Exon definition facilitates reliable control of alternative splicing

Authors: Mihaela Enculescu[1], Simon Braun[1], Samarth Thonta Setty[2], Kathi Zarnack[2], Julian König[1]*, Stefan Legewie[1]*

Affiliations:

[1]Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany.

[2]Buchmann Institute for Molecular Life Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt a.M., Germany

*Correspondence to: s.legewie@imb-mainz.de, j.koenig@imb-mainz.de.

Abstract:

Alternative splicing is a key step in eukaryotic gene expression that allows the production of multiple protein isoforms from the same gene. Even though splicing is perturbed in many diseases, we currently lack insights into regulatory mechanisms promoting its precision and efficiency. Using mechanistic mathematical modeling, we show that alternative splicing control is facilitated if spliceosomes recognize exons as functional units ('exon definition'). We find that exon definition is crucial to prevent the accumulation of partially spliced retention products during alternative splicing regulation. Furthermore, it modularizes splicing control, as multiple regulatory inputs are integrated into a common net input, irrespective of the location and nature of the corresponding cis-regulatory elements in the pre-mRNA. These predictions of our model are qualitatively and quantitatively supported by high-throughput mutagenesis data obtained for an alternatively spliced exon in the proto-oncogene *RON* (*MST1R*). Our analysis suggests that exon definition has evolved as the dominant splice-regulatory mechanism in higher organisms to promote robust and reliable splicing outcomes.

One Sentence Summary: Exon definition is required for alternative precise splicing control without accumulation of undesired retention isoforms.

Keywords: alternative splicing, exon definition, mechanistic mathematical modeling, combinatorial regulation, high-throughput mutagenesis

Introduction

Eukaryotic gene expression is controlled at multiple levels. One important step in post-transcriptional gene regulation is splicing, the removal of intronic sequences from pre-mRNA precursors to yield mature mRNAs. Spliced mRNAs are then exported from the nucleus and translated into protein. In alternative splicing, certain exons are either included or excluded (skipped) to yield distinct mRNA and potentially protein isoforms. Alternative splicing is thought to be key to proteome complexity in higher eukaryotes and is perturbed in multiple diseases including cancer (*1*,*2*).

Splicing is catalyzed by a complex molecular machine, the spliceosome, that recognizes splice consensus sequences in nascent pre-mRNAs. The resulting splicing reaction generates mature mRNAs by removing intronic and joining exonic sequences. The catalytic cycle is initiated by recruitment of the U1 and U2 small nuclear ribonucleoprotein (snRNP) subunits to the 5' and 3' splice sites, respectively. Upon joining of further subunits (U4-U6 snRNPs) and extensive remodeling, a catalytically active higher-order complex is formed. Alternative splicing is commonly regulated by differential recruitment of the U1 and U2 snRNPs. In most cases, such modulation occurs by auxiliary RNA-binding proteins (RBPs) which promote or inhibit U1 or U2 snRNP recruitment by binding to intronic or exonic *cis*-regulatory elements (*1*,*3*,*4*).

Spliceosome assembly may occur by two conceptually different mechanisms: In 'intron definition', the U1 and U2 snRNPs directly assemble across the intron to form a catalytically competent spliceosome. Alternatively, a cross-exon complex of U1 and U2 snRNPs forms first in a process termed 'exon definition' and is then converted into the catalytic cross-intron complex. The simpler intron definition scenario is thought to be the default mechanism of splicing provided that introns are short enough (<200 bp) for efficient cross-intron spliceosome complex formation (*5*,*6*). Accordingly, intron definition is prevalent in lower organisms such as *S. cerevisae* and *Drosophila* that often display just one or few short introns per gene (*6*-*8*). In contrast, exon definition seems to be required for splicing of most mammalian genes, as these typically contain long introns and short exons (*6*-*8*). The predominant role of exon definition in mammals is supported by mutation effects on splice outcomes and by the co-evolution of *cis*-regulatory elements across exons (*6*,*8*). Furthermore, mathematical models accurately described human splicing kinetics when assuming an exon definition mechanism (*9*,*10*).

Here, we study the precision and efficiency of alternative splicing regulation, and focus on the role of intron and exon definition. We compare both mechanisms using mathematical modeling, study their functional implications and test the model predictions against comprehensive high-throughput mutagenesis data. As a model system, we use a cancer-relevant human splicing decision in the *RON* receptor tyrosine kinase gene, in which the alternative exon and its flanking introns are short (<150 bp), implying that both intron and exon definition scenarios are possible (*5*,*6*). We find that only exon definition quantitatively explains concerted splice product changes

upon sequence mutations. Thus, in human cells, the more complicated exon definition pathway may be the default mode of splicing and dominates if both mechanisms are in competition. This suggests that exon definition has additional benefits beyond spliceosome assembly across long introns. Indeed, we show that this mechanism greatly simplifies alternative splicing regulation compared to intron definition and efficiently prevents the generation of intron retention products, which frequently contain premature stop codons and are potentially toxic to cells. Our results provide a framework for the systems-level analysis of complex splice isoform patterns, and offer insights into the evolution of alternative splicing regulation.

Results

Mutations in the RON minigene induce concerted splice isoform changes

Using high-throughput mutagenesis and next-generation sequencing, we recently quantified the splice products originating from a splicing reporter minigene of the *MST1R* (*RON*) gene for 1942 single point mutations (*11*). The three-exon minigene covers *RON* alternative exon (AE) 11 (147 nt), the two flanking introns (87 and 80 nt, respectively) as well as constitutive exons 10 and 12 (210 and 166 nt, respectively; Fig. 1A). In HEK293T cells, the major splice product for the unmutated wildtype minigene is exon 11 inclusion (~59%), followed by full intron retention (~21%), exon 11 skipping (~12%), first intron retention (~4%) and second intron retention (~4%) (Figs. 1A and B). 510 out of the 1942 single point mutations quantified in our study induced significant changes in the isoform distribution (Fig. 1C). These mutation effects reflect a complex cis-regulatory landscape that we will use to train and test mathematical models of splicing regulation.

To learn more about the principles of splice isoform regulation, we focused on the 510 point mutations inducing the strongest changes on *RON* splicing (> 10% change in the relative abundance of any isoform w.r.t. wildtype). Using hierarchical clustering, we sorted these mutations according to their effect on all isoform frequencies, and found three types of splice isoform changes (Fig. 1D): in cluster 1, mutations induced anti-correlated changes in exon 11 inclusion and skipping, with little changes in intron retention isoforms. The remaining mutations additionally affected intron retention, either together with correlated changes in exon 11 inclusion and skipping (cluster 2), or with anti-correlated changes in inclusion and skipping (cluster 3). Taken together, these results indicate that mutation effects in *RON* converge on a small set of splice isoform patterns and may contain information about the underlying regulatory mechanisms.

Mathematical modeling discriminates intron and exon definition

We turned to mathematical modeling to mechanistically explain mutation-induced changes in splice isoforms. We assumed that mutations influence the recognition of splice sites by the spliceosome, and modeled the binding of spliceosomes to the pre-mRNA (Figs. 2A, 2B and S1). For simplicity, we only described the initial binding events, i.e., U1 and U2 snRNP binding to the 5' and 3' splice sites, respectively. Subsequent spliceosome maturation steps were not modeled explicitly, and it was assumed that splicing decisions are made based on the U1 and U2 snRNP recognition patterns (see below).

In our model, each U1 or U2 snRNP binding step to one of the six splice sites in the three-exon minigene is characterized by a recognition probability p_i. We assumed that U1 and U2 snRNP binding is fast compared to the subsequent spliceosome maturation and splicing catalysis. Then, the probabilities p_i are given by $k_i^{on}/\left(k_i^{on}+k_i^{off}\right)$, where k_i^{on} and k_i^{off} are the binding and unbinding rates of U1 or U2 snRNP to splice site i (see Materials and Methods, Section 2).

For each pre-mRNA molecule, multiple splice sites can be occupied at a time and depending on the individual recognition probabilities (p_i) such simultaneous binding may occur in different combinations. We describe the combinatorial nature of spliceosome binding by combining the individual recognition probabilities p_i into joint probabilities, one for each of the 64 (2^6) possible U1 and U2 snRNP binding configurations (Fig. S1). For instance, the joint probability of all six splice sites being simultaneously occupied is given by the product $p_1 \cdots p_6$, and this term changes to $\left(1-p_1\right)p_2 \cdots p_6$ if the first splice site is not occupied.

In the next step, we assigned a splicing outcome to each of the 64 binding states, and summed up the probabilities over all binding states yielding the same splicing outcome (Fig. S2). We thereby describe the frequency of five splice isoforms as a function of six spliceosome recognition parameters (p_i). By fitting this model to the measured mutation-induced isoform changes, we infer how mutations affect spliceosomal recognition of splice sites (see below).

In two alternative model variants, we implemented splicing decisions based on intron definition and exon definition mechanisms (Figs. 2A and 2B): For the intron definition model, it was assumed that an intron can be spliced out as soon as its flanking 5' and 3' splice sites are simultaneously occupied by U1 and U2 snRNPs (Fig. 2A, left). If multiple competing splicing reactions are possible in a binding configuration, we assumed that splicing occurs across the shortest distance (Fig. 2B and S2). The exon definition model involves an additional layer of regulation: before catalytic cross-intron complexes can form, transitory cross-exon U1-U2 snRNP complexes are required to stabilize initial U1/U2 snRNP binding to splice sites (Fig. 2A, right). We implemented this additional requirement for cross-exon complexes by assuming that an intron can only be spliced if all splice sites flanking the adjacent exons are occupied ('defined'). For example, splicing of the first intron requires full definition of neighboring exons 10 and 11, i.e., simultaneous recognition of splice sites 1-4 in the three-exon minigene (Fig. 2B and S2). Importantly, 26 out of 64 binding configurations generate distinct splicing outcomes in

the exon and intron definition models (Fig. S2). Hence, we expect that concerted isoform changes in our mutagenesis dataset (Fig. 1) will discriminate between intron and exon definition mechanisms.

High-throughput mutagenesis data supports the exon definition model

To investigate whether our mutagenesis data evidence intron and/or exon definition, we separately fitted these model variants to the measured frequencies of five splice isoforms for the wildtype sequence and 1854 single point mutations (see Table S1). During fitting, we assumed that mutations affect the recognition of one or multiple splice sites. In exon definition, U1 and U2 snRNP affect splicing only if they are simultaneously bound to both splice sites flanking an exon. Therefore, splicing outcomes depend only on three effective parameters (p_{12}, p_{34}, p_{56}), each reflecting the recognition probability of the complete exon. Thus, in exon definition there are three free parameters per mutation variant, whereas intron definition results in four independent parameters (see Materials and Methods and below).

Despite its lower degree of freedom, the exon definition model provides an overall much better fit to the mutagenesis data when compared to the intron definition model (Pearson correlation coefficients = 0.85 vs. 0.99, respectively; Fig. 2C, left and middle panels). The fit quality of the exon definition model can be further improved if we additionally allow five global parameters (shared between all mutation variants) to accommodate that long intron retention products may be under-represented in the RNA sequencing library due to metabolic instability (faster degradation of un-spliced transcripts (*12*)) and/or sequencing biases (such as PCR amplification or clustering problems for long fragments on the Illumina flowcell). Taken together, these quantitative results strongly favor exon definition as the predominant mechanism of *RON* splicing.

Qualitative arguments based on the algebraic sign of mutation-induced splice isoform changes further disfavor the intron definition model: first, isoform changes in the best-fit intron definition model frequently occur in opposite direction compared to the data, whereas this is not the case for the best-fit exon definition model (Fig. 2C, insets). Second, using analytical calculations, we show that the direction of isoform changes for splice site mutations completely abolishing spliceosome binding is fully consistent with exon definition, but partially disagrees with intron definition (Fig. 2D, Fig. S3 and Material and Methods, Section 2). This is particularly evident for mutations of the last splice site (5’ splice site of exon 12) which induce characteristic changes in all splice isoforms (Fig. 2D, left panel). Likewise, mutations in both splice sites flanking the alternative exon have the same effects on first and second intron retention which further supports the exon definition scenario (Fig. 2D, right panel).

Taken together, these results strongly support that *RON* exons 10-12 are spliced via exon definition, even though they are flanked by two short introns (80 nt and 87 nt). Thus, in human

cells exon definition may be the preferred and more efficient splicing mechanism, even if the gene structure (intron length) permits the simpler intron definition mode.

Notably, our conclusions concerning *RON* splicing are robust to the precise implementation of the exon definition mechanism: in our model, we assume that U1 and U2 snRNP independently recognize splice sites, and that cross-exon and cross-intron complexes form only later during spliceosome maturation. Alternatively, exon definition may already occur at the level of initial U1 and U2 snRNP binding, because both subunits cooperate across exons during splice site recognition *(9,13)*. In Materials and Methods, we show that both scenarios lead to the same splice isoform probability equations, implying that our fitting results also apply for strong cross-exon cooperation of U1 and U2 snRNP binding.

Modeling infers spliceosome relocation upon mutations and RBP knockdowns

To further validate the biological plausibility of the exon definition model, we analyzed how the exon recognition probabilities (p_{12}, p_{34}, p_{56}) are perturbed by point mutations in the best-fit model. In line with the intuitive expectation, we find that strong changes in exon recognition require point mutations to be located within or in close vicinity to the respective exon (Fig. 2E). For the outer constitutive exons, strong mutation effects are mostly confined to the corresponding splice sites, whereas the alternative exon is additionally regulated by a large number of non-splice-site mutations. This reflects the extensive regulation of alternative (but not constitutive) exons by nearby *cis*-regulatory elements. The recognition probability landscapes also provide plausible explanations for the concerted splice isoform changes we had identified by clustering (Fig. 1): Concerted changes in exon 11 inclusion and skipping (cluster 2) are explained by changes in constitutive exon recognition (p_{12} and p_{56}). On the contrary, any type of anticorrelated change in exon 11 inclusion and skipping (clusters 1, 3) is assigned to perturbed AE recognition (parameter p_{34}), p_{34} being affected with opposing directionality in each of the clusters (decreased p_{34} in cluster 1 and increased p_{34} in cluster 3).

Besides the effects of single point mutations, our model also allows us to quantify the effects of knockdowns of *trans*–acting RNA-binding proteins that control *RON* splicing. As a proof-of-concept, we fitted the exon definition model to human HEK293 data, in which the RBP *HNRNPH* was knocked down and splicing outcomes were quantified for the population of unmutated minigenes. As shown in Fig. S5A, we found that *HNRNPH* mainly controls the recognition of the alternative exon (p_{34}), with minor effects on the recognition of the outer constitutive exons (p_{12} and p_{56} , respectively). This agrees well with the expectation, since we previously showed by iCLIP and a genetic interaction screen in MCF7 cells that *HNRNPH* is bound throughout the minigene sequence, but primarily acts on splicing via a cluster of binding sites in the alternative exon (*11*). In Fig. S5B, we confirm for HEK293 cells that *HNRNPH* affects splicing outcomes by binding to the AE.

Hence, fitting the exon definition model to splice-perturbing experimental conditions allows us to reconstruct how these perturbations affect the splice site recognition by the spliceosome. This constitutes a first step towards reconstruction and mechanistic modeling of combinatorial splicing networks, in which many RBPs jointly control splicing.

Benefits of splicing regulation by an exon definition mechanism

To explore benefits of exon definition beyond the recognition of exons flanked by long introns, we used our models to perform splicing simulations. Interestingly, these simulations revealed that exon definition facilitates alternative splicing control when compared to intron definition. In our models, we simulate alternative splicing regulation by modulating the recognition probability of exon 11 at its 3’ or 5’ splice site. This mimics point mutations or the binding of regulatory RBPs nearby these splice sites. In the intron definition mechanism, splicing outcomes are very distinct, depending on whether p_3 and p_4 are regulated separately or jointly (Fig. 3A, left and Materials and Methods). In contrast, in the exon definition model, splicing outcomes are identical, irrespective of how the recognition of the 3' and 5' splice site of the alternative exon is regulated (Fig. 3A, right). Thus, only for exon definition, the alternative exon serves as a regulatory module which integrates inputs on both exon-flanking splice sites into a joint recognition probability p_{34} of the alternative exon 11. This modularization simplifies alternative splicing control and ensures that splicing outcomes are robust to the precise location and nature of cis-regulatory elements in the pre-mRNA sequence.

Exon definition further seems beneficial, as it prevents the accumulation of potentially toxic intron retention products during splicing regulation: using simulations and analytical calculations, we find that the sum of all intron retention products remains constant in the exon definition model if splicing is regulated by the AE recognition parameter p_{34} (Fig. 3A, red lines and Materials and Methods, Section 6). In these simulations, the degree of intron retention is solely determined by the recognition probabilities of the outer constitutive exons (p_{12} and p_{56} , see Materials and Methods). On the contrary, the intron definition mechanism inevitably leads to a strong accumulation of retention products during alternative splicing regulation, especially if the splice site recognition probabilities p_3 or p_4 are regulated separately (Fig. 3A, left and Materials and Methods, Section 6). In fact, in the intron definition model, pronounced switching from inclusion to skipping isoforms is only possible if p_3 and p_4 are concurrently regulated. However, even in this scenario, intron retention species account for $\geq$50% of the splice products during the splicing transition (Materials and Methods, Section 6).

Using analytical calculations, we confirm that exon modularity and suppression of intron retention also occur for pre-mRNAs containing more than three exons (see Materials and Methods, Section 7 and Discussion). This suggests that exon definition is generally beneficial from a regulatory point of view.

Exon definition modularizes splicing regulation

Our simulations predict that exon definition modularizes splicing control and prevents the accumulation of intron retention products. To confirm the predicted modularity of alternative splicing regulation, we compared the effects of point mutations located at 3’ and 5’ ends of the alternative exon. Since the exon functions as a module in the exon definition model, we expect that these mutations should have very similar effects on the abundance of splice products. We considered all mutations within a +/- 30 nt window around the 3' and 5' splice sites of exon 11. To account for mutation strength, we sorted mutations according to their effect on the AE recognition probability (p_{34}) in the best-fit model. Then, we plotted the experimentally measured splice isoform abundances as a function of the assigned mutation strength (Fig. 3B). As predicted by simulations of the exon definition model, the observed mutation-strength-dependent isoform changes are almost identical for 3’- and 5’-associated mutations. Furthermore, the measured isoform patterns quantitatively agree with simulations of an exon definition model, in which the AE recognition parameter p_{34} is systematically varied at otherwise constant recognition probabilities (Fig. 3B, second row). In contrast, corresponding simulations of the intron definition model completely fail to match the data (Fig. 3B, third row). In further support for the exon definition model, we observe highly similar flanking mutation effects not only around the alternative exon but also for the constitutive exon 12 (Fig. S4). In contrast, the intron definition model would predict congruence of mutation effects flanking a common intron, but this behavior is not supported by the experimental data (Fig. 3B and Fig. S4). These observations confirm that exon definition allows exons to function as dominant regulatory modules in alternative splicing control.

To further support that modular exons integrate regulation at the 3’ and 5’ splice site into a joint splicing outcome, we turned to the analysis of combined mutation effects. We reasoned that an exon definition model trained only on single point mutations should be able to quantitatively predict splicing effects of combined mutations. Therefore, we fitted the model to the subset of minigenes harboring only a single mutation, and predicted splicing outcomes of minigenes with a combination of two mutations. The exon definition model accurately predicted how two simultaneous mutations in the vicinity of the 3’ and 5’ splice site of exon 11 (each having a strong effect on splicing) jointly affect splicing outcomes (Fig. 3C, left panel). More generally, the exon definition model accurately predicted the combined outcome of any two mutations throughout the minigene (Fig. 3C, right panel). In contrast, a similarly trained intron definition model fails to correctly predict combined mutation effects (Fig. 3C, red dots).

Taken together, by analyzing sequence mutations, we find that integration of splice-regulatory signals in *RON* follows an exon definition scenario which has profound impact on the controllability of alternative splicing.

Exon definition prevents the accumulation of undesired intron retention products

An important observation in our splicing simulations of the exon definition model is that intron retention products remain constant if splicing is regulated at the alternative exon (by the AE recognition parameter p_{34}, Fig. 3A). In contrast, intron retention products inevitably accumulate during alternative splicing regulation in the intron definition model (Fig. 3A). To intuitively understand why intron and exon definition differentially affect retention products, consider discrete spliceosome binding configurations (Fig. 2B). If all six splice sites in the three-exon pre-mRNA are occupied by U1 and U2 snRNPs, the splicing outcome is exon inclusion for both mechanisms (Fig. 2B, I). In the next step, alternative splicing can be induced by reducing the recognition of one or both splice sites of the alternative exon. In exon definition, such regulation yields only exon skipping because the middle exon is always incompletely recognized and this impairs splicing of both introns (Fig. 2B, II-IV). In contrast, retention products accumulate in intron definition, as one of the introns remains defined and is therefore spliced (Fig. 2B, II-IV). Our model translates these qualitative arguments into continuous and quantitative predictions of splicing outcomes for five isoforms. For instance, it predicts that in intron definition, retention products strongly accumulate even if the recognition probability of both splice sites is reduced coordinately, e.g., to 50% ($p_3 = p_4 = 0.5$). This is due to the fact that combinatorial spliceosome binding to the 3' and 5' splice sites results in an equally distributed mixture of four binding configurations, two of which result in retention products (Fig. 2A, V). Hence, exon definition seems superior when compared to intron definition, as it prevents the accumulation of potentially deleterious retention products.

To verify that alternative splicing of *RON* exon 11 is controlled without intron retention, we compared the predictions of our exon definition model to the experimental data. To this end, splicing of the alternative exon was quantified for each point mutation using the PSI metric (percent spliced-in; PSI = AE inclusion/(AE inclusion + AE exclusion)) and then plotted against the corresponding total intron retention level, i.e., the sum of full, first and second intron retention isoforms (Fig. 3D, left panel). In line with the exon definition model, we observed that the majority of point mutations (red and blue dots) induce shifts in alternative splicing (PSI) at almost constant intron retention levels, when compared to wildtype (grey dots). Only a minority (< 2%) of mutations showed strong effects on intron retention, but these had at the same time only minor effects on the PSI.

These orthogonal changes in either exon inclusion or intron retention could be explained by simulations of the exon definition model, in which we randomly perturbed one of the splice site recognition probabilities, while sampling the others close to their reference value (Fig. 3D, middle panel). The model traces changes in PSI at constant retention levels back to altered splice site recognition of alternative exon 11 (red dots), whereas intron retention enhancement at constant PSI involves reduced recognition of the outer constitutive exons (blue dots, Materials

and Methods, Section 6). Consistently, we find in the experimental data that mutations with strong effect on intron retention map to the constitutive exons (Fig. 3D, left; blue dots), whereas mutations affecting PSI located to the AE (Fig. 3D, left; red dots). Simulations of the intron definition model fail to reconcile the data as the PSI cannot be modulated without accumulation of retention products (Fig. 3D, right panel; Materials and Methods, Section 6). In conclusion, modeling and comprehensive mutagenesis data suggest that alternative splicing is designed to prevent mis-splicing over a wide range of exon inclusion levels, likely due to an exon definition mechanism.

Discussion

Cellular regulatory networks should not only produce a certain outcome, but need to achieve it in highly precise and controllable fashion. Mathematical models are valuable tools to gain insights into design principles of cellular networks that ensure robustness and precision (*14-16*). To date, only a handful of mechanistic modeling studies on alternative splicing have been published which mainly focused on the quantification of mutation effects (*9,11,17*), studied the impact of co-transcriptional splicing (*10,18*) and analyzed cell-to-cell variability of the process (*19,20*). Here, we approach splicing regulation from a different angle and mechanistically describe how splice site recognition by the spliceosome shapes the splicing outcome. We systematically compare intron and exon definition mechanisms, and find that exon definition ensures robust, yet simple regulation of alternative splicing. Thereby, we gain general insights into the efficiency and controllability of splicing.

Using data-based modeling, we identify exon definition as the mechanism of *RON* exon 11 splicing. The prevalence of exon definition is surprising, given that exon 11 and its flanking introns are short, and therefore exon definition should be in competition with the simpler intron definition mechanism. Previous work showed that vertebrate exons flanked by short introns can switch to an intron definition mechanism if cross-exon spliceosome complexes are inhibited, e.g., by artificially lengthening the exon (*6*) or by the lack of exonic splice enhancer elements (*21*). Our data indicates that this switch does not occur in a natural context, and that exon definition is more efficient than intron definition in human cells (*22*). This suggests that exon definition does *not* merely serve an auxiliary role in productive spliceosome assembly across long introns, but may also be beneficial from a regulatory point of view.

Accordingly, we find that exon definition leads to a modularization of splicing regulation. Hence, regulation at one splice site of an exon is transferred to the other splice site, such that exons act as functional units. This has important consequences for the robustness and control of alternative splicing: Our simulations highlight that for pure intron definition splicing outcomes would be very distinct if splice-regulatory inputs affect spliceosome recruitment to the 3’ or the 5’ splice site of the alternative exon (Fig. 3). Furthermore, exon skipping may be difficult to achieve with intron definition, unless both splice sites are coordinately regulated (Fig. 3). Accordingly, a

global survey of *Drosophila* alternative splicing indicated that exons with short flanking introns (likely spliced by an intron definition mechanism) show a strong trend against exon skipping (*21*). In contrast, in the modular exon definition, inputs at the 3' and 5' splice site (or combinations thereof) produce the same splicing outcome. Such signal integration by exon definition is likely to be physiologically relevant, as RBPs frequently control alternative exons by binding nearby only one of the flanking splice sites (*23*). Arguably, a given RBP can repress or activate splicing depending on its binding position relative to an intron-exon junction (*23*). Our model does not exclude such a scenario, but solely predicts that the net effect an RBP has is integrated in a simple way with signals from other RBPs. Thereby, our mechanistic model may also explain why combined mutation effects on splicing outcomes can be accurately quantified using additive regression models when expressed as log-fold changes (*11*,*17*). In conclusion, exon definition allows for reliable splicing, even though alternative exons are typically influenced by a whole battery of distinct *cis*-regulatory elements (*4*,*5*).

Exon definition further prevents the accumulation of potentially non-functional intron retention products, and thereby improves the fidelity and efficiency of alternative splicing. In line with our observation that intron retention is difficult to achieve by an exon definition mechanism, human splice site mutations most often cause exon skipping and rarely result in intron retention (*6*). If retention occurs the mutations are typically located to short introns or affect terminal introns at the beginning or the end of a pre-mRNA which may be more prone to splicing by an intron definition mechanism (*6*). During granulocyte differentiation, intron retention is enhanced for dozens of genes as a means of active cellular regulation. Interestingly, this involves a switch from exon to intron definition, as splice factors which favor intron definition complexes are upregulated which promotes the retention of short introns with weak splice sites (*24*,*25*). Thus, intron definition indeed seems to promote retention. Our work transfers these earlier findings into a continuous and quantitative description of all splice products, and shows that intron retention is an inevitable consequence of the intron definition mechanism. This may explain why exon definition dominates for human exons, and raises the question how simple organisms with prevalent intron definition accurately control alternative splicing decisions.

In this work, we analyzed prototypical splicing unit consisting of three exons. Most human genes contain more than three exons, raising the question whether the described regulatory principles also apply for these more complex scenarios. In the Materials and Methods, we analyze an extended exon definition model containing more than three exons and show that the inclusion frequency (PSI) of each internal exon is solely determined by its own recognition probably (Materials and Methods, Section 7). Thus, the inclusion of an exon is regulated independently of the neighboring exons, i.e., each exon is targeted in a modular fashion. Importantly, in the multi-exon case, modularity therefore not only involves reliable signal integration on an exon, but also ensures insulation of this exon from other alternative splicing events. In similarity to the three-exon scenario, total intron retention is solely determined by the recognition probabilities of the two outer exons also for long pre-mRNAs containing multiple exons. Thus, inclusion levels and intron retention are again uncoupled, i.e., alternative splicing regulation occurs without the

accumulation of retention products. Taken together, this suggests that the regulatory benefits of exon definition described in this work continue to hold for longer pre-mRNAs.

Genome-wide sequencing indicates that ~80% of human exons are spliced co-transcriptionally while RNA polymerase is elongating the transcript (*7*). While this may affect the principles and efficiency of splicing regulation, several lines of evidence suggest that our post-transcriptional view well approximates co-transcriptional splicing regulation in human cells: Splicing of human introns occurs with a delay after nascent RNA synthesis (*26*) and begins only several kilobases after an intron-exon junction leaves the RNA polymerase complex, with the lag being especially pronounced for alternatively spliced exons (*27*). Given the median length of human introns and exons (145 and 1964 bp, respectively (*1*)), the splicing machinery thus likely generates splicing decisions based on sequence stretches containing multiple exons as we assumed here. Otherwise, it would be difficult to explain why neighboring human introns tend to be spliced concurrently (*27*,*28*) or even in inverse order relative to transcription (*23*,*27*,*28*). In fact, concurrent splicing of introns further supports exon definition, as this mechanism prevents the accumulation of partially spliced retention products (Fig. 3B). In simpler organisms, splicing is tightly coupled to the exit from RNAP (*27*,*29*). Thus, the kinetics of splicing may have co-evolved with mechanisms of splice decision making, slower kinetics being beneficial for exon definition and thus for precise alternative splicing.

Our modeling framework integrates and quantifies the effect of sequence mutations and knockdowns of *trans*-acting RBPs on spliceosome recruitment and splicing outcomes (Fig. 2 and Fig. S5). Thus, it constitutes a first step towards a comprehensive network model of splicing which mechanistically describes the integration of multiple splice-regulatory inputs into a net splicing outcome. In fact, we could successfully predict how multiple point mutations jointly control splicing outcomes (Fig. 3C), and the same type of predictions are possible for combined RBP knockdowns and combination of RBP knockdown and sequence mutations. Conceptually, the modeling framework resembles thermodynamic models of transcriptional gene regulation (*30*-*32*). However, for the case of splicing, regulation is more complex compared to transcription, as both the regulators (RNA-binding proteins) and the effectors (spliceosomes) show combinatorial binding to multiple sequence elements. Owing to this high level of complexity at multiple levels, we believe that mechanistic splicing models like the one we propose here will be essential to fully disentangle the intricate networks of splicing regulation.

References and Notes:

1. Y. Lee, D.C. Rio, Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* **84**, 291-323 (2015).

2. H.Y. Xiong, B. Alipanahi, L.J. Lee, H. Bretschneider, D. Merico, R.K. 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).

3. Z. Wang, C.B. Burge, Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802-13 (2008).
4. F.X.R. Suthandy, S. Ebersberger, L. Huang, A. Busch, M. Bach, H.S. Kang, J. Fallmann, D. Maticzka, R. Backofen, P.F. Stadler, K. Zarnack, M. Sattler, S. Legewie, J. König, In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res*. **28**, 699-713 (2018).
5. K.J. Hertel, Combinatorial control of exon recognition. *J. Biol. Chem.* **283**, 1211-5 (2008).
6. S.M. Berget, Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411-4 (1995).
7. L. De Conti, M. Baralle, E. Buratti, Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* **4**, 49-60 (2013).
8. S. Ke, L.A. Chasin, Context-depending splicing regulation: exon definition, co-occurring motif pairs and tissue specificity. *RNA Biology* **8**, 384-8 (2011).
9. M.A. Arias, A. Lubkin, L.A. Chasin, Splicing of designer exons informs a biophysical model for exon definition. *RNA* **21**, 213-29 (2015).
10. J. Davis-Turak, T.L. Johnson, A. Hoffmann, Mathematical modeling identifies potential gene structure determinants of co-transcriptional control of alternative pre-mRNA splicing. *Nucleic Acids Res.* **46**, 10598-10607 (2018).
11. S. Braun, M. Enculescu, S.T. Setty, M. Cortés-López, B.P. de Almeida, F.X.R. Sutandy, L. Schulz, A. Busch, M. Seiler, S. Ebersberger, N.L. Barbosa-Morais, S. Legewie, J. König, K. Zarnack, Decoding a cancer-relevant splicing decision in the *RON* proto-oncogene using high-throughput mutagenesis. *Nature Communications* **9**, 3315 (2018).
12. S. Lykke-Andersen, T.H. Jensen, Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16**, 665-77 (2015).
13. J. E. Braun, L. J. Friedman, J. Gelles, M. J. Moore, Synergistic assembly of human pre-spliceosomes across introns and exons. *Elife* **7**, e37751 (2018).
14. N. Blüthgen, S. Legewie, Robustness of signal transduction pathways. *Cell Mol Life Sci.* **70**, 2259-69 (2013).
15. J. Kamenz, T. Mihaljev, A. Kubis, S. Legewie, S. Hauf, Robust ordering of anaphase events by adaptive thresholds and competing degradation pathways. *Mol. Cell* **60**, 446-59 (2015).
16. M. Enculescu, C. Metzendorf, R. Sparla, M. Hahnel, J. Bode, M.U. Muckenthaler, S. Legewie, Modelling systemic iron regulation during dietary iron overload and acute

inflammation: Role of hepcidin-independent mechanisms. *PloS Comput. Biol.* **13**, e1005322 (2017).

17. 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 (2019).

18. J.C. Davis-Turak, K. Allison, M.N. Shokhirev, P. Ponomarenko, L.S. Tsimring, C.K. Glass, T.L. Johnson, A. Hoffmann, Considering the kinetics of mRNA synthesis in the analysis of the genome and epigenome reveals determinants of co-transcriptional splicing. *Nucleic Acids Res.* **43**, 699-707 (2015).

19. Z. Waks, A.M. Klein, P.A. Silver, Cell-to-cell variability of alternative RNA splicing. *Mol Syst. Biol.* **7**, 506 (2011).

20. U. Schmidt, E. Basyuk, M.C. Robert, M. Yoshida, J.P. Villemin, D. Auboeuf, S. Aitken, E. Bertrand, Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. *J. Cell Biol.* **193**, 819-29 (2011).

21. K.L. Fox-Walsh, Y. Dou, B.J. Lam, S.P. Hung, P.F. Baldi, K.J. Hertel, The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. *Natl. Acad. Sci. USA* **102**, 16176-81 (2005).

22. To further exclude splicing by an intron definition mechanism, we also considered mixed intron and exon definition models, in which only a subset of the three exons acts as a functional unit, whereas the remainder affects splicing already when partially defined. Interestingly, only the full exon definition model was consistent with the mutagenesis data, further suggesting that none of the two *RON* introns is spliced by a direct cross-int ron spliceosome complex (data not shown).

23. J.T. Witten, J. Ule, Understanding splicing regulation through RNA splicing maps. *Trends Genet.* **27**, 89-97 (2011).

24. J.J. Wong, W. Ritchie, O.A. Ebner, M. Selbach, J.W. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T.L. Khoo, C.G. Bailey, J. Holst, J.E. Rasko, Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**, 583-95 (2013).

25. A. G. Jacob, C. W. J. Smith, Intron retention as a component of regulated gene expression programs. *Hum. Genet.* **136**, 1043-1057 (2017).

26. T. Nojima, T. Gomes, M. Carmo-Fonseca, N.J. Proudfoot, Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat. Protoc.* **11**, 413-28 (2016).

27. H.L. Drexler, K. Choquet, L.S. Churchman, Human co-transcriptional splicing kinetics and coordination revealed by direct nascent RNA sequencing. https://www.biorxiv.org/content/10.1101/611020v2 (2019).

28. S.W. Kim, A.J. Taggart, C. Heintzelman, K.J. Cygan, C.G. Hull, J. Wang, B. Shrestha, W.G. Fairbrother, Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Res.* **45**, 9503-9513 (2017).

29. F.C. Oesterreich, L. Herzel, K. Straube, K. Hujer, J. Howard, K.M. Neugebauer, Splicing of nascent RNA Coincides with intron exit from RNA polymerase II. *Cell* **165**, 372-381 (2016).

30. L. Bintu, N.E. Buchler, H.G. Garcia, U. Gerland, T. Hwa, J. Kondev, R. Phillips, Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev*. **15**, 116-24 (2005).

31. G. Casanovas, A. Baneji, F. d'Alessio, M.U. Muckenthaler, S. Legewie, A multi-scale model of hepcidin promoter regulation reveals factors controlling systemic iron homeostasis. *PloS Comput. Biol.* **10**, e1003421 (2014).

32. P. Schulthess, A. Löffler, S. Vetter, L. Kreft, M. Schwarz, A. Braeuning, N. Blüthgen, Signal intergration by the CYP1A1 promoter – a quantitative study. *Nucleic Acids Res.* **43**, 5318-30 (2015).

Acknowledgments: The authors would like to thank the members of all participating labs for their support and discussion; **Funding:** This work was funded by a joint DFG grant (ZA 881/2-1 to K.Z., KO 4566/4-1 to J.K. and LE 3473/2-1 to S.L.). K.Z. was also supported by the LOEWE program Ubiquitin Networks (Ub-Net) of the State of Hesse (Germany) and the Deutsche Forschungsgemeinschaft (SFB902 B13). S.L. acknowledges support by the German Federal Ministry of Research (BMBF; e:bio junior group program, FKZ: 0316196). The Institute of Molecular Biology (IMB) gGmbH is funded by the Boehringer Ingelheim Foundation; **Author contributions:** ME, SL and JK conceived and designed research; ME and SL performed data analysis and modeling; SB and JK performed experiments; STS and KZ analyzed sequencing data; ME and SL wrote the paper with input from JK and KZ; **Competing interests:** Authors declare no competing interests; **Data and materials availability:** This study uses previously published sequencing data (*11*) that are available from ArrayExpress under the accession numbers E-MTAB-6216, E-MTAB-6217 (RNA-seq), and E-MTAB-6219 (DNA-seq).

Supplementary Materials:

Materials and Methods

Figures S1-S5

Table S1

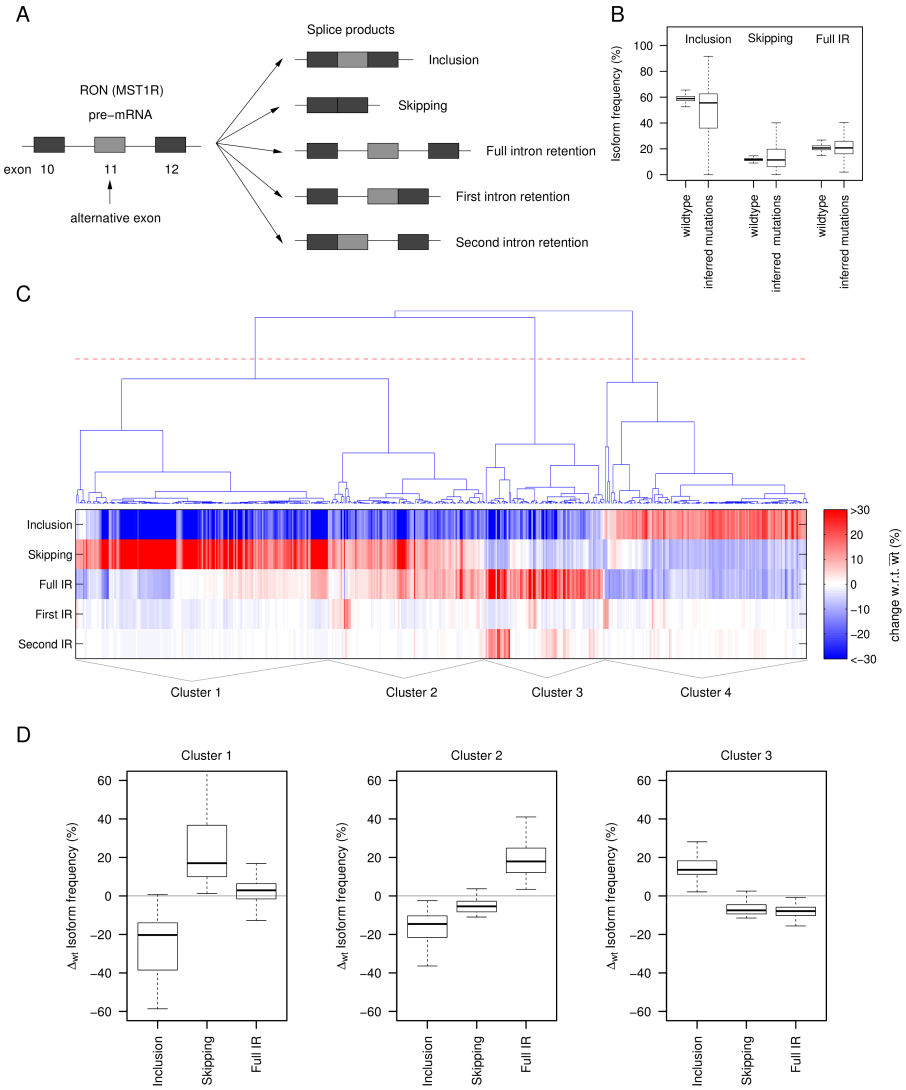

Fig. 1. Sequence mutations in a three-exon minigene containing *RON* AE exon 11 induce concerted changes in the distribution of splice isoforms. A The studied three-exon-minigene (704 bp) contains *RON* AE exon 11 and the complete adjacent introns and constitutive exons 10 and 12. Using next-generation sequencing, five different splice products were quantified (as % of

all splice products) for wildtype minigenes as well as for single point mutations (see Materials and Methods, Section 1 and (*11*)). **B** Point mutations induce strong changes in the splice isoform distribution, as visible by the much broader isoform frequency distributions (mut) compared to the population of ~500 unmutated wildtype (wt) minigenes. Full IR: full intron retention. **C** Heatmap of splice isoform difference between mutant and wildtype is plotted for 510 point mutations (columns) with a strong effect on the splicing (more than 10% change in at least one isoform frequency with respect to wildtype). Mutations are sorted using hierarchical clustering (Euclidean distance) and three main clusters are defined (using the red line in the dendrogram as a threshold). **D** Same data as in C, represented as boxplots summarizing the isoform distribution of each cluster for the three main isoforms. Mutations in clusters 1 and 3 induce anti-correlated changes in inclusion and skipping. In cluster 1, these changes are most pronounced in absolute terms, and intron retention is only slightly changed compared to wt. Cluster 3 shows weaker changes and altered intron retention, though in opposite direction. Mutations assigned to cluster 2 decrease both inclusion and skipping and simultaneously increase full intron retention.

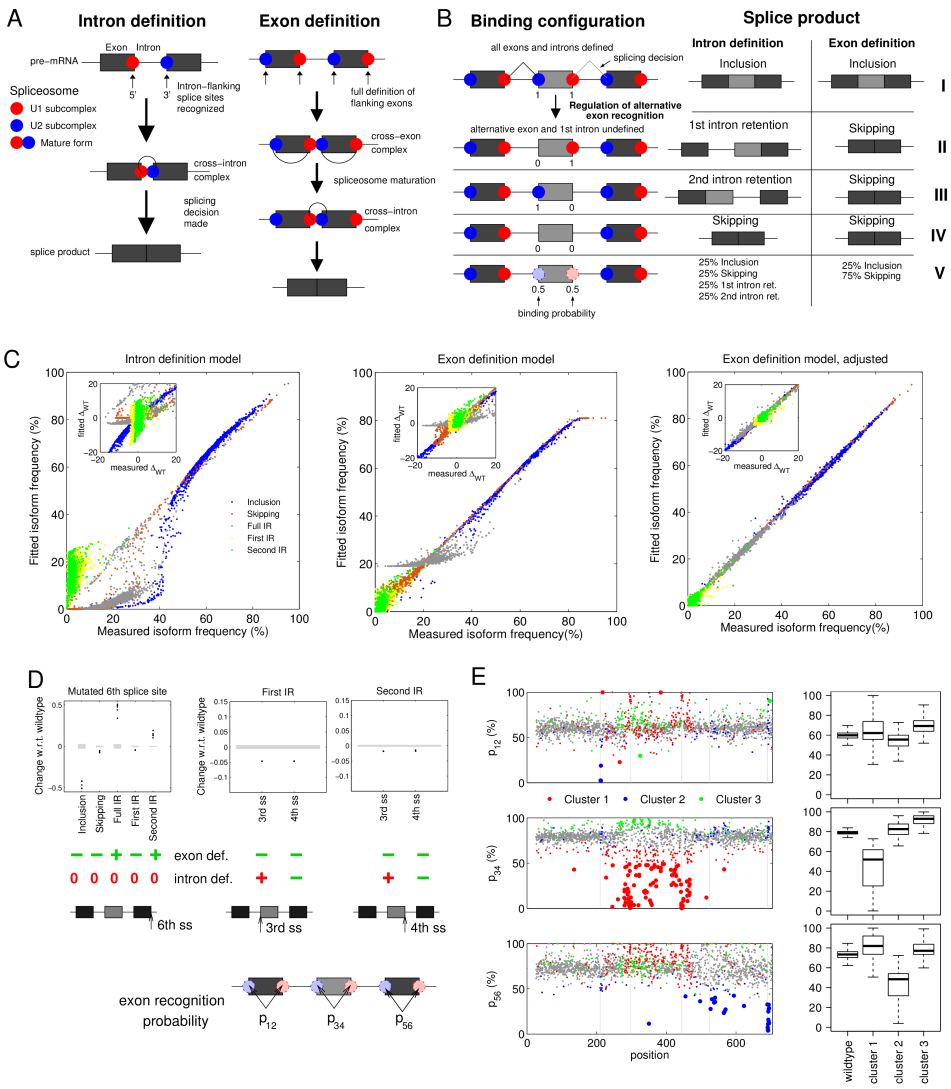

Fig. 2. The exon definition model quantitatively explains isoform changes in the mutagenesis screen. A Intron definition model: an intron is spliced out, as soon as its 3' and 5' ends splice sites are simultaneously bound by the U1 and U2 snRNP spliceosomal subcomplexes. Exon definition model: full definition of flanking exons is additionally required for the splicing of an intron, as transitory cross-exon complexes are involved in spliceosome maturation. **B** Different spliceosome binding configurations (I-IV) of the U1 and U2 subcomplexes may result in distinct splice products in the intron and exon definition models, respectively (see main text for details). In configuration V, binding at the 3' and 5' ends of the AE is assumed to take place with 50% probability, giving rise to a equimolar mixture of binding states I-IV. **C** The exon definition model (middle) shows a better quantitative agreement with the mutagenesis data when compared to the intron definition model (left), as judged by the scatter of

model fit against measurements of all five splice isoforms for 1854 point mutations. The performance of the exon definition model is further improved by allowing three global parameters (common to all point mutations) to model under-representation of long intron retention products due to metabolic instability and/or sequencing biases (right). The insets compare model fit (y-axis) vs. data (x-axis) as the difference in splicing outcomes between point mutations and wildtype (zero: no change relative to wt). In terms of directionality of changes, the exon definition model provides a better qualitative match to the measured mutation effects. **D** Defined point mutations in the 3rd, 4th and 6th splice sites (see bottom sketches) allow for a categorical discrimination between the intron and exon definition mechanisms. Shown are measured splice isoform differences (relative to wt) for minigenes harboring individual point mutations in the indicated splice sites (dots) alongside with the wildtype standard deviation (gray shadows). Directionalities of mutation effects according to the intron and exon definition models – as derived from analytical calculations (Material and Methods, Section 2) – are indicated below (green for matches with the data, red for contradicting results). **E** Landscapes and corresponding boxplots showing the exon recognition probabilities (p_{12} , p_{34} and p_{56} , see scheme) expressed as % recognition for point mutations along the minigene sequence (x-axis) according to the best-fit adjusted exon definition model. The dot color indicates to which cluster a mutation was assigned (see legend and Fig.1; mutations with weak effects, not included in clustering are plotted in gray). Mutations with a recognition probability shift of more than 20% relative to wildtype are highlighted in bold (only the strongest effect being highlighted for each mutation). Mutations in cluster 2 mainly affect the recognition probability of constitutive exons (p_{12} or p_{56}), while mutations in the other two clusters mainly affect alternative exon recognition (p_{23}), although in different direction and to a different extent.

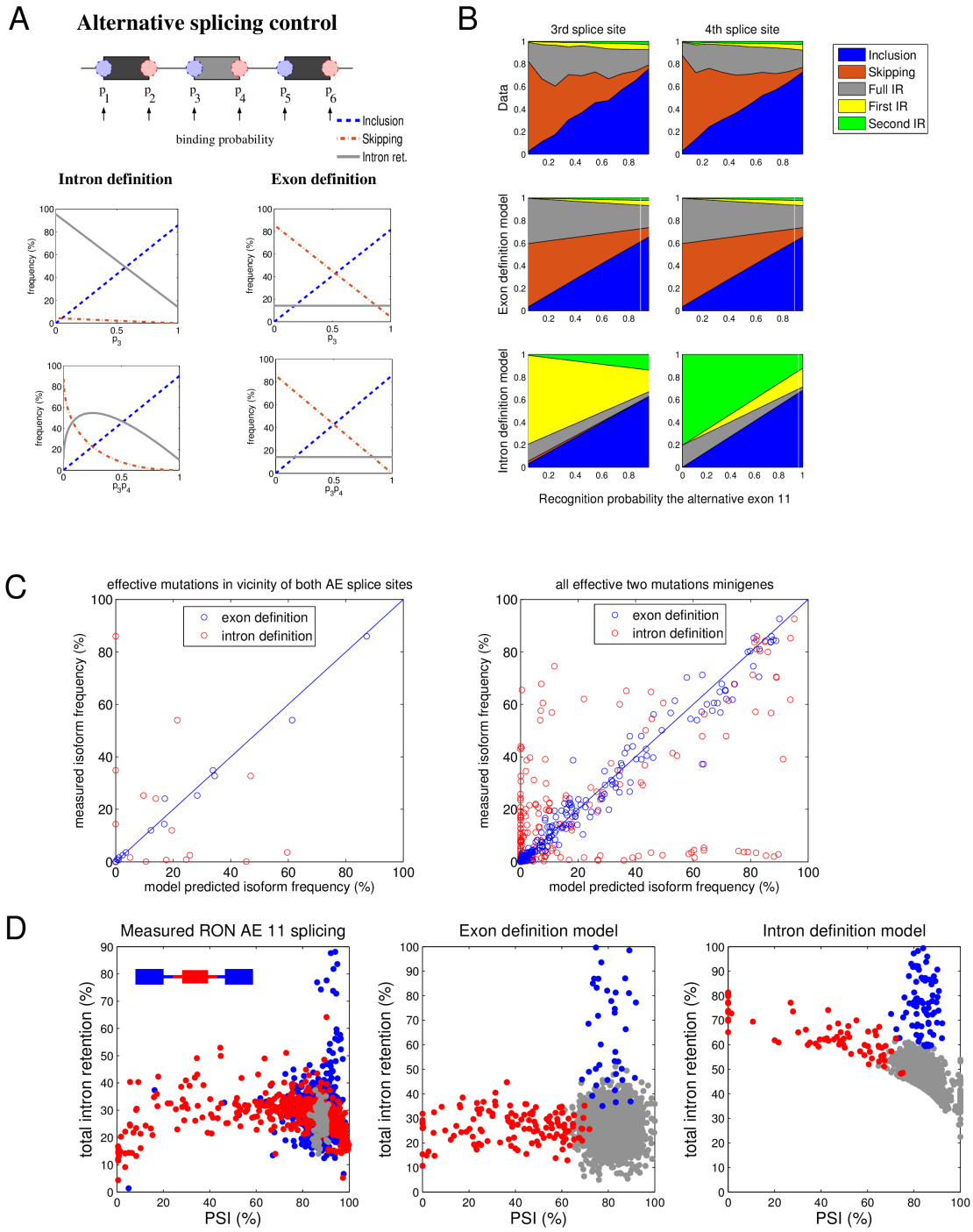

Fig. 3. Exon definition allows for modular control of alternative splicing and prevents accumulation of intron retention products. A Simulated splice product frequency of inclusion, skipping and total intron retention (sum of first, second and full IR isoforms) in response varying recognition of the splice sites flanking the alternative exon. Simulations of the intron and exon definition models are shown in the left and right columns, respectively. The upper plots show the splicing probabilities obtained when only the binding probability of the 3' splice site (p_3) is

varied (similar plots, are obtained for variation of p_4, not shown). The bottom plot displays the effects of concerted variation of the two recognition probabilities ($p_3 = p_4$, x-axis: product $p_3 p_4$). In the intron definition scenario, skipping is possible only if p_3 and p_4 are simultaneously regulated, and is accompanied by enhanced retention. For the exon definition model, separate or joint changes in p_3 and/or p_4 lead to a switch from skipping to inclusion, without accumulation of retention isoforms. **B** Mutagenesis data confirms the modular control of AE 11 splicing predicted by the exon definition model. Point mutations at positions in a +/- 30-nt window around the 5' (left column) or 3' (right column) splice sites of the AE 11 were selected and sorted according to their effect on the recognition probability p_{34} of the AE in the best-fit model (adjusted exon definition model). The measured changes in the splice isoform fractions with varying mutation strength (1st row) are similar for both splice sites, and agree with simulations of the exon definition model in which p_{34} is systematically varied (2nd row), but disagree with the intron definition model (3rd row). See also main text and Materials and Methods, Section 4. **C** An exon definition model trained on single point mutation data accurately predicts the abundance of five splice isoforms for combined mutations. Measured isoform frequencies in minigenes containing two mutations are plotted against values predicted by an exon definition model fitted only to single point mutation measurements. The left panel shows three combinations of mutations in a +/- 30 nt window around the 3' and 5' splice sites of the AE, and the right panel shows all 45 present combinations of two arbitrary mutations throughout the minigene. Only single mutations that induce strong changes were considered (sum of absolute changes in all five isoforms > 20%). See Material and Methods, Section 5 for details. **D** Alternative splicing occurs at low levels of intron retention in the mutagenesis data (left panel). Shown is the sum of all retention products as a function of the PSI-metric (PSI= AE inclusion / (AE inclusion + AE skipping)) which measures alternative splicing of exon 11. The distribution of wildtype minigenes is shown by grey dots and each colored dot represents a single point mutation. Mutations located to the outer constitutive (and adjacent intron halves) are highlighted in blue, whereas the red dots show corresponding mutation effects in or around the AE (see legend). The middle and left panels show 2,000 simulations of the exon and intron definition models, respectively. In these simulations, the splice site recognition parameters ($p_1 - p_6$) were randomly perturbed, one randomly chosen parameter being more affected than the others to mimic the experimentally measured PSI and intron retention values (see Materials and Methods, Section 6 for details). Only exon definition allows for alternative splicing at low retention levels.