the harmonic mean of precision and recall, on a training set of human genes (Methods). Application to a short human gene, *ZNF575*, is shown (Fig. 1C).

The activities of splice site motifs in human genes are known to be quite sensitive to the local sequence context, often being influenced by nearby SREs (*1*), which can act from exonic or intronic locations. Exonic SREs are typically designated ESEs when they promote the use of upstream 3′SSs/downstream 5′SSs and/or inclusion of the exon in which they reside and as ESSs when they have the opposite

effect on splicing. Similarly, intronic SREs that promote the use of downstream 3′SSs or upstream 5′SSs are known as intronic splicing enhancers (ISEs), while those that have the opposite effect are intronic splicing silencers (ISSs). To represent the behavior of SREs enhancing or silencing splicing in local regions, we defined an exonic splicing regulatory element, or "ESR" score, for each possible hexanucleotide (hexamer), representing its impact on the splicing of upstream 3′SSs and downstream 5′SSs, as well as an intronic splicing regulatory element, or "ISR" score, representing its impact on downstream 3′SSs and upstream 5′SSs. In essence, we treat hexamers with positive or negative ESR scores as akin to ESEs and ESSs, respectively, and those with positive or negative ISR scores as ISEs or ISSs. These scores, determined by a learning procedure describe below, are added to the "core SS scores" (from the MaxEnt procedure described above) of adjacent SS motifs as shown (Fig. 1A). We refer to these SRE-modulated SS scores as CASS scores.

Mathematically, the CASS scores are defined by the formulas

$$s_t^5 = m_t^5 + \sum_{j=t-9-r}^{t-9} \sigma_{h(j)}^e + \sum_{j=t+6}^{t+6+r} \sigma_{h(j)}^i$$

$$s_t^3 = m_t^3 + \sum_{j=t+4}^{t+4+r} \sigma_{h(j)}^e + \sum_{j=t-25-r}^{t-25} \sigma_{h(j)}^i$$

Here, *t* refers to the position of the base in question within the wider sequence and the superscripts 3 and 5 differentiate between the 3′ and 5′SS. So, s_t^5 refers to the 5′ CASS score at position *t* of a sequence and s_t^3 refers to the 3′ CASS score at that same position. The first terms in these equations, the *m*, are the core SS scores, with similar notation as the context-aware scores. For the summation terms, *r* indicates the range of sequence context considered, and the bounds of the summation are chosen to avoid overlapping the sequences used to determine the SS scores. We set *r* equal to 80, which proved optimal or near optimal in a variety of tests in all organisms studied. The summands are the relevant SRE scores, the superscripts *e* and *i* differentiate between exonic and intronic context, and the subscript function *h*(*j*) is a hash function, indicating the index used for σ^e or σ^i to get the value associated with the hexamer beginning at position *j* in the sequence.

ESR and ISR scores were learned as follows. Starting with the predictions made using our core SS motif models, we considered all positions that disagreed with the canonical annotation for the associated gene. We reasoned that these false positives (FPs) and false negatives (FNs) represented sequences where SREs were likely involved. An FP represents a strong SS motif, which might require silencing elements in the flanking sequence, whereas an FN represents a weaker SS motif that might require splicing enhancers nearby to promote its use. In particular, we took the flanking regions that would be incorporated into the CASS scores, separating the exonic and intronic sides, and updated the SRE scores, encouraging the hexamers frequently present near FPs to be silencing and those frequently near FNs to be enhancing (Fig. 1D and Methods). We then used these updated SRE scores to make new predictions on the training set and repeated this process until performance on the validation set peaked (Methods).

We refer to SRE scores calculated in this manner as CASS-learned SRE scores. Generally speaking, the ESR scores had greater magnitudes, both positive and negative, than the ISR scores (Fig. 1E). Furthermore, we were able to quantify the benefit provided by the context in CASS scores. We applied the CASS framework to a test set of genes and, using the prediction cutoffs learned from the training set, calculated the F_1 scores (Fig. 1F). The ensuing change in performance was
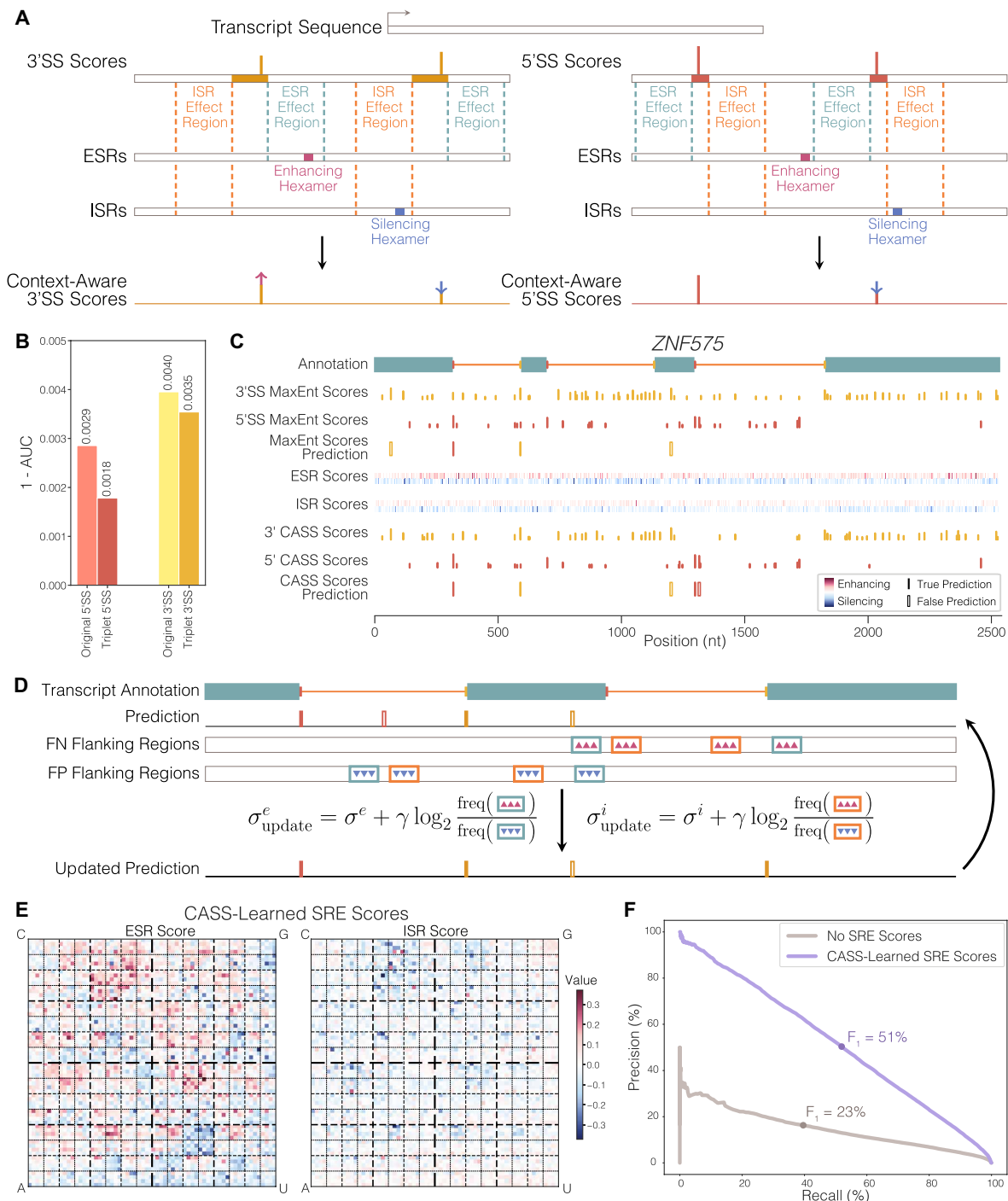
Fig. 1. CASS scores improve predictions. (**A**) CASS scores for an RNA sequence are determined by scoring every position as a potential 5′SS, 3′SS, ESR, and ISR. The 3′ CASS score is the sum of its original 3′SS score plus ISR scores of hexamers in the upstream ISR effect region and ESR scores of hexamers in the downstream ESR effect region. The 5′ CASS score is calculated similarly but with upstream ESR scores and downstream ISR scores. (**B**) Reduced $1 - \mathrm{AUC}$ values for the updated MaxEnt methods compared with the originals show improvement in classification of SS versus non-SS sequences. (**C**) Illustration of MaxEnt scores, SRE scores, CASS scores, predictions, and annotations for the gene *ZNF575*. 5′ and 3′SS scores above zero are shown, with height indicating strength. For ESR and ISR scores, positive scores are above (pink), and negative scores are below (blue), with color intensity indicating strength. Solid/hollow bars represent correct/incorrect predictions, respectively. (**D**) Iterative learning of scores uses the frequency of hexamers in regions flanking FP and FN predictions. SRE scores of hexamers that appear frequently in exonic and intronic FP regions compared to the respective FN regions are decreased, while scores of those with the opposite pattern of enrichment are increased. (**E**) Chaos plots for CASS-learned SRE scores. For each hexamer, the overall quadrant in the heatmap indicates the first base, indicated at the corners. The second letter determines the subquadrant and so on. The score for each hexamer is indicated by color. (**F**) Precision-recall curves on the test set for predictions made using no SRE scores and the CASS-learned SRE scores. The dots represent the location on the curve determined from the cutoff learned on the training data, and the associated F_1 score on the test data is indicated.

substantial, increasing from 23% with SS scores alone to 51% with CASS-learned SREs. Thus, more than half of the performance of the CASS model in human is attributable to the local SRE context of splice sites.

SMsplice captures structural and sequence features of splicing and improves over CASS

CASS scores represent SS strength in local context but do not account for the pairing of SSs across introns and/or exons that occurs in splicing. For instance, the CASS scores could predict a 3′SS upstream of the first plausible 5′SS in the gene (e.g., Fig. 1C), although such a site could not participate in conventional cis-splicing. Furthermore, there are known minimum intron lengths in human and other organisms below which splicing is not observed (33), and both exon and intron lengths can affect recognition by the spliceosome, often via exon or intron definition (31). We therefore sought to make a more general model that could incorporate both SS pairing and exon/intron length preferences, as well as the preferred numbers of introns per transcript. To this end, we developed a directed graphic structure that could describe the splicing pattern of an arbitrary transcript in terms of exons, introns, and SSs beginning from the start of the transcript (Fig. 2A).

This structure reflects the observation that the number of introns per human gene fits approximately a geometric (exponential) distribution in the training set, with parameter $p_{EO} = 0.093$ (Fig. 2B). We additionally characterized the length distributions of introns and different types of exons in human genes, which may partly reflect preferences involved in spliceosome assembly (Fig. 2C) (31). To model these length distributions, we smoothed the empirical length distributions on the training set for each type, then substituted a geometric tail for each distribution to avoid issues with data sparsity at long lengths (Methods).

Our SMsplice model combines the above "structural" features (relating to exon/intron order/size), which we call the SMsplice structure, with the CASS scores. This model assumes that the exons and introns in a transcript are chosen in proportion to: (i) the strengths of the involved SSs in local context, represented by (exponentiated versions of) the CASS scores; (ii) the exon and intron length preferences represented in Fig. 2C; and (iii) the preferred number of introns represented by a geometric distribution (with parameter from Fig. 2B). Together, these three features define the score for any "parse" (splicing pattern) π for the sequence of interest ("seq"). If seq is of length T, then an arbitrary parse π has $N > 0$ introns of lengths $d_1^I, \ldots, d_N^I$, a first exon with length d_F^E, $N - 1$ internal exons with lengths $d_1^E, \ldots, d_{N-1}^E$, and a last exon with length d_L^E (with the lengths of all exons and introns summing to T). The SMsplice score for this parse, SM[seq, π], is then defined as

$$\begin{aligned}
SM[seq, \pi] = &\sum_{t \text{ where } \pi \text{ has a } 5'SS} s_t^5 + \sum_{t \text{ where } \pi \text{ has a } 3'SS} s_t^3 \\
&+ \log_2\left[p_{ELF}(d_F^E) \cdot p_{ELL}(d_L^E) \cdot p_{IL}(d_N^I)\right] + \sum_{n=1}^{N-1} \log_2\left[p_{ELM}(d_n^E) \cdot p_{IL}(d_n^I)\right] \\
&+ \log_2\left[(p_{EO})(1 - p_{EO})^{N-1}\right]
\end{aligned}$$

Here, the first line represents the strengths of the SSs (in context, i.e., using CASS-style scores), the middle line represents length preferences for the involved exons and introns, and the last line represents the preference for the specific number of introns. The predicted splicing

pattern for the sequence, π^*, is the one that has the highest SMsplice score of all possible parses, which can be obtained by adapting the Viterbi algorithm for HSMMs (Methods) (34). While SMsplice is similar in structure to an HSMM, it is a discriminative rather than generative model, meaning that it discriminates different parses but cannot generate RNA sequences. Therefore, it is technically a semi-Markov conditional random field rather than an HSMM (35, 36).

While we can use the previously determined CASS scores with SMsplice, we can also learn an additional set of SRE scores using SMsplice predictions. As before, we began by setting all SRE scores to zero but now used the FP and FN from SMsplice predictions to update the scores rather than the CASS predictions (Fig. 2D). We refer to the resulting SRE scores as SMsplice-learned. Comparing these scores with the CASS-learned SRE scores, we observed fairly strong correlation between the scores learned by both methods, more so for the ESR scores than the ISR scores (Fig. 2E).

We then explored the predictions made using either set of SRE scores within either the CASS framework or SMsplice on a test set of genes (Fig. 2F). This analysis showed that F_1 performance improved with the addition of the SMsplice structure, regardless of the SRE scores considered, up to an increase of 21% over the CASS predictions. Furthermore, SMsplice had the highest performance with the SMsplice-learned SRE scores at 69%, substantially higher than with the CASS-learned scores. This improvement may result partly from the ability of SMsplice to exclude potential SSs that are structurally incompatible with splicing; for example, a predicted 3′SS that occurs before the first putative 5′SS in a transcript, which is considered by the CASS model (Fig. 1C). In addition, comparing SMsplice and CASS predictions, SMsplice is more apt to predict SS pairs with typical exon/intron spacings and less apt to predict lone, high-scoring SS in introns. Therefore, the false predictions considered during SMsplice learning may be more enriched for regions considered by the splicing machinery whose exclusion is modulated by nearby SREs.

We further defined a "local score" (LS) for any intron or internal exon based on the associated terms in the SM[seq, π] expression. For an internal exon whose flanking 3′ and 5′SS occur at $t1$ and $t2$, respectively, resulting in an exon of length $d^E = t2 + 1 - t1$, the LS is

$$s_{t1}^3 + s_{t2}^5 + \log_2\left[p_{ELM}(d^E)\right]$$

For an intron whose 5′ and 3′SS occur at $t1$ and $t2$, respectively, resulting in an intron of length d^I, the LS is

$$s_{t1}^5 + s_{t2}^3 + \log_2\left[p_{IL}(d^I)\right]$$

Applying these scores to predicted exons and introns for the human test set, we observed that exons with larger local scores were more likely to be predicted correctly, with a fairly smooth and nearly monotonic relationship, and similarly for introns (fig. S1A).

Application of SMsplice to other animals and a plant

Splicing is thought to have been present in the last common ancestor (LCA) of eukaryotes (1), and the assumptions about splicing underlying the CASS and SMsplice models are reasonable for a wide range of species. Furthermore, many protein families known to modulate splicing via the binding of SREs are conserved between animals and plants, although some have been lost, especially in particular fungal lineages (4). Therefore, it was of interest to explore the application of our models to genes from diverse organisms. To capture a range of evolutionary distances and address important model organisms, we selected mouse, zebrafish, fruit fly (*Drosophila melanogaster*), silkworm moth (*Bombyx mori*),
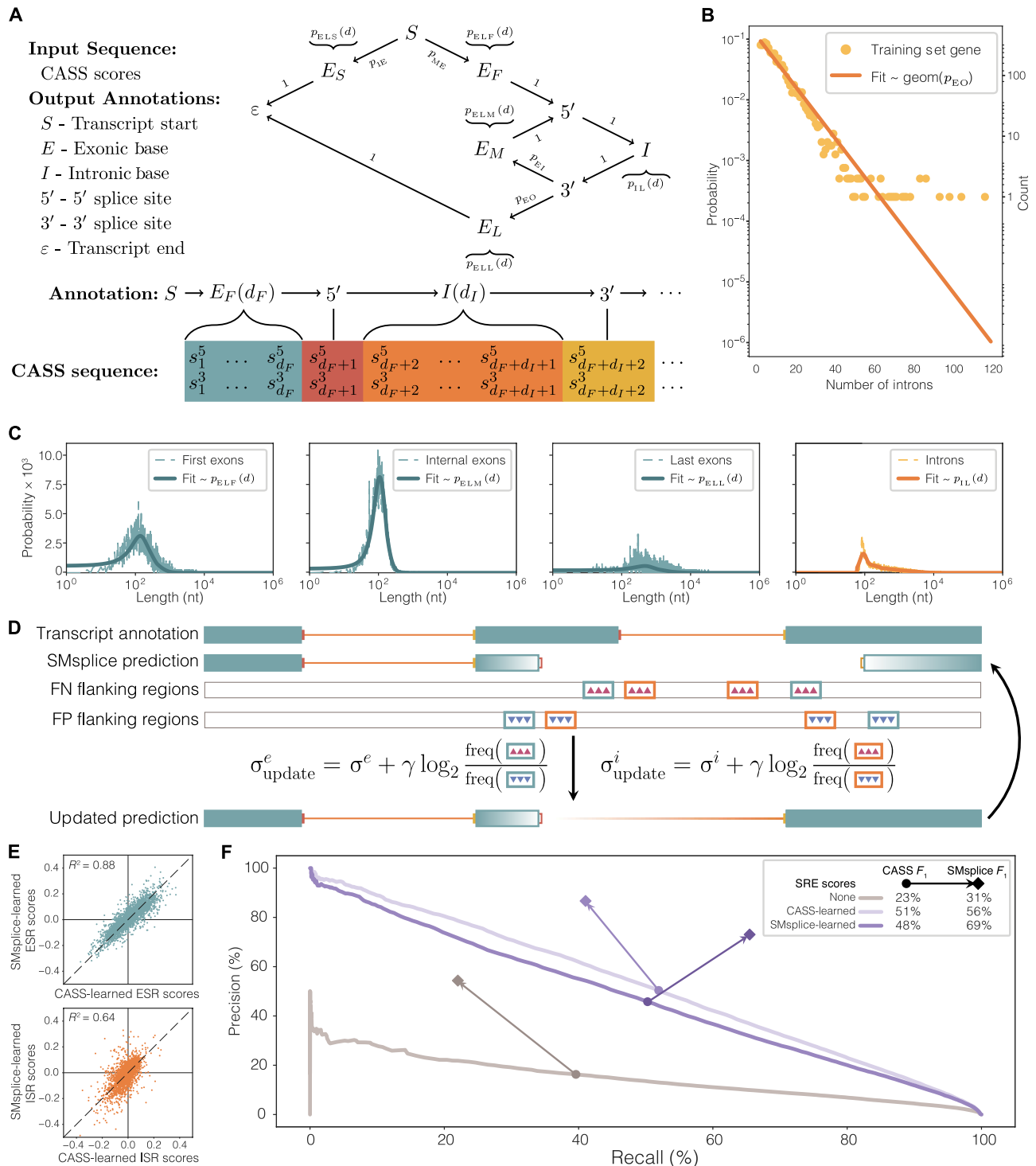
Fig. 2. SMsplice structure, parameters and performance in human. (**A**) In the upper diagram, the labeled arrows represent the possible transitions with their associated probabilities, and the curly braces represent the association between states and their length distributions. Below, a possible parse of a sequence of CASS scores is shown, which would be determined from some RNA sequence of interest. (**B**) Geometric fit to the empirical counts of genes with a certain number of introns in the training set. (**C**) Empirical length distributions for the introns and first, internal (or middle), and last exons in the training set of genes along with the smoothed distributions used in the SMsplice model. (**D**) Iterative learning of scores is performed as before but using SMsplice predictions to define the FP and FN flanking regions. (**E**) SMsplice-learned ESR scores correlate better with the CASS-learned ESR scores than their ISR counterparts. (**F**) Precision-recall curves on the test set for the CASS framework predictions using different sets of SRE scores. The dots along the curves represent the cutoff learned on the training data to maximize the F_1 score and are connected by an arrow to a diamond marking the precision and recall for predictions made using the SMsplice model with the same SRE scores.

and the model plant *Arabidopsis thaliana*. We devised separate training, validation, and test sets for each organism analogous to those used for human (Methods).

As the SS motifs differ somewhat between these organisms, we trained a new organism-specific third-order MaxEnt model for each. These models were trained on all of the genes not in the organism's test set using the third-order constraints as was done with the updated human MaxEnt models (Methods). We then compared the classification performance of these organism-specific MaxEnt models to that of the original (human-trained) second-order MaxEnt model to assess the change in SS classification performance. We found that these new models provided improved discrimination compared with the original MaxEnt models for all six organisms, especially in *Arabidopsis* (Fig. 3A).

Using these new splice site models, we next learned organism-specific SREs using the CASS-learning and SMsplice learning approaches. We also determined structural parameters for the SMsplice model, including exon and intron length distributions, for each organism (Methods, Fig. 3B, and fig. S2) (*3*). Modeling exon and intron lengths were of particular interest as the length distributions for exons and introns are related to the prevalence of intron definition versus exon definition, with the latter being especially prevalent in mammals (*1*, *31*).

When we applied these newly determined parameters to the task of predicting splicing on test sets from the respective organisms, we observed several trends (Fig. 3C). In every organism, the use of CASS- or SMsplice-learned SRE scores improved performance substantially over SS scores alone. The SMsplice structure further improved performance, with best overall performance observed with SMsplice-learned SRE scores for every organism. While the highest F_1 value for mouse (72%) was similar to the 69% seen in human, accuracy in the four other organisms was substantially higher (83 to 85%). Some variation was seen in the relative performance of the different score/framework combinations, with a larger gap between SMsplice- and CASS-learned SRE scores for mammals than other organisms, perhaps reflecting the longer distances involved in mammalian splicing. The local scores of exons and introns in these organisms showed the same positive association with accuracy as was observed for human (fig. S1), indicating that this score can be used to distinguish predictions of higher and lower confidence. Thus, our framework and learning approach generalize well to other animal and plant species, enabling a variety of comparative investigations.

We also assessed performance at the level of individual genes, again using the F_1 measure. For most organisms, the median F_1 accuracy for individual genes in the test set was close to or slightly above the overall F_1 value (fig. S3). The exception was fly, whose median test gene had an F_1 of 95%, reflecting that >49% of the test genes were predicted with perfect accuracy. The proportion of perfectly predicted genes in other organisms varied from 8% (human) to 33% (*Arabidopsis*) (Fig. 3D). A representative splicing prediction for each organism—with F_1 value within 1% of the median value across genes—is visualized in fig. S4. Examination of these visualizations reveal several common features of splicing, including the high density of decoy SS, the somewhat lower density of locations with high CASS scores, and a complex landscape of SRE scores. In *Arabidopsis*, but not other organisms, there was often a visually apparent shift in ISR and/or ESR scores at or very near exon/intron and intron/exon boundaries, often persisting throughout the succeeding exon or intron (fig. S4).

Splicing in organisms with longer introns is more dependent on SREs

One of our fundamental goals was to understand the relative contributions of different features to SS recognition. The CASS framework was designed to mirror the concept of SRE regulation of SS motifs, and these two components can be separated into the SRE scores and MaxEnt scores, respectively. Then, the SMsplice structure additionally includes the structural constraints of the spliceosome. So, by examining their effects within the model, we might gain some insight into how important these different types of splicing information are in these organisms. To do so, we considered the change in performance provided by the use of the SMsplice structure in the absence of SREs and then the further benefit that adding the SMsplice-learned SRE scores granted (Fig. 3E). From this analysis, we noted that there was substantial variety in the proportions across these organisms. In human and mouse, for instance, most of the performance comes from the SRE scores, while *Arabidopsis*, with the shortest mean intron length, had the most reliance on core SS motif scores. The largest structural contribution was seen in fly, which has a large proportion of introns within a very narrow length range of 40 to 80 nt (*37*).

Intron length varied substantially more than exon length across the organisms considered and determined to a large degree the number of decoy SS per transcript (*5*, *38*). Noting this, we explored the relationship among intron length, performance, and the contributions of different features. In general, SMsplice performance declined as average intron length increased (Fig. 3F). However, the decline was much shallower when SREs were included in the model. Consistent with this observation, we found that the relative contribution of the SS motifs to performance decreased with average intron length, while the relative contribution of SREs increased markedly (Fig. 3G). These trends suggest that, in lineages where introns lengthen, dependence on the presence of SREs becomes stronger.

Learned SRE scores distinguish real and decoy SS

To further explore the idea of distinguishing real and decoy SSs via SRE regulation within the model, we began by determining a suitable set of decoy SSs. To create such a set, for each SS in the canonical training set that flanked an internal exon, we selected a non-SS base whose associated SS score was within half a bit of the score of the true SS (Methods). Thus, we were able to create a decoy set of similar size and score distribution to the set of real SS for each organism. We considered the distribution of SMsplice-learned SRE scores relative to the selected decoy 5′ and 3′SSs, as well as real 5′ and 3′SSs. In our models, 5′SS scores are affected by upstream ESRs and downstream ISRs, and 3′SS scores are affected by upstream ISRs and downstream ESRs. For each position in the flanking regions, we considered the average ESR or ISR score (Fig. 4A). Dividing the ESRs into ESEs (if score > 0) or ESSs (if score < 0) and similarly dividing ISRs into ISEs and ISSs, we separately tallied the averages of each of these categories of SREs.

In human, the average positional ESR and ISR scores flanking real SSs were relatively smooth, as were the average ESE, ESS, ISE, and ISS values, suggesting that SRE information is, on average, fairly evenly distributed (Fig. 4A). We did observe a modest increase in ESE scores at both exon boundaries, as has been observed previously (*39*). Similar slightly sloping ESE distributions were observed in mouse, zebrafish, and *Arabidopsis* (fig. S5A). ESS scores were also fairly smooth, with slightly increased magnitude further from both 5′ and 3′SSs. The average SRE scores flanking decoy SSs were also fairly smooth, with
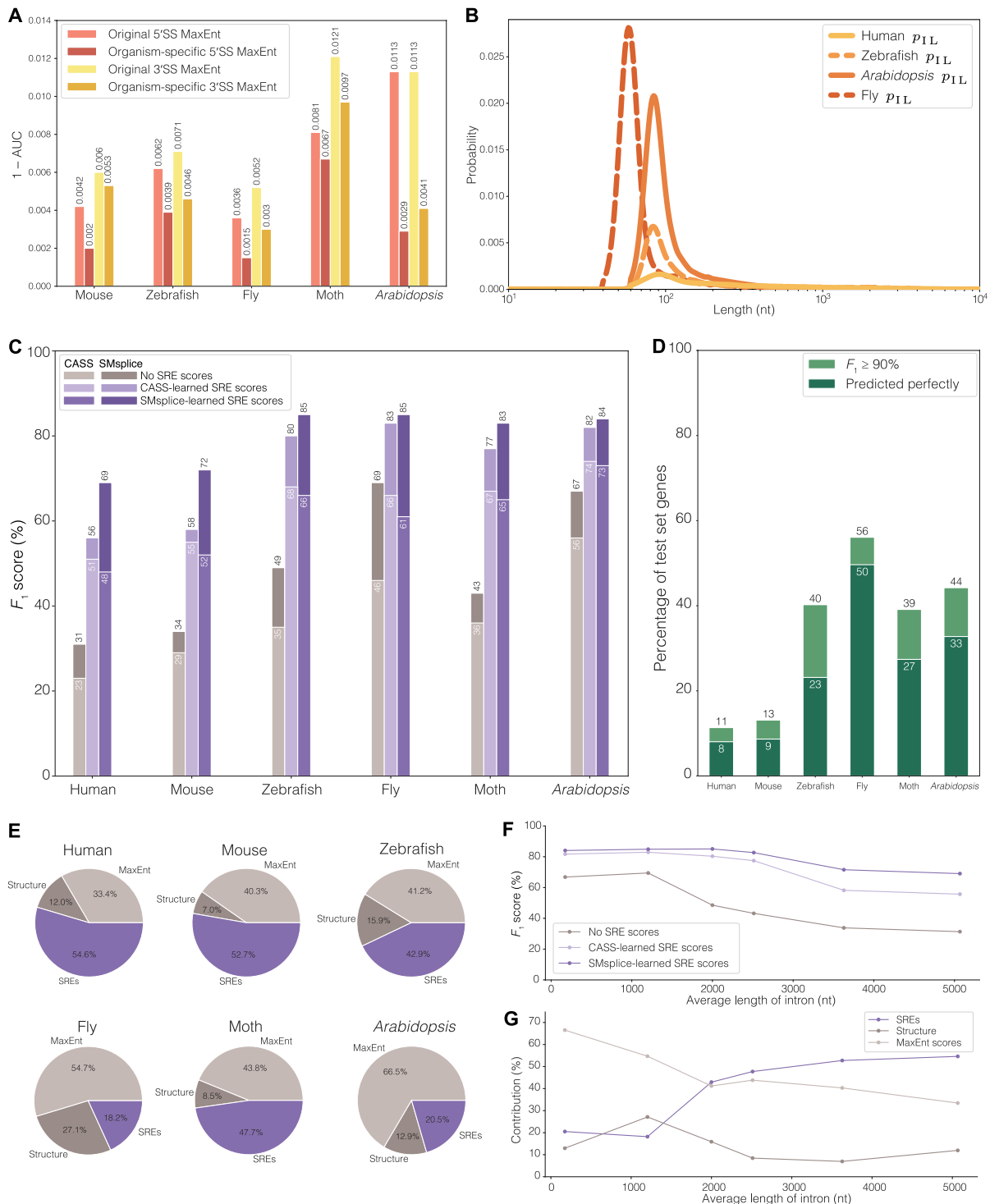
Fig. 3. Application of SMsplice to six organisms show varied contributions of different features. (**A**) Smaller $1 - \text{AUC}$ values for the updated, organism-specific MaxEnt methods compared with the original human version show their improvement on classifying SS versus non-SS sequences. (**B**) Intron length distribution fits used within the SMsplice model for human, zebrafish, *Arabidopsis*, and fly. (**C**) Test set F_1 performance for each organism with no SRE scores, CASS-learned SRE scores, and SMsplice-learned SRE scores within both the CASS framework and SMsplice model. (**D**) Proportions of the test set predicted perfectly and with F_1 score at least 90% for each organism. (**E**) Pie chart breakdowns of the contributions to F_1 performance of the organism-specific MaxEnt scores, the SMsplice structure, and the SMsplice-learned SRE scores. (**F**) F_1 performance of the SMsplice model for each organism with no SRE scores, CASS-learned SRE scores, and SMsplice-learned SRE scores as a function of the average length of introns in the test set. (**G**) Contributions to F_1 performance of the organism-specific MaxEntScan scores, the SMsplice structure, and the SMsplice-learned SRE scores as a function of the average length of introns in the test set.
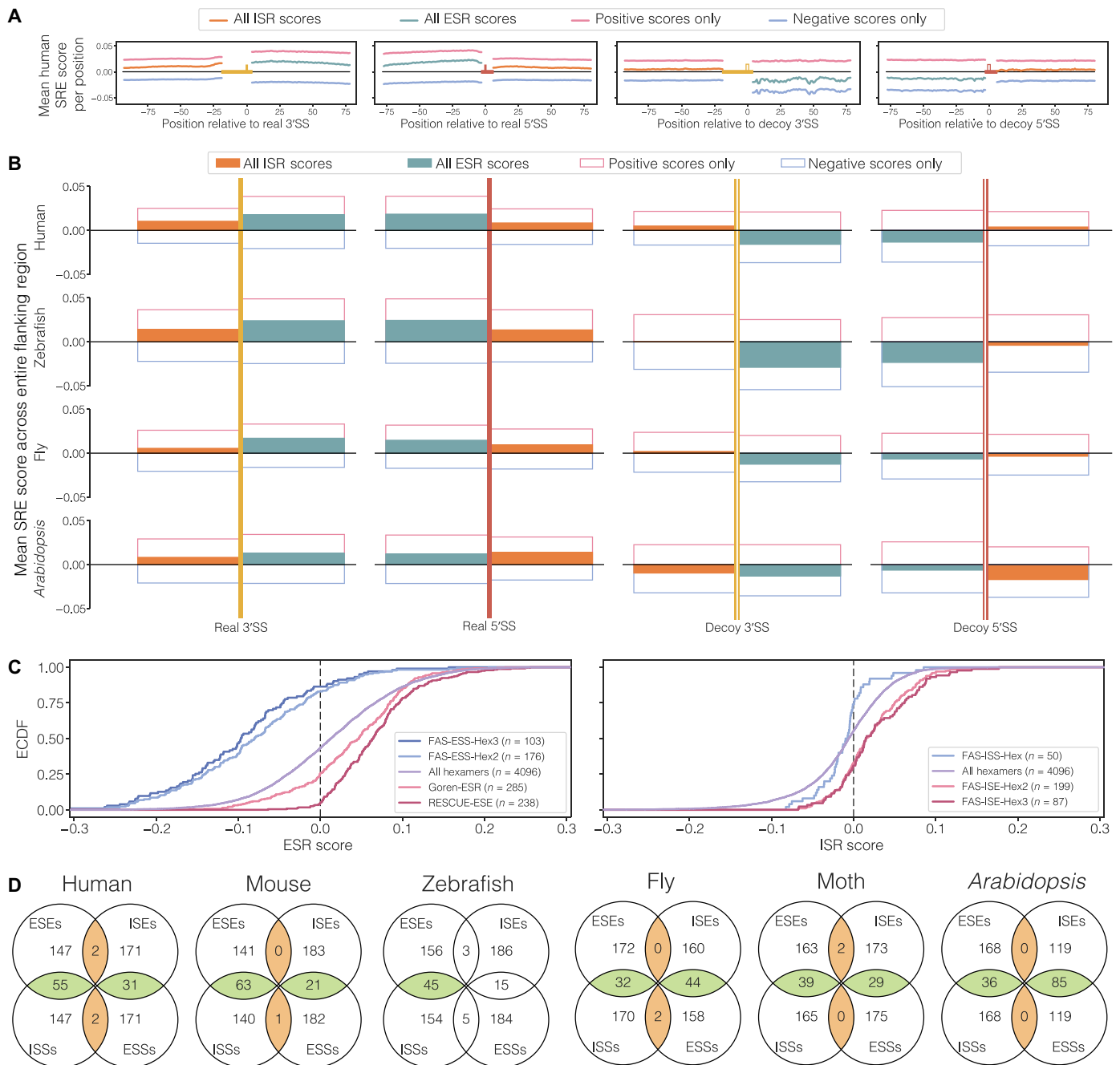
Fig. 4. SMSplice-learned SRE scores have features associated with known SREs. (A) Relevant SMsplice-learned ESR and ISR scores, along with their positive and negative components, for each position in the flanking regions, averaged separately across all real and decoy 5′ and 3′SSs in the human training set. **(B)** For human, the average of the scores shown in 4A across the associated regions. For zebrafish, fly, and *Arabidopsis*, the equivalent values on their training sets. **(C)** Plots of the ECDFs for the human SMsplice-learned ESR and ISR values of all hexamers along with subsets of hexamers reported in the literature as particular types of splicing regulators in human. **(D)** Agreement between the top and bottom 5% hexamers based on their SMsplice-learned ESR and ISR scores. Overlaps that are significantly underrepresented are colored in brown, while overlaps that are significantly overrepresented are colored in green.

the exception of ESSs near decoy 3′SSs, which showed a visually distinctive pattern of peaks and valleys, with a less pronounced pattern observed near decoy 5′SSs. Suspecting that a common repetitive element might be involved, we observed that 18% of decoy 3′SS and 14% of decoy 5′SS overlapped a short interspersed nuclear element (SINE) of the Alu class, often at particular locations in the Alu consensus (*1*).

Examining the SRE distributions in the Alu versus non-Alu decoys, we found that the decoy SSs in Alus had a more exaggerated version of the overall pattern of peaks/valleys, indicating that Alu decoys are driving this pattern (fig. S5B). In some other organisms (e.g., mouse and moth), similar bumpiness in ESS distributions was observed (fig. S5A), likely for similar reasons.

To summarize the impacts of different SRE score categories on SS prediction, we also considered the average of the relevant scores for the entire flanking regions (Fig. 4B and fig. S5C). For every organism, there was positive net score for real versus decoy 5′ and 3′SSs in both the exonic and intronic flanking regions, as expected. Furthermore, the average ESR score was of greater magnitude than the average ISR score for real 5′ and 3′SSs in every case, except the 5′SS of *Arabidopsis*. For the decoys, we observed a negative average ESR score on the exonic side for both SS types in all organisms. ISR scores on the intronic side of decoy SSs were more variable in sign, with smaller magnitude than the corresponding ESR scores in all cases, except for the decoy 5′SS in *Arabidopsis*.

The patterns we observed for the SRE scores in regions flanking decoy SSs suggested that these regions might be useful in SRE score learning. To explore this idea, we scored each hexamer according to its frequencies of appearance in the CASS-relevant regions flanking the real and decoy SSs in the training set and found that these "real versus decoy" scores correlated strongly with the SMsplice-learned SRE scores [R^2 (coefficient of determination) values between 0.34 and 0.79; fig. S6A and Methods]. Furthermore, when appropriately weighted, these real versus decoy scores provided useful seeds to SMsplice learning, providing a small but consistent improvement in validation performance of ~1% over no seeding (fig. S6B and Methods). Going forward, we used the scores learned from these seeds as our SMsplice-learned SRE scores. As observed for the CASS scores, the new ESR scores tended to have greater magnitudes than the ISR scores in human (Fig. 1E and fig. S6C). For the non-mammals, ESE scores tended to exceed ISE scores, but ISSs often had greater average magnitudes than ESSs.

To ask how the SRE scores learned by our SMsplice model relate to known sets of SREs, we gathered several groups of hexamers determined to function as ESSs, ESEs, ISSs, and ISEs using splicing reporter–based screens or validations (5–7, 9, 10). We compared the empirical distributions of SRE scores associated with these different groups of hexamers to the relevant overall empirical distribution of SMsplice-learned SRE scores (Fig. 4C). All of the ESR groups were significantly different from the overall ESR score distribution [$P < 10^{-9}$, two-sample Kolmogorov-Smirnov (KS) test], and the ISR groups were also significantly different from the overall ISR score distribution ($P < 0.01$, two-sample KS test). Additionally, RESCUE-ESE hexamers predominantly had positive ESR scores (226 of 238, 95%), and Goren-ESR hexamers were also strongly enriched for positive scores (both $P < 10^{-9}$, one-sided binomial test). Furthermore, both sets of FAS-ESS hexamers were strongly overrepresented for negative ESR scores (89 of 103 or 86% for the FAS-hex3 set) ($P < 10^{-18}$ for both sets, one-sided binomial test). These trends were also observed for intronic elements, with FAS-ISEs overrepresented for hexamers with positive ISR scores and FAS-ISSs overrepresented for negative ISR scores ($P < 0.01$, one-sided binomial test). These observations support that positive/negative SRE scores learned by our model are predictive of splicing enhancing/silencing activity, respectively.

It has been previously observed that human SREs have a pattern of agreement where ESEs and ISSs tend to overlap as do ISEs and ESSs, likely reflecting common mechanisms of the associated RNA-binding SFs (6, 7). To explore whether our SMsplice-learned SRE scores recapitulated this pattern, we considered the hexamers with the most extreme ESR and ISR scores: The top 5% scoring hexamers we considered our enhancers (ESEs and ISEs), and the bottom 5% scoring hexamers we considered our silencers (ESSs and ISSs). While the 5% threshold is arbitrary, it yields sets of 204 hexamers, similar in size to the SRE sets from the literature discussed above. Examining the hexamers in common between these four groups of 204 hexamers, we saw that there was indeed a great deal more overlap between the ESEs and ISSs, as well as the ISEs and ESSs than between the other types of overlap. Tests for over- and underrepresentation found that all of these overlaps were significant in the expected direction ($P < 0.05$, one-sided binomial test, Bonferroni corrected) aside from the comparisons involving zebrafish ISEs and ESSs (Fig. 4D). These observations provide further support that strongly scoring hexamers in our model have properties expected of the associated SRE class.

Clustering human hexamers recovers known splicing RBP motifs and additional splicing motifs

Known SREs predominantly function by recruitment of SFs that bind motifs of approximately 3 to 7 nt in length (15). To ask whether the SMsplice-learned SRE scores correspond to known SF motifs, we clustered hexamers from the sets of 5% most extreme positive and negative ESR and ISR scores considered above by sequence similarity and aligned them to create a position weight matrix (PWM) for each of the resulting clusters (figs. S7 to S10 and Methods). The ESS clusters were notably poor in cytosine, consistent with a previous cell-based screen for ESSs (5). For each of the clusters obtained from human SRE hexamers, we identified the best-matching motif from the set of RNAcompete in vitro RBP-binding motifs (15) using a simple permutation approach to assess which matches had significant similarity (Methods).

From this analysis, we identified at least one significantly similar RNAcompete match for the majority of human SRE clusters (Fig. 5, A and B, and fig. S11). The matching was particularly strong for the ESE and ESS clusters, where 15 of 20 had a significant RNAcompete match. For the ESE clusters, matches were made to motifs for known SFs including RBM45 and SRSF3, as well as other RBPs with homologs involved in splicing (PCBP4) or other roles in RNA metabolism (e.g., the PolyA Polymerase Star-PAP). For the ESS clusters, matches were made to several known SFs, including HNRNPA1L2, HNRNPA3, HNRNPDL, the HNRNPF/H family, the MSI family (whose motif resembles that of HNRNPA0), and QKI.

Thus, most of the ESR motifs obtained by this clustering matched known SFs motifs. ESS motifs primarily matched hnRNP motifs, most of which are known to repress splicing when bound to exonic locations, and two ESE clusters matched SR proteins, which are known to activate splicing from exonic locations (4). Many of the ISR clusters also matched known SF motifs, including ISS clusters that matched RBM42, SRSF3, and SRSF7 motifs, and an ISE cluster that matched the QKI motif (which also resembles BPS and SF1 motifs). For each cluster where at least one splicing regulator was a significant match, we further considered short hairpin RNA (shRNA) knockdown RNA-seq data from ENCODE, where available (19). Of the 13 clusters where such data were available, we asked whether the presence of that cluster in the relevant CASS region was significantly associated with changes in PSI values in the expected direction following knockdown (Methods). For 11 of the 13 clusters we found such an association for at least one RBP (table S1). In addition, where enhanced cross-linking and immunoprecipitation (eCLIP) data were available from ENCODE in the same cell line, we found that the presence of the cluster in the relevant CASS region was significantly associated with binding of the RBP in all but one case (table S1 and Methods). Together, these observations suggest that our score learning and hexamer clustering approach identifies
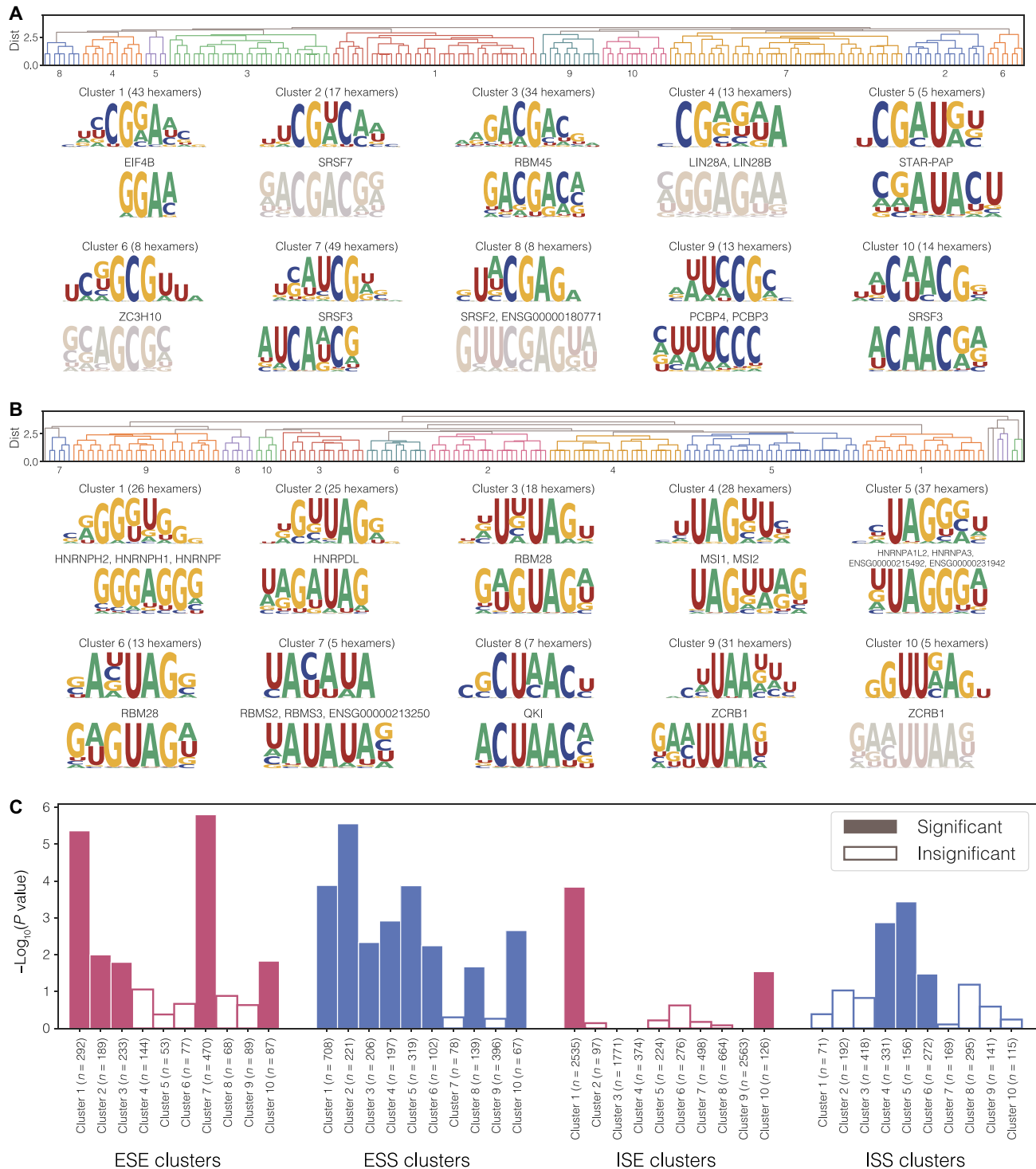
Fig. 5. Human SRE cluster motifs often match known RBP binding sites. (**A**) Dendrogram for clustering hexamers with the 5% most positive ESR scores in human and the resulting aligned clusters and best RNAcompete matches below those clusters. Desaturation indicates that the best match did not satisfy our threshold. (**B**) Dendrogram for clustering hexamers with the 5% most negative ESR scores in human, as in (A). (**C**) For each of the clusters made from the sets of the most extreme scoring hexamers, the support for that cluster being more associated with changes in splicing outcomes. The number of events available for each cluster is indicated, and whether the association is deemed significant.

authentic SRE motifs without prior knowledge of RBP binding. Clusters that fail to match RNAcompete motifs may represent known SFs that have not yet been characterized in vitro, motifs for unknown SFs, or artifacts of the algorithm or clustering process.

We next assessed the splicing regulatory activity of our clusters to see if the hexamers that we identified as high scoring were more associated with changes in splicing outcomes. To do this, we used fine-mapped splicing quantitative trait locus (sQTL) data from the Genotype-Tissue Expression (GTEx) database to ask whether variants that disrupted hexamers in each cluster were more likely to be causal relative to control hexamers (Methods). This analysis supported the regulatory activity of five ESE clusters, eight ESS clusters, two ISE clusters, and three ISS clusters (Fig. 5C). These supported clusters included ESS cluster 10 and ESE cluster 2, which did not have significant matches in the RNAcompete set. This observation suggests that these clusters represent bona fide SREs that function through other RBPs not analyzed by RNAcompete. The greater number of exonic clusters than intronic clusters that reached significance could reflect stronger activity of ESRs than ISRs or better modeling of ESR than ISR elements by our model. In addition, cross-referencing this fine-mapped data with other GTEx data allowed us to explore the relationship between variants and splicing events (Methods and table S2). Of the identified events, the SMsplice local scores were able to correctly predict the direction of change in 71% of cases.

Similar regulatory motifs identified across organisms allows for generalization of the SMsplice model

Key splicing proteins of the SR and hnRNP classes are conserved from animals to plants (4), suggesting that similar motifs may exert similar effects on splicing. However, intron-containing genes are often spliced differently when moved between mammals and fish (40) and potentially even more so across greater evolutionary distances. Examining the clusters for each SRE class in the non-human organisms (figs. S7 to S10), we observed many similar motifs within each class.

We repeated our motif comparison analysis for each examined organism, matching the clusters to RNAcompete RBP motifs associated with the relevant organism (figs. S12 to S16). These comparisons identified a number of RBP matches in each species, generally to very similar classes of proteins as observed in human, with several ESEs matching SR proteins, ESSs often matching hnRNPs or the MSI family, and QKI family proteins appearing as ISE matches. These observations suggest that clusters identified by our algorithm in other species often correspond to motifs bound by SFs, many of which are conserved.

To explore the extent of similarity of ESS clusters across organisms, we clustered all the ESS PWMs across species using the same distance measure used to compare our PWMs with the RNAcompete PWMs (Fig. 6A). This comparison identified several motifs, including UAG- and UAA-containing motifs, that were present in all six organisms, suggesting that these could represent ancient ESSs present in the LCA of animals and plants. Other motifs, including poly-G motifs, were present in all metazoans but were not observed in *Arabidopsis*. Some of the more CG-rich ESSs appeared more lineage specific. For instance, a CGCG motif only identified as an ESS in zebrafish resembled motifs identified as ISSs in other organisms (fig. S17).

We performed similar clustering analyses of ESEs, ISEs, and ISSs across the six organisms (fig. S17). For each class, we observed clear clusters, with some spanning all six organisms or the five metazoans.

For example, we observed a cluster of ESE motifs related to "CYACG" (Y = C or U) in all six organisms, a cluster of ISE motifs containing "UAAC" in five of six organisms (all but moth), and an ISS motif cluster containing CGG and/or GGA, also present in all organisms except moth. Other clusters were more lineage-restricted, including a poly-G ISE motif in vertebrates and moth, and a purine-rich ISS motif in the four non-mammals. These observations suggest the presence of both conserved and lineage-specific SRE motifs across the species considered.

As the motifs were derived from subsets of hexamers with the most extreme scores, we wondered whether the full sets of SMsplice-learned SRE scores would also exhibit conservation across species. To compare ESR and ISR conservation, we calculated the correlation of ESR and ISR scores as a function of evolutionary divergence time (Fig. 6B) (41).

We observed a positive correlation of ESR and ISR scores across all pairs of organisms, further supporting the notion that some SREs have maintained similar activity across >1.5 billion years of evolution. While very close organisms had more similar ESR scores, the similarity plateaued at a moderate correlation at evolutionary distances of ~400 Ma ago and beyond. Notably, ESRs tended to be somewhat more conserved than ISRs, consistent with the high conservation of ESS and ESE motifs observed above (39), and the fact that the protein-coding function of exons provides a strong constraint on their compositional drift that is not present in introns.

We were curious whether the SRE similarity we observed above was strong enough that one organism's SMsplice-learned SRE scores could be used effectively to predict splicing in the genes of another organism when paired with core SS motif scores and structural parameters of the other organism. Applying SMsplice in this manner with all possible combinations of SREs/genes (Methods), we observed that SREs from the non-mammalian species performed well on other non-mammals but poorly on the mammals. This finding may relate to our observations that mammals derive far more information from SREs than other organisms (Fig. 3D) and have some lineage-restricted SRE motifs (Fig. 6A and fig. S17). Perhaps the intrinsic difficulty of identifying SSs using local information in longer mammalian transcripts necessitates more species-optimized SRE scores than in other lineages. On the other hand, the mammalian-trained SREs performed reasonably well on all organisms, always outperforming the organism-specific SMsplice model without SREs (Fig. 3C), suggesting a reduced prominence of lineage-specific SREs in non-mammals.

The variations between organisms in the other components of the SMsplice model play a substantial role in the ability to generalize across species. For instance, using the human SMsplice-learned SRE scores within an organism's SMsplice model always outperformed the fully human-specific model (Fig. 6D). In every case, the organism-specific SMsplice model had the best performance, as expected. As a point of comparison, we also applied the neural net model CI-SpliceAI to the same test sets (42). This black box model was trained on human sequences, so the most direct comparison is to the SMsplice model with fully human parameters. Comparing performance, CI-SpliceAI had substantially higher performance on human and mouse genes than our human SMsplice model but did not generalize as well to non-vertebrates. Furthermore, our organism-specific SMsplice models outperformed CI-SpliceAI on all non-vertebrate organisms. As CI-SpliceAI is a black box model, we cannot consider extracting parameters related to SREs, for example, as we could for the SMsplice model, so it was not feasible to mix and match CI-SpliceAI with SMsplice. One
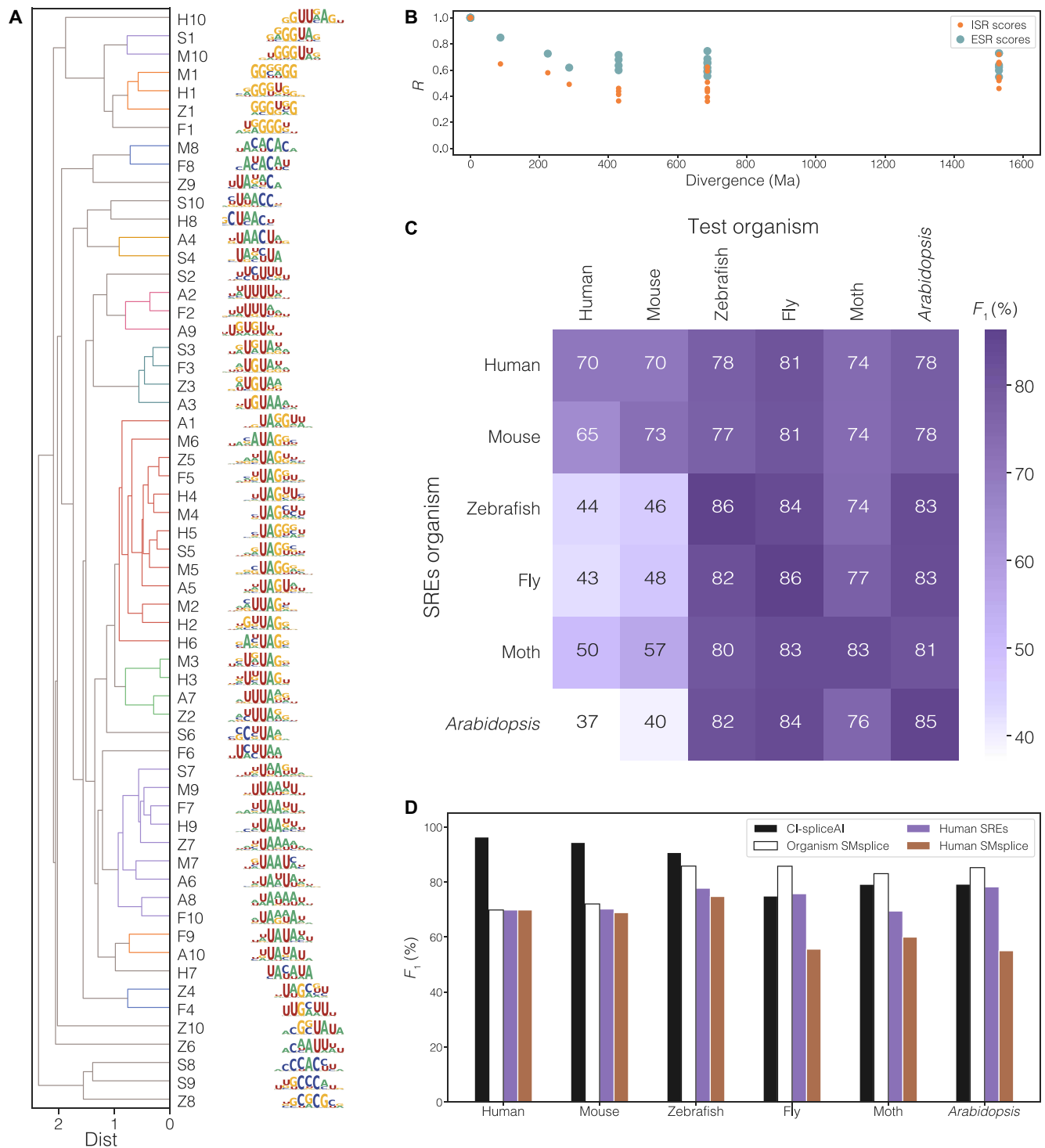
Fig. 6. Comparing SRE motifs and splicing models across organisms. (**A**) All ESS clusters for all organisms clustered by motif similarity. Motifs are labeled using the first letter of the organism, using "S" for silkworm moth, followed by the number of the ESS cluster in fig. S8. (**B**) Correlation between ESR scores and ISR scores for each pair of organisms (as well as fugu, data otherwise not shown) as a function of their evolutionary divergence. (**C**) F_1 performance on each organism using weighted SREs from each other organism. (**D**) F_1 performance on each organism for CI-SpliceAI, the SMsplice model with all organism-specific parameters, the SMsplice model with SRE scores from human, but otherwise all organism-specific parameters, and the human SMsplice model.

of the advantages in structuring a model in an explicit manner as we have done is that it allows for the exploration and interpretation of the different parameters of the model as shown in Fig. 6 (C and D).

DISCUSSION

A few simple assumptions about features important for splicing underly SMsplice, principally that SS recognition depends on a core motif modulated by nearby SREs and that SSs are recognized in pairs with favored lengths for both exons and introns. The model depends on just three types of parameters: the triplet-based core SS motif models, structural parameters capturing preferences for particular exon and intron lengths and numbers, and scores of hexamers as exonic and intronic SREs. The core SS motif and structural parameters are directly estimated from frequencies in the training data, with no adjustments or "learning" involved, thus generating a scaffold for SS recognition in the absence of SFs. Scores of hexamers as exonic and intronic SREs were then learned on the basis of their ability to predict splicing patterns within this scaffold.

The organism-specific SMsplice models enabled us to explore many facets of splicing across species. For example, we found that organisms with longer introns have greater reliance on SRE scores and that *Drosophila* had the greatest reliance on the structural parameters, likely related to strong preferences for a narrow range of short intron lengths in flies (*43*). In addition, we generally observed that ESR scores had greater prominence than ISR scores, except in *Arabidopsis* where ESRs and ISRs appeared comparably important, consistent with early studies showing that U-rich sequences distributed throughout plant introns are important in intron recognition (*44*).

In human, where SREs and the effects of mutations on splicing outcomes have been extensively studied, we observed significant agreement between our learned parameters and experimentally validated sets of SREs (Fig. 4C), and the top and bottom 5% of ESR and ISR hexamers yielded clusters that match well to binding motifs of many known SFs (Fig. 5 and figs. S11 to S16). In other organisms, agreement with the somewhat sparser sets of known RBP-binding motifs available was also observed, with additional clusters representing candidate novel SREs in each lineage suitable for experimental tests of function. Unlike qualitative catalogs of SREs identified previously, our approach learns a score for each hexamer, distinguishing stronger and weaker ESEs, ESSs, ISEs, and ISSs. This feature could be particularly useful for predicting the most essential SREs governing splicing of a given exon, intron, or splice site, with potential applications for interpretation of genetic variants or design of splice-switching antisense oligonucleotides (*45*).

In addition to the full SMsplice model, the new core SS motif models and CASS framework may have useful applications in assessing the strength of a SS in local context. These models can be applied even on short sequences of a few hundred bases (for CASS) or as short as a few dozen bases (for SS motifs), where neural models typically struggle. For example, CASS scores might find use in design or troubleshooting of splicing in minigenes or in synthetic biology applications. The SMsplice structure, which normally uses our new MaxEnt SS scores, could potentially also be applied to SS scores derived in other ways (*23*) because of its modular design.

While we have learned model parameters for several organisms, a predictive tool to study splicing might be useful in many additional organisms. In some cases, using the pretrained model for the evolutionarily closest organism may suffice, e.g., using the zebrafish model to predict splicing in other fish species seems reasonable. However, given sufficient high-quality annotated gene data, organism-specific parameters can be generated using the provided code. New structural parameters and MaxEnt models are straightforward to derive, while SRE score learning is more computationally intensive. Our experiments have shown reasonable generalization of SRE scores within mammals and within the non-mammals studied, while structural features vary substantially across lineages. Therefore, simply deriving new SS models and structural parameters with pretrained SRE scores may provide most of the benefits of a fully organism-specific splicing model with relatively modest effort.

Limitations of this work

The SMsplice framework was designed with typical animal and plant splicing systems in mind and is likely less suitable for modeling splicing in fungi, where the BPS motif is of much greater importance (*46*). Furthermore, our training and testing procedures are dependent on the accuracy of the underlying genomic sequences and annotations, potentially limiting application to organisms where annotations are less accurate. In addition, we have considered canonical annotations, which represent predominantly constitutive exons, and have not explicitly explored alternative splicing in this study. Another limitation is the fact that, while we have attempted to construct and learn the SRE scores in such a way that they reflect splicing regulatory activity (e.g., emphasizing hexamers within ~80 nt of the SS), it is inherently challenging to isolate this activity from biases in sequence composition of exons and introns that reflect other facets of gene expression, such as protein coding, mRNA export or stability, and/or processes of DNA mutation and repair. While we have shown that hexamers with the most extreme SRE scores are strongly associated with SF binding and splicing activity, weaker-scoring SRE hexamers may be less enriched for splicing regulatory activity. To enhance tractability and interpretability, we have also made the strong assumption that SREs act only over short distances and interact additively. This assumption ignores potential long-range activities and synergistic or antagonistic relationships that undoubtedly occur in some cases. These limitations represent worthwhile challenges for the next generation of interpretable splicing models.

METHODS
Datasets

The datasets used for analyses of human splicing were based on the canonical_dataset.txt used in training and testing of SpliceAI, in the hg19/GRCh37 genome version (*29*). We excluded from consideration any genes whose sequence contained "Ns" or nonstandard bases, whose total length was less than 100 bases, or whose shortest annotated intron was less than 25 bases (unrealistically short, suggesting an annotation error). From the genes not in the test set, we defined a validation set by randomly selecting 1000 nonparalogous genes. We then randomly selected 4000 genes from the remaining genes not in the test set to act as the training set. An additional set of 1000 randomly selected genes not in the test set was chosen as our SRE weight learning set. For practical (speed) reasons, we excluded genes longer than 200 kb from the validation and weight learning sets above. We used the same test set as SpliceAI, applying the filters as above, which removed ~1% of genes, resulting in a set of 1629 genes. Aside from the MaxEntScan training described below, these sets were used throughout the paper.

For each nonhuman organism, to focus on the most highly used and reliable splice sites, we created a canonical dataset containing a canonical splicing pattern for each protein-coding, intron-containing transcript that met the filtering criteria used for human genes. We also created an "all-SS" dataset containing the full complement of splice sites present within the respective canonical transcript, analogous to the gtex_dataset.txt from SpliceAI. Training, validation, and test sets were derived from the canonical dataset. The all-SS datasets were used for MaxEnt training and the real versus decoy splice site analyses.

For mouse, we used the knownCanonical table for the GRCm38/mm10 assembly to create the mouse canonical dataset, excluding the handful of genes with more than one annotation, and used the known-Gene table to create the all-SS mouse dataset. We downloaded paralog annotations from Ensembl's BioMart. The training, SRE weight learning, and validation sets were also selected from the genes not used in the test set as described for human. Nonparalogous genes on chromosomes 1, 3, 5, 7, and 9 were used to form the test set, resulting in a set of 1212 genes.

For zebrafish, splicing annotations for GRCz11 were downloaded from Ensembl's BioMart along with paralog status and APPRIS annotations (47). To select the canonical transcript, we filtered for APPRIS principal transcripts for each gene and selected the longest transcript for genes with multiple APPRIS principal transcripts. Genes without a unique longest APPRIS principal transcript were discarded. The training, SRE weight learning, and validation sets were selected as in human from genes that were not used in the test set, which was made of nonparalogous genes on chromosomes 1, 3, 5, 7, and 9, resulting in a set of 825 genes.

For *Drosophila*, we downloaded splicing annotations, APPRIS annotations, and paralog status for dm6 from Ensembl's BioMart. The canonical and full set of SS datasets were created in the same manner as zebrafish. The training, weight training, and validation sets were selected from the genes not used in the test set in the same manner as human, and the test set was made from all the nonparalogous genes on chromosomes 2L and 3L, resulting in a set of 1938 genes.

For silkworm moth, we downloaded splicing annotations and paralog status for Bmori_2016v1.0 from EnsemblMetazoa's BioMart and defined the canonical transcripts using the Ensembl canonical set. For this genome, we filtered out genes that were annotated as mitochondrial rather than filtering for genes annotated as chromosomal and made the training, SRE weight learning, and validation sets from the genes not used in the test set as described for human. The test set was made from the nonparalogous genes with the same filters as human on the primary assemblies named BHWX01000012.1, BHWX01000013.1, BHWX01000018.1, BHWX01000021.1, BHWX01000022.1, BHWX01000027.1, BHWX01000038.1, BHWX01000074.1, and BHWX01000097.1, resulting in a set of 920 genes.

For *Arabidopsis*, we used EnsemblPlant's BioMart to download the paralog status and splicing annotations for genome version TAIR10. As with silkworm moth, we used the Ensembl canonical set to define the canonical transcripts. The training, SRE weight learning, and validation sets were selected from the genes not used in the test set as described for human. The test set was made of nonparalogous genes on chromosomes 2 and 4, resulting in a set of 1117 genes.

New MaxEntScan models

To create the training and testing data for our third-order model of human SSs, we downloaded the GCF_000001405.39 NCBI assembly for hg38/GRCh38. For each intron-containing, protein-coding gene at least 300 bases in length (to ensure sufficient null examples), we used all 5′SSs and 3′SSs for true examples and then randomly selected 30 random non-SS positions for each SS to act as background examples. We removed any sequences that contained noncanonical bases, randomly selected one third of each set to act as test data for Fig. 1B, and then used the remaining two-thirds to train on.

To generate new SS models, we reimplemented the iterative scaling procedure described in the original MaxEntScan paper (22), but used all empirical trinucleotide frequencies (from adjacent as well as non-adjacent triples of positions) as constraints, in addition to the mono- and dinucleotide constraints used previously. For the 5′SS, we used the same 9mer around the SS, but rather than considering the consensus dinucleotide and remaining positions separately, we used the entire 9-nucleotide oligomer sequence. Similarly for the 3′SS, we incorporated the consensus dinucleotide into the considered subsequences. So, in a similar notation as the original paper, where the subscripts refer to sequence position, the probability of generating a 23-nt potential 3′SS sequence X is

$$P_{\text{overlap}}(X) = \frac{P(X_1, \dots, X_7)P(X_8, \dots, X_{14})P(X_{15}, \dots, X_{23})P(X_5, \dots, X_{11})P(X_{12}, \dots, X_{18})}{P(X_5, \dots, X_7)P(X_8, \dots, X_{11})P(X_{12}, \dots, X_{14})P(X_{15}, \dots, X_{18})}$$

Furthermore, for both the 5′SS sequence and the 3′SS subsequences, the constraints on the maximum entropy distributions include all third-order frequency constraints—i.e., the frequencies of triples of nucleotides at all possible sets of three consecutive or non-consecutive SS positions—in addition to the second- and first-order constraints. The original MaxEnt methods are used in this paper only in the comparison analyses shown in Figs. 1B and 3A.

For the other organisms, the same training procedure and constraints were used, with the training data coming from the all-SS datasets for all genes at least 300 base pair (bp) in length that are not present in the respective test sets described above. (No validation set is needed for SS training since the parameters are obtained through a deterministic iterative scaling procedure.) These test sets were then used in the comparison analysis in Fig. 3A.

Iterative learning of scores

To iteratively learn SRE scores, we began by dividing training set genes (excluding those >200 kb in length) into four equally sized subsets (plus or minus one gene). For the first subset, we made predictions for the set, either using the CASS model or the SMsplice model. Comparing these predictions to the canonical annotations allowed us to define sets of FPs and FNs. We counted the occurrences of kmers in the intronic and exonic flanking regions for these sets in the same manner as the decoy SS flanking regions described above. This yielded four sets of kmer counts: exonic regions flanking FNs ($C^{e\,\text{fn}}$), exonic regions flanking FPs $C^{e\,\text{fp}}$), intronic regions flanking FNs $C^{i\,\text{fn}}$), and intronic regions flanking FPs $C^{i\,\text{fp}}$). We then added pseudocounts as follows:

$$C_{\text{kmer}}^{e\,\text{fn}} \mathrel{+}= \frac{\sum C^{e\,\text{fn}}}{\sum C^{e\,\text{fn}} + \sum C^{e\,\text{fp}}} \qquad C_{\text{kmer}}^{e\,\text{fp}} \mathrel{+}= \frac{\sum C^{e\,\text{fp}}}{\sum C^{e\,\text{fn}} + \sum C^{e\,\text{fp}}}$$

$$C_{\text{kmer}}^{i\,\text{fn}} \mathrel{+}= \frac{\sum C^{i\,\text{fn}}}{\sum C^{i\,\text{fn}} + \sum C^{i\,\text{fp}}} \qquad C_{\text{kmer}}^{i\,\text{fp}} \mathrel{+}= \frac{\sum C^{i\,\text{fp}}}{\sum C^{i\,\text{fn}} + \sum C^{i\,\text{fp}}}$$

Pseudocounts were added to smooth the empirical frequencies and to ensure that when we normalize each class by the total count to obtain the frequency, any hexamer that did not appear in either the exonic regions flanking FNs or the exonic regions flanking FPs will be assigned the same frequency for both sets and likewise in the intronic case. Thus, the $\log_2$ of the frequency ratio will be zero. This is a desirable property because when we use these resulting frequency ratios to update the SRE scores, we do not want the scores of $k$mers that did not appear near false predictions to be affected. Letting the frequencies for a particular kmer be represented by $f_{kmer}^{e\ fn}$, $f_{kmer}^{e\ fp}$, $f_{kmer}^{i\ fn}$ and $f_{kmer}^{i\ fp}$; the update is done as follows

$$\sigma_{h(kmer)}^e + = \gamma \log_2\left(\frac{f_{kmer}^{e\ fn}}{f_{kmer}^{e\ fp}}\right) \qquad \sigma_{h(kmer)}^i + = \gamma \log_2\left(\frac{f_{kmer}^{i\ fn}}{f_{kmer}^{i\ fp}}\right)$$

Here, γ is a learning rate. We used to $\gamma = 0.01$ throughout this study because this value performed best in early tests.

Length distributions

We used two general methods of smoothing empirical length distributions to create the length distributions for SMsplice. The first method was Gaussian kernel smoothing, accomplished using the neighbors. KernelDensity class of scikitlearn. The second method was adaptive width KDE smoothing, for which we used the GaussianKDE class in the awkde package available at https://github.com/mennthor/awkde with an α value of 1 and otherwise default parameters. For both of these methods, we modified the tail of the smoothed distribution to allow introns and exons of arbitrarily long lengths. This was done by taking the density of the distribution past a cutoff point and replacing the tail with an appropriately scaled geometric distribution. So if the value of the distribution at the cutoff point is p, the value of the distribution for the next integer length is $p \cdot s$, and then for the next integer length, it is $p \cdot s^2$, and so on, where s is chosen so that the total density still sums to one. For the intronic distributions, we additionally imposed a minimum length, forcing the probabilities of all the lengths below a certain value to be zero, after which we renormalized the distribution. As all the genes considered in our analyses contained introns, we did not learn length distributions for single-exon genes. The choices of smoothing parameters below were made in each case to improve the fit to the empirical distributions on the training sets, as judged by eye.

For human, the intronic length distribution was smoothed via Gaussian kernel smoothing with a bandwidth 15, a steric constraint of 60, and transitioning to a geometric tail at a length of 1000 nt. All the exonic length distributions were smoothed using adaptive width KDE smoothing and transitioning to a geometric tail at the 80th percentile of the empirical lengths.

For mouse, exon and intron length distributions were smoothed as in human. For zebrafish, the intronic length distribution was smoothed via Gaussian kernel smoothing with a bandwidth 5, a steric cutoff of 50, and a geometric tail at 5000. All the exonic length distributions were smoothed using adaptive width KDE smoothing and a geometric transition at the 80th percentile of the empirical lengths.

For fly, all the length distribution were smoothed via Gaussian kernel smoothing. The intronic length distribution used a bandwidth 5, a geometric tail transition of 2000, and a steric cutoff of 40. The first exon length distribution used a bandwidth 30 and a geometric tail transition of 300. The internal exon length distribution used a bandwidth 30 and a geometric tail transition of 500. The last exon

length distribution used a bandwidth 100 and a geometric tail transition of 750.

For moth, the intronic length distribution was smoothed via Gaussian kernel smoothing with a bandwidth 15, a geometric tail transition of 3000, and a steric cutoff of 60. The first and internal exonic length distributions were smoothed using adaptive width KDE smoothing and a geometric transition determined at the 80th percentile of the empirical lengths. The last exon length distribution was smoothed using adaptive width KDE smoothing and a geometric transition determined at the 85th percentile of the empirical lengths.

For *Arabidopsis*, the intronic length distribution was smoothed via Gaussian kernel smoothing with a bandwidth 5, a geometric tail transition of 200, and a steric cutoff of 60. All the exonic length distributions were smoothed using adaptive width KDE smoothing and a geometric transition determined at the 80th percentile of the empirical lengths.

SMsplice and relation to HSMM models

The SMsplice model is most easily described in relation to an HSMM associated with the hidden structure shown in Fig. 2A. This structure involves seven hidden states: E_S, E_F, E_M, E_L, I, 5′, and 3′, with S and ε representing the traditional start and end states, from which all valid parses must begin and end, respectively. The transition probabilities between states are all set to 0 or 1, as indicated by the arrows, except the transitions from 3′ to E_L and E_M, which are set to p_{EO} and $p_{EI} = 1 - p_{EO}$, respectively, where p_{EO} is the parameter fit in Fig. 2B. Here, we have set the S to E_F transition probability, p_{ME}, to 1 and p_{1E} to 0, effectively assuming that genes have at least one intron, but these transitions can be set differently if one wishes to include intronless genes as a possibility. The length distributions for each of the hidden states were learned as described above; the 5′ and 3′SS states were assigned a length of 1 with probability 1 for convenience, although of course the associated core SS motifs extend upstream and downstream of these positions.

To be fully specified, a standard HSMM would require a set of possible emissions and probability distributions over these emissions for each hidden state, and one could then find the highest probability parse for any sequence of emissions using the HSMM version of the Viterbi algorithm (*34*). However, our main goal here was to provide a framework to apply the structural constraints of the spliceosome, particularly length distributions and SS pairing, for use in discriminating more and less plausible splicing patterns, rather than building a generative model. Therefore, where an HSMM would consider the probability of some parse π for a given sequence, P[seq, π], for SMsplice, we define SM[seq, π] an expression analogous to the base-2 logarithm of P[seq, π], normalized to a background model of sequence composition. For example, our SS-only model uses the MaxEnt log-odds ratios (representing the log of the probability of generating a sequence segment under the SS model to that of generating it under a background model) in place of the terms an HSMM would use for emissions from the SS hidden states. Also, it assigns a log-odds value of 0 in place of the emissions for other states (representing introns and different types of exons), effectively treating these states as not different from background (i.e., odds ratio of 1). The CASS model makes similar substitutions, replacing SS emission terms by CASS scores, which derive from log-odds ratios but do not necessarily correspond to the log-odds ratios of any specific pair of sequence-generative models; again, the emissions terms for other

states are replaced by log-odd values of 0, effectively ignoring the composition of exons and introns outside of the local regions that contribute to CASS scores.

Because the SMsplice model is defined in log-odds rather than generative terms, it is technically a semi-Markov conditional random field (CRF) (35, 36). It satisfies the Markov condition that the odds of hidden state $i + 1$ given seq and any subset of hidden state values depend only on seq and the previous state i (via the transition probability from the state i to state $i + 1$). The CASS framework defines the SS score at some position j as a function of sequences up to ~100 bp upstream and downstream of j (via the core SS motif and ESR and ISR scores). Such a definition is compatible with a CRF framework, where conditional probabilities (or log-odds) of hidden states are defined conditionally on the entire input sequence.

Consider a parse π of a sequence of length T, which has $N > 0$ introns of lengths $d_1^I, \ldots, d_N^I$, and therefore $N - 1$ internal exons, which have lengths $d_1^E, \ldots, d_{N-1}^E$. Let the first exon have length d_F^E and the last exon have length d_L^E. Then, otherwise using the notation discussed above and shown in Fig. 2, the base-2 logarithm of the complete data likelihood of the fully specified HSMM would be

$$\log_2(\mathbb{P}[\text{seq}, \pi])$$

$$= \log_2(p_{\text{ME}}) + \log_2\left[p_{\text{ELF}}(d_F^E)\right] + \log_2\left[p_{\text{IL}}(d_1^I)\right] + \log_2(p_{\text{EO}}) + \log_2\left[p_{\text{ELL}}(d_L^E)\right]$$

$$+ \sum_{n=1}^{N-1} \left\{ \log_2(p_{\text{EI}}) + \log_2\left[p_{\text{ELM}}(d_n^E)\right] + \log_2\left[p_{\text{IL}}(d_{n+1}^I)\right] \right\}$$

$$+ \sum_{t=1}^{T} \left[\log_2(\mathbb{P}[\text{seq}_t \mid \text{state at position } t])\right]$$

The SMsplice function is reproduced here, with minor modifications to emphasize correspondence with the HSMM

$$SM[\text{seq}, \pi] = \log_2(p_{\text{ME}}) + \log_2\left[p_{\text{ELF}}(d_F^E)\right] + \log_2\left[p_{\text{IL}}(d_1^I)\right] + \log_2(p_{\text{EO}}) + \log_2\left[p_{\text{ELL}}(d_L^E)\right]$$

$$+ \sum_{n=1}^{N-1} \left\{ \log_2(p_{\text{EI}}) + \log_2\left[p_{\text{ELM}}(d_n^E)\right] + \log_2\left[p_{\text{IL}}(d_{n+1}^I)\right] \right\}$$

$$+ \sum_{t \text{ where } \pi \text{ has a 5'SS}} s_t^5 + \sum_{t \text{ where } \pi \text{ has a 3'SS}} s_t^3$$

Thus, SM[seq, π] uses the same sums of log probabilities as in an HSMM, except that the emissions term is replaced by CASS scores, which involve log-odds of SS-associated sequences relative to background or decoy sequences. Therefore, SMsplice is a discriminative model whose maximum can be found by applying the Viterbi algorithm for the associated HSMM. The parse π^* that maximizes SM[seq, π] is taken to be the SMsplice prediction for the input sequence.

Decoy SS set and real versus decoy scores

To identify decoy SSs, we scored every position in transcripts of the training set using the third-order MaxEntScan methods for the organism in question. For 5'SS and 3'SS, we collected the scores of all the positions annotated in the canonical splicing pattern, which flanked an internal exon. For each of these real SS scores, ordered from strongest to weakest, we selected a random gene in the training set. If that gene had a position that was not annotated as a splice site (from gtex_dataset.txt for human and from the all-SS datasets for other organisms), was not already selected as a decoy, and had an associated MaxEnt score within 0.5 bits of the real SS in question, then we designated that position as a decoy. If the gene lacked a site with appropriate score,

then another gene was selected at random without replacement. If none of the genes in the training set had such a position, then the SS was left unmatched. All SSs were matched by decoys for human, mouse, zebrafish, and moth, while fly had just 13 unmatched 3'SSs (less than 1%) and 1351 unmatched 5'SSs (11%), and *Arabidopsis* had 2131 unmatched 3'SSs (11%) and 1610 unmatched 5'SSs (8%).

To score $k$mers using the matched decoy SS set, we counted the occurrence of each $k$mer in the flanking regions, the regions of length r that abut but do not overlap the MaxEntScan regions. Flanking regions were also required to remain within the boundaries of the gene; if after this shortening the region was reduced to less than k bases, then the region was discarded. The $k$mer counts for real SSs were similarly calculated, further shortening the flanking region to avoid overlap with any other SS motifs in the vicinity, if necessary (considering the segment scored by MaxEntScan to represent the SS motif). So, for instance, in the case of an exon of length $< r$, the flanking region upstream of the 5'SS and downstream of the 3'SS would be the same exonic sequence. Overall, this procedure yielded $k$mer counts upstream and downstream of real and decoy 5'SSs and 3'SSs. $k$mers upstream of real and decoy 5'SSs were considered exonic, while downstream $k$mers were considered intronic, with the opposite convention used for real and decoy 3'SSs. A pseudocount of two was added to each of these counts for smoothness, and then each category was normalized to obtain the frequency of each kmer in each region.

Representing the resulting frequency values for a particular $k$mer as $f_{k\text{mer}}^{e\ \text{real}}$, $f_{k\text{mer}}^{e\ \text{decoy}}$, $f_{k\text{mer}}^{i\ \text{real}}$, and $f_{k\text{mer}}^{i\ \text{decoy}}$, we seeded exonic and intronic scores by assigning σ^e and σ^i as

$$\sigma_{h(k\text{mer})}^e = c \cdot \log_2\left(\frac{f_{k\text{mer}}^{e\ \text{real}}}{f_{k\text{mer}}^{e\ \text{decoy}}}\right) \qquad \sigma_{h(k\text{mer})}^i = c \cdot \log_2\left(\frac{f_{k\text{mer}}^{i\ \text{real}}}{f_{k\text{mer}}^{i\ \text{decoy}}}\right)$$

Here, the "SRE weight constant" $c > 0$ was chosen using a binary search to maximize the F_1 performance on the SRE weight learning set when using these SRE scores. This binary search was initialized by considering the values $c = 0$ and $c = 1$, and our binary search was limited to 16 values. The resulting scores are used in all the relevant analyses following Fig. 4B and fig. S5.

Clustering hexamers

For the clustering analyses shown in Fig. 5, we defined the "distance" (dissimilarity) between two $k$mers by checking the agreement of the sequences for each possible amount of overlap. For an overlap of one base, the base distance score was $k - 1$; for an overlap of two bases, the distance was $k - 2$ and so on. Then, for every base that did not agree in the overlapping potion, we incremented the distance by 1. Last, we defined the distance as the minimum base distance score obtained from evaluating all $2k - 1$ possible overlaps between the two $k$mers. These distances were used to construct a distance matrix for all the $k$mers in the relevant set, and the $k$mers were clustered on the basis of this matrix using average linkage hierarchical clustering to define 10 clusters of at least four $k$mers. As in the RESCUE-ESE paper, the $k$mers in each cluster containing at least four $k$mers were aligned using ClustalW with default parameters (10). A PWM for each cluster was calculated from the frequency of each base at each position in the resulting alignment; these frequencies may sum to <1 at positions where one or more sequences were not aligned.

PWM comparisons

We first pruned any position where fewer than one-third of the sequences aligned. We then replaced these positions with a uniform distribution (frequency of ¼ for all four bases) and also added padded each PWM with uniform positions so that all PWMs extended for the full extent of the alignment. To compare two PWMs, we considered all possible ungapped alignments of the two PWMs, defining the PWM distance for each specific alignment as the sum of the Jensen-Shannon divergence values between corresponding positions; the distance between the two PWMs was then defined as the minimum PWM distance across all alignments of the PWMs. When clustering the PWMs within each category for all organisms, we used this distance for average linkage hierarchical clustering.

RNAcompete comparisons

Using the PWM distance measure described above, we compared each of the PWMs determined for our clusters with each of the RNAcompete PWMs from the relevant organism and considered the smallest value as the best match. When that best matching motif was associated with multiple proteins, all of those proteins are reported. To determine a significant distance, we considered all of the PWMs for all of the clusters and changed their bases using the permutation A ➔ G ➔ T ➔ C ➔ A, e.g., a PWM with consensus sequence ATCCG would be converted to one with consensus GCAAT. Then, we scored these permuted PWMs against all the RNAcompete motifs for the relevant organism, plotted the distribution of minimum distances, and took the first percentile of this distribution as our distance cutoff, corresponding to $P < 0.01$.

Analysis of SRE cluster activity using RBP knockdowns

For each of the RNAcompete PWMs that was a significant match to any of the human clusters, we considered those with annotated splicing regulatory according to the ENCORE paper (*19*). For each of these splicing regulators, we downloaded the associated exon skipping RMATS results from ENCORE where available and removed any dataset where the number of splicing changes with a false discovery rate < 0.1 was less than 50. These datasets had the following identifiers: ENCFF394AJI, ENCFF461PND, ENCFF258VNP, ENCFF290I-YV, ENCFF964QGK, ENCFF542IZE, ENCFF177CCV, ENCFF692CYY, ENCFF103PLX, ENCFF433SCM, ENCFF094KRM, ENCFF707RAF, ENCFF651GFP, and ENCFF459QWA (encodeproject.org).

We imposed a minimum read coverage threshold on these exons by excluding exons where the mean sum of skipping and inclusion junction reads, averaged across conditions and replicates, was less than 20. Then, we associated exons with RBPs based on the presence of a hexamer from the associated cluster in the relevant CASS region, with exons lacking these hexamers considered nonassociated. For RBPs with multiple cluster matches, such as the two ESS clusters that matched HNRNPA1, we analyzed the clusters separately based on the same knockdown data. To ask whether exons associated with an RBP were more affected by the knockdown of that RBP than those that were not, we used a one-sided KS test to compare the distributions of inclusion level changes (delta PSI values) between knockdown and control conditions. For enhancer clusters, we asked for increased exclusion following knockdown, while for silencer clusters, we looked for increased inclusion following knockdown. Applying Benjamini-Hochberg multiple test correction with $\alpha = 0.1$ on the resulting comparisons, 13 of 26 were significant, with at least one significant result for 11 of 13 clusters analyzed (table S1).

We additionally downloaded any eCLIP dataset available from ENCODE for these RBPs that was performed in the same cell line as the corresponding shRNA knockdown experiment, which had the following identifiers: ENCFF549HTU, ENCFF604WJZ, ENCFF178XWA, ENCFF704OCI, ENCFF805LKH, and ENCFF969OHE (encodeproject.org). For these eCLIP datasets, we classified each exon as being associated with an eCLIP peak if the 5′ end of any peak was within 75 bases of a relevant CASS region. This allowed us to consider overlaps between eCLIP-associated exons and RBP-associated exons using the cluster member *k*mers as above. To explore whether these associations were in agreement as expected, we performed one-sided hypergeometric tests and corrected them with Benjamini-Hochberg multiple test correction with $\alpha = 0.1$, yielding significant results in all but one case (table S1).

Analysis of SRE cluster activity using sQTLs

Fine mapping of sQTLs and eQTLS in GTEx was performed by Barbeira and coworkers using the DAP-G (deterministic approximation of posteriors) fine-mapping method (*48*). We accessed this dataset using their zenodo link. Note that fine-mapping was only done on European Ancestry samples of GTEx. We filtered for intron clusters with two or three introns per cluster to filter for splicing events that can be meaningfully described as either alternative splice site exons or skipped exons. We created pan-tissue clusters by merging the locations of each intron in a cluster, dropping tissue-specific cluster IDs. We then sorted by posterior inclusion probability (PIP) values in descending order and dropped duplicates based on pan-tissue cluster IDs and variant pairs, keeping the first value. This insured that we obtained the top variant-phenotype pair for all splicing events.

From this set, we further filtered for only SNP variants and splicing events that represented skipped exons and choices between exactly two alternative 5′SSs or 3′SSs. For each of these events, we considered only the variants that fell in the regions that would be relevant to the SRE contribution to CASS scores of any of the SSs, and used the regions it fell in to determine whether the variant was intronic, exonic, or both. For each cluster, we determined the set of relevant variants by selecting those that fell in a relevant CASS region of an annotated SS and for which the (reference or alternative) variant fell within a hexamer of the cluster and changed the score of the resulting hexamer to the opposite sign. For instance, for an ESE cluster, we looked for variants that fell in exonic CASS regions where the variant overlapped a component hexamer of that cluster, and the variant changed the hexamer to one with an ESR score < 0.

We determined a control set of variants by considering variants that fell in CASS regions but did not fall in any of the most extreme 5% hexamers, i.e., none of the hexamers that overlapped the variant in either condition were in the top or bottom 5% of ESRs or ISRs. We then compared the distributions of PIP values associated with this control set of variants to those associated with each of the clusters. If the clusters are indeed associated with splicing regulation, then we would expect their PIP values to be generally larger than those for the control set. To measure this, we used a one-sided KS test to assess whether the distributions were significantly shifted in the appropriate direction, indicating increased likelihood of causality. We used Benjamini-Hochberg multiple test correction with $\alpha = 0.1$ to determine final significance shown in Fig. 5C.

We further considered the subset of these events to which we could assign the direction of the change in the tissue associated with the highest PIP by cross-referencing with the data contained

in GTEx_Analysis_v8_sQTL_EUR.tar, downloaded from the GTEx portal (www.gtexportal.org/home/downloads/adult-gtex/qtl). For each such variant, we considered the changes in local scores between the two sequences to determine a difference or "delta" value. For the exon skipping events, we subtracted the local exon score for the new sequence from the score for the reference sequence to determine delta. For the alternative splice sites, we calculated the local intron scores for both the longer and shorter introns for both the variant and reference sequences. We then defined delta for these events by subtracting the change in the score of the intron-proximal alternative SS (i.e., associated with splicing of a shorter intron) pattern from the change in score for the intron-distal site, meaning that a positive delta suggests a change toward splicing of the longer intron isoform.

Supplementary Materials

This PDF file includes:
Figs. S1 to S17
Legends for tables S1 and S2

Other Supplementary Material for this manuscript includes the following:
Tables S1 and S2

REFERENCES AND NOTES

1. H. Keren, G. Lev-Maor, G. Ast, Alternative splicing and evolution: Diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–355 (2010).
2. M. C. Wahl, C. L. Will, R. Lührmann, The spliceosome: Design principles of a dynamic RNP machine. *Cell* **136**, 701–718 (2009).
3. L. P. Lim, C. B. Burge, A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11193–11198 (2001).
4. A. Busch, K. J. Hertel, Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev. RNA.* **3**, 1–12 (2012).
5. Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, C. B. Burge, Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831–845 (2004).
6. Y. Wang, M. Ma, X. Xiao, Z. Wang, Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–1052 (2012).
7. Y. Wang, X. Xiao, J. Zhang, R. Choudhury, A. Robertson, K. Li, M. Ma, C. B. Burge, Z. Wang, A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat. Struct. Mol. Biol.* **20**, 36–45 (2013).
8. A. B. Rosenberg, R. P. Patwardhan, J. Shendure, G. Seelig, Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
9. A. Goren, O. Ram, M. Amit, H. Keren, G. Lev-Maor, I. Vig, T. Pupko, G. Ast, Comparative analysis identifies exonic splicing regulatory sequences–The complex definition of enhancers and silencers. *Mol. Cell* **22**, 769–781 (2006).
10. W. G. Fairbrother, R. F. Yeh, P. A. Sharp, C. B. Burge, Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013 (2002).
11. P. Baeza-Centurion, B. Miñana, J. M. Schmiedel, J. Valcárcel, B. Lehner, Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176**, 549–563.e23 (2019).
12. P. Baeza-Centurion, B. Miñana, J. Valcárcel, B. Lehner, Mutations primarily alter the inclusion of alternatively spliced exons. *eLife* **9**, e59959 (2020).
13. S. Ke, V. Anquetil, J. R. Zamalloa, A. Maity, A. Yang, M. A. Arias, S. Kalachikov, J. J. Russo, J. Ju, L. A. Chasin, Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).
14. X. D. Fu, M. Ares, Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15**, 689–701 (2014).
15. D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, T. R. Hughes, A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
16. D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, C. B. Burge, Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol. Cell.* **70**, 854–867.e9 (2018).
17. M. Jens, M. McGurk, R. Bundschuh, C. B. Burge, RBPamp: Quantitative modeling of protein-RNA interactions in vitro predicts in vivo binding. bioRxiv 515616 [Preprint]. 9 November 2022.
18. H. T. Rube, C. Rastogi, S. Feng, J. F. Kribelbauer, A. Li, B. Becerra, L. A. N. Melo, B. V. do, X. Li, H. H. Adam, N. H. Shah, R. S. Mann, H. J. Bussemaker, Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nat. Biotechnol.* **40**, 1520–1527 (2022).
19. E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. L. Cody, D. Dominguez, S. Olson, B. Sundararaman, L. Zhan, C. Bazile, L. P. B. Bouvrette, J. Bergalet, M. O. Duff, K. E. Garcia, C. Gelboin-Burkhart, M. Hochman, N. J. Lambert, H. Li, M. P. M. Gurk, T. B. Nguyen, T. Palden, I. Rabano, S. Sathe, R. Stanton, A. Su, R. Wang, B. A. Yee, B. Zhou, A. L. Louie, S. Aigner, X.-D. Fu, E. Lécuyer, C. B. Burge, B. R. Graveley, G. W. Yeo, A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
20. M. S. Wong, J. B. Kinney, A. R. Krainer, Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol. Cell.* **71**, 1012–1026.e3 (2018).
21. M. B. Shapiro, P. Senapathy, RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155–7174 (1987).
22. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
23. J. Cheng, T. Y. D. Nguyen, K. J. Cygan, M. H. Çelik, W. G. Fairbrother, Ž. Avsec, J. Gagneur, MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
24. H. Y. Xiong, Y. Barash, B. J. Frey, Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–2562 (2011).
25. Y. Barash, J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, B. J. Frey, Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
26. H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, B. J. Frey, RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
27. M. Mort, T. Sterne-Weiler, B. Li, E. V. Ball, D. N. Cooper, P. Radivojac, J. R. Sanford, S. D. Mooney, MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15**, R19 (2014).
28. X. Jian, E. Boerwinkle, X. Liu, In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42**, 13534–13544 (2014).
29. K. Jaganathan, S. K. Panagiotopoulou, J. F. M. Rae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, K. K.-H. Farh, Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
30. K. Gupta, C. Yang, K. McCue, O. Bastani, P. A. Sharp, C. B. Burge, A. Solar-Lezama, Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing. *Genome Biol.* **25**, 25 (2024).
31. L. De Conti, M. Baralle, E. Buratti, Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev. RNA.* **4**, 49–60 (2013).
32. C. R. Sibley, L. Blazquez, J. Ule, Lessons from non-canonical splicing. *Nat. Rev. Genet.* **17**, 407–421 (2016).
33. B. Wieringa, E. Hofer, C. Weissmann, A minimal intron length but no specific internal sequence is required for splicing the large rabbit beta-globin intron. *Cell* **37**, 915–925 (1984).
34. Y. Shun-Zheng, H. Kobayashi, An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Process. Lett.* **10**, 11–14 (2003).
35. S. Sarawagi, W. W. Cohen, Semi-markov conditional random fields for information extraction. *Adv. Neural. Inf. Process. Syst.* **17**, (2004).
36. J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (Icml, 2001), vol. 1, p. 3.
37. A. A. Pai, T. Henriques, K. McCue, A. Burkholder, K. Adelman, C. B. Burge, The kinetics of pre-mRNA splicing in the *Drosophila* genome and the influence of gene architecture. *eLife* **6**, e32537 (2017).
38. D. L. Black, Finding splice sites within a wilderness of RNA. *RNA* **1**, 763–771 (1995).
39. E. F. Cáceres, L. D. Hurst, The evolution, impact and properties of exonic splice enhancers. *Genome Biol.* **14**, R143 (2013).
40. G. Yeo, S. Hoon, B. Venkatesh, C. B. Burge, Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15700–15705 (2004).
41. S. Kumar, M. Suleski, J. M. Craig, A. E. Kasprowicz, M. Sanderford, M. Li, G. Stecher, S. B. Hedges, TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

42. Y. Strauch, J. Lord, M. Niranjan, D. Baralle, CI-SpliceAI—Improving machine learning predictions of disease causing splicing variants using curated alternative splice sites. *PLOS ONE* **17**, e0269159 (2022).

43. M. Talerico, S. M. Berget, Intron definition in splicing of small Drosophila introns. *Mol. Cell. Biol.* **14**, 3434–3445 (1994).

44. Z. J. Lorkoví, D. A. W. Kirk, M. H. L. Lambermon, W. Filipowicz, Pre-mRNA splicing in higher plants. *Trends in Plant Science* **5**, 160–167 (2000).

45. M. A. Havens, M. L. Hastings, Splice-switching antisense oligonucleotides as therapeutic drugs. *Nucleic Acids. Res.*, 6549–6563 (2016).

46. D. M. Kupfer, S. D. Drabenstot, K. L. Buchanan, H. Lai, H. Zhu, D. W. Dyer, B. A. Roe, J. W. Murphy, Introns and splicing elements of five diverse fungi. *Eukaryot. Cell* **3**, 1088–1100 (2004).

47. J. M. Rodriguez, P. Maietta, I. Ezkurdia, A. Pietrelli, J. J. Wesselink, G. Lopez, A. Valencia, M. L. Tress, APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).

48. A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, GTEx v8 fine mapping on eQTL and sQTL (Zenodo, 2019).