Sequence determinants and evolution of constitutive and alternative splicing in yeast species

Dvir Schirman[1], Zohar Yakhini[2,3], Orna Dahan [1], Yitzhak Pilpel[1*]
1 - Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 7610001, Israel.
2- School of Computer Science, Herzliya Interdisciplinary Center, Herzliya 4610101 , Israel.
3 - Computer Science Department, Technion, Haifa 3200003, Israel.
* Corresponding Author: pilpel@weizmann.ac.il

Abstract

RNA splicing is a key process in eukaryotic gene expression. Most Intron-containing genes are constitutively spliced, hence efficient splicing of an intron is crucial for efficient gene expression. Here we use a large synthetic oligo library of ~20,000 variants to explore how different intronic sequence features affect splicing efficiency and mRNA expression levels in *S. cerevisiae*. Using a combinatorial design of synthetic introns we demonstrate how non-consensus splice site sequences affect splicing efficiency in each of the three splice sites. We then show that *S. cerevisiae* splicing machinery tends to select alternative 3' splice sites downstream of the original site, and we suggest that this tendency created a selective pressure, leading to the avoidance of cryptic splice site motifs near introns' 3' ends. We further use natural intronic sequences from other yeast species, whose splicing machineries have diverged to various extents, to show how intron architectures in the various species have been adapted to the organism's splicing machinery. We suggest that the observed tendency for cryptic splicing is a result of a loss of a specific splicing factor, U2AF1. Lastly, we show that synthetic sequences containing two introns give rise to alternative RNA isoforms in *S. cerevisiae*, exposing intronic features that control and facilitate alternative splicing. Our study reveals novel mechanisms by which introns are shaped in evolution to allow cells to regulate their transcriptome.

Introduction

RNA splicing has a major role in eukaryotic gene expression. During splicing, introns are removed from a pre-RNA molecule towards creation of a mature and functional mRNA. In human, splicing is central to gene expression, as a typical gene contains 8 introns (Lander et al. 2001), and these introns can be alternatively spliced to create different alternative splicing isoforms, with a potential to also contribute to proteomic diversity (Andreadis et al. 1987; Graveley 2001; Nilsen and Graveley 2010; Kalsotra and Cooper 2011). However, most introns are constitutively spliced (Chen et al. 2006; Ding and Elowitz 2019) and their contribution to gene expression is not through increasing proteome diversity. Nevertheless, because a pre-mRNA must undergo splicing to produce a functional mRNA, the efficiency of this process directly affects the efficiency of the overall gene expression process. Hence, regulation of constitutive splicing can be a mechanism to regulate gene expression (Gotic et al. 2016) and a target for evolution to act on (Frumkin et al. 2019).

The budding yeast *Saccharomyces cerevisiae*, like other *hemiascomycetous* fungi, has a low number of intron-containing genes compared to other eukaryotes (Stajich et al. 2007). Most of these genes have a single intron which is constitutively spliced. Yet, although they occupy a small part of the genome, these intron-containing genes are highly expressed and are common among cellular functions such as ribosomal genes (Parenteau et al. 2011). Hence, yeast cells tightly regulate the splicing process and devote a lot of cellular resources for its accurate execution.

Recent investigations of splicing efficiency in yeast focused on analyzing natural introns in the genome, either by using RNA-seq data (Oesterreich et al. 2016; Douglass et al. 2019; Xia 2019), by studying a library based on a single reporter gene containing natural introns from the

S. cerevisiae genome (Yofe et al. 2014), or by investigating intron sequence features and examining the evolution and conservation of natural introns (Schwartz et al. 2008; Hooks et al. 2014). Other studies have utilized large synthetic libraries to study how sequence features affect alternative splicing decisions (Rosenberg et al. 2015; Baeza-Centurion et al. 2019; Cheung et al. 2019; Mikl et al. 2019).

In this work, we systematically study cis-regulatory features that affect splicing efficiency of an intron by using a large synthetic oligonucleotide library. This enables a much larger scale exploration of intron features compared to existing studies in yeast. In addition, as opposed to previous library based studies of splicing, the present library is mostly focused on constitutive splicing regulation. This technique, based on on-array synthesis (LeProust et al. 2010), was used previously to explore different elements involved in regulation of transcription, translation, RNA stability, and other regulatory elements (Sharon et al. 2012; Kosuri et al. 2013; Shalem et al. 2015; Weingarten-Gabbay et al. 2016; Levy et al. 2017; Maricque et al. 2018). Here we designed and synthesized a library of approximately 20,000 oligos, each consisting of a unique intronic sequence aimed to explore a range of sequence determinants that may affect and regulate splicing efficiency. We then measured splicing efficiency of each oligo using targeted RNA sequencing. Using this oligo library we cover and explore many sequence features that can affect splicing, expanding far beyond the repertoire of natural introns.

Introns are defined by three sequence elements, the 5' donor site (5'SS), the branch site (BS), and the 3' acceptor site (3'SS). The mechanism of 5'SS and BS recognition by the splicing machinery in *S. cerevisiae* are well understood (Madhani and Guthrie 1994). However, the exact mechanism of interaction of the spliceosome with the 3'SS is not yet fully understood, as *S. cerevisiae* lacks a splicing factor which is present and crucial for 3'SS recognition in higher eukaryotes (U2AF1) (Wu et al. 1999). Hence it was suggested that the 3'SS is recognized

through a scanning mechanism and that any HAG (i.e. [A/C/T]AG) site could be recognized (Smith et al. 1993). However, a vast majority of 3'SS of *S. cerevisiae* natural introns are in fact YAG (i.e. [C/T]AG), just like observed in higher eukaryotes (Wilkinson et al. 2020).

Working with a synthetic library of introns allows us to also study the evolution of introns and their splicing. In particular, natural introns from 11 different yeast species were incorporated in our library design . This enabled us to examine how intron architecture co-evolves with changes in the splicing machinery. Specifically we compare how the *S. cerevisiae* splicing machinery splices introns coming from species with or without the U2AF1 splicing factor encoded in their genomes. Additionally we show that *S. cerevisiae* has a tendency to select alternative downstream 3' splice sites and produce cryptic splice isoforms. We suggest that this is related to the loss of U2AF1, and that it has shaped the evolution of the intron architecture.

Lastly, we examine the potential of the budding yeast to feature alternative splicing. In *S. cerevisiae*, a vast majority of intron-containing genes have a single intron, and there are only a handful of examples of regulated alternative splicing in this organism (Howe et al. 2003; Grund et al. 2008; Juneau et al. 2009; Hossain et al. 2011; Meyer et al. 2011). We examine the extent to which the splicing machinery produces multiple splice isoforms when given a new synthetic two-intron gene. Meaning, how easy it is for *S. cerevisiae* to produce alternative splicing if the necessary information is embedded within genes?

Results

<u>High-throughput splicing efficiency measurements of thousands of synthetic introns</u>

To explore how the intron architecture affects splicing efficiency, we designed a synthetic oligonucleotide library (LeProust et al. 2010) of 18,061 variants. All the oligonucleotides were cloned into the same location inside a synthetic non-coding gene that was then integrated into

the yeast genome. We chose to explore splicing in the context of a non-coding gene, to avoid any differences between variants that might result from differences in translation.

Each designed oligo consists of fixed sequences for amplification and cloning, a unique 12nt barcode, and a 158 nt-long sequence that contains a unique intron design. All synthetic intron sequences were introduced into a mutated version of the natural *S. cerevisiae* gene *MUD1* lacking any ATG codon in all reading frames. The expression of the mutated *MUD*1 is driven by a synthetic promoter that was chosen from an existing promoter library (Sharon et al. 2012, 2014) based on its high expression and low noise characteristics. The entire intron-containing gene library was integrated into the YBR209W open reading frame in *S. cerevisiae* genome using a high-throughput Cre-Lox based method (Levy et al. 2015)). A schematic description of oligos structure and library creation is shown in Figure 1.

Each intron design in the library is characterized by the sequence of its three functional sites (5'SS, BS, 3'SS), its length, the distance between its BS and 3'SS, and by the length of a short U-rich element upstream to the 3'SS.

The library was composed of four major subsets , and a fifth set of negative control variants (see Table 1). The first subset represents a combinatorial design introducing different splice site sequences with their exact sequence as observed in the genome, on the background of the *MUD1* derived gene. Introns were created with different lengths and different BS-to-3'SS length that represent the length characteristics of introns from *S. cerevisiae* non-ribosomal genes. In each oligo, an intron was created by replacing the background sequence at positions 6-11 with a 5'SS sequence, and then according to the choice of intron length, and BS-to-3'SS a BS sequence and a 3'SS were placed instead of the background sequence at corresponding positions (see Table S1). In addition, three versions of short poly uracil sequence were inserted upstream to the 3'SS. A second subset of the library was composed of introns that naturally

occur in *S. cerevisiae* and in other yeast species. The sequences of these introns were inserted into the 158nt synthetic oligo, replacing the existing background sequence at its 5' end (introns longer than 158 nucleotides were not used for this set). The third subset was based on perturbations to the two former subsets by introducing mutations to the genome-observed splice site sequences. Splice site mutations to synthetic intron variants were introduced only to specific intron length and BS-to-3'SS length (89 and 30 nucleotides respectively). Mutated variants were created for each of the individual three sites separately, and for all four possible combinations of the three sites. This subset also included an additional set of synthetic variants with a different background sequence that was created by introducing only consensus splice site sequences at varying length properties, within different background sequences, similar to the first synthetic subset. Lastly, the fourth subset of variants was composed of designs with two short introns, one next to the other, separated by a short exon. The intron sequences used for this subset were natural introns taken from the *S. cerevisiae* and *S. pombe* genomes that are short enough to fit with another intron inside the 158 oligo, together with 5 short synthetic variants. This last subset enables us to study the potential of the *S. cerevisiae* splicing machinery to process genes with multiple introns and to produce alternative splice variants. In addition to the above four subsets a set of negative control variants was created by introducing mock randomly chosen sequences instead of the three splice sites, while using the same design principles as the first combinatorial subset, creating negative control variants with variable distances between the mock splice sites. The same randomly chosen mock sites were used for all the variants in this subset.

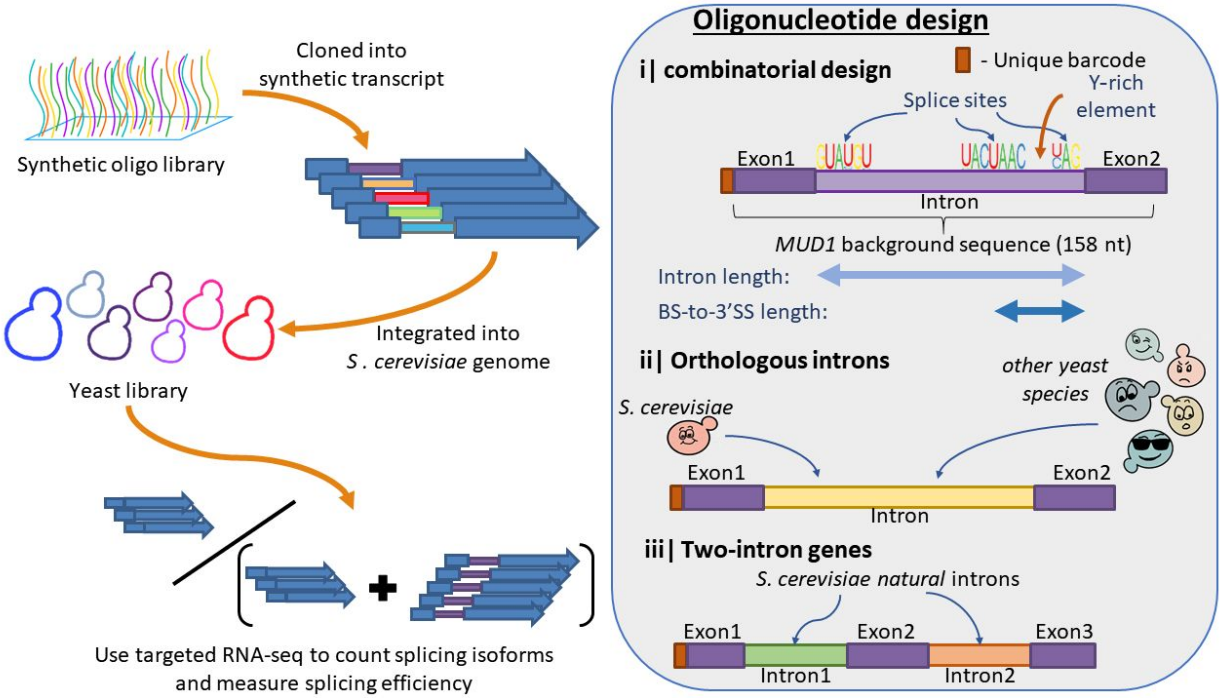


Figure 1 - A designed synthetic intron library in budding yeast

A large oligonucleotide library of designed introns was synthesized and cloned into a synthetic non-coding gene. The gene was then integrated into the budding yeast *S. cerevisiae* genome, to produce a pooled yeast library. Then splicing efficiency was measured using targeted RNAseq of the intronic region, identification of RNAseq reads according to the barcode, and of spliced isoforms using alignment of exon-intron and exon-exon junctions. Inset: oligonucleotide design strategy - All oligos were identified using a unique 12nt random barcode at their 5' end; i) Synthetic introns based on a combinatorial design representing different splice site sequences and other intronic features; ii) A set of natural introns from S. cerevisiae and other 10 yeast species was introduced into the library; iii) a set of synthetic two-intron genes produced by pairing together short intron sequences and placing an exon between them.

The splicing efficiency of each variant was measured using targeted (PCR based) RNA sequencing of the library's variable region. The sequence amplicon that was deep-sequenced included both the unique barcode of each intron design and its entire variable region, in either its unspliced or spliced forms. This allowed us to calculate splicing efficiency for each intron design. Shortly, each variant was identified by its unique barcode, and then the relative abundances of the spliced and unspliced isoforms were determined by aligning exon-intron and exon-exon junction sequences against the RNAseq reads. The splicing efficiency of a design was defined as the ratio between the spliced isoform abundance and total RNA abundance of

this design. If for a specific RNA read, neither an unspliced isoform nor an intended spliced isoform of the designed intron were identified, we searched for an mRNA isoform that might have been a result of a novel unintended splicing event (for details see Material and Methods).

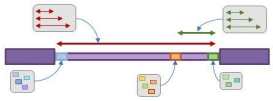
Subset number	Description	Graphical description	Number of variants
1	Synthetic combinatorially designed introns		4,713
2	Natural introns' sequences from 11 additional yeast species		1,173
3	Synthetic introns with splice sites mutations		4,505
	Synthetic introns with consensus splice sites and different background sequences		1,377
	Natural introns with mutated splice sites		1,328
4	Two-intron designs		823
5	Negative control		4,142

Table 1 - Summary of the different subsets composing the library and the number of variants in each of them

<u>Synthetic introns are successfully spliced within the genomic construct</u>

We used targeted RNA sequencing to measure splicing efficiencies, 99.43% of the library's variants were identified using this process. 13,096 variants in the library were designed with a single intron (The remaining 4,695 are negative control variants or two-intron variants). For

33.5% of these single intron variants, we observed the designed splice isoform with median splicing efficiency of 0.428 (on a scale from 0 to 1). For comparison, of the 4,142 variants that were designed to serve as negative controls (as they miss regulatory sites for splicing), only 1.9% yielded a spliced isoform and the median splicing efficiency there was 0.039. In addition, when examining the natural introns of *S. cerevisiae* that were included in our library, we saw that 84% of them yielded spliced isoforms, with median splicing efficiency of 0.675. We examined the total splicing efficiency distribution (i.e. splicing of any possible intron including cryptic introns) for all the variants that present greater than zero splicing efficiency (We chose to ignore variants with zero splicing efficiency as most of them are a result of non-functional intron designs). We observe a bi-modal distribution of splicing efficiency, when most variants are either mostly spliced, or rarely spliced (Fig. S1A). This suggests that the splicing machinery acts mostly as a binary switch.

To verify that the introduction of a 12nt barcode upstream of the intron does not have a significant effect on splicing efficiency, we attached four different barcodes to each of 517 randomly chosen designs. We then computed the variance in splicing efficiency between each quartet of barcoded designs (considering only designs with non-zero splicing efficiency) and compared it to the variance obtained with random quartets of barcodes that do not belong to the same design. Reassuringly we see that the mean variance among the correct quartets is much lower than the mean variance of each of 10^4 randomly shuffled variants quartets (Fig S1B). We further examine the pairwise correlations between splicing efficiency of pairs of library variants that share the same sequence yet different barcodes (Fig S1C). Reassuringly, we found a significant positive correlation between pairs of variants with different barcodes (mean Pearson correlation, r=0.76). Nevertheless we notice that in ~33% of the cases we observe significantly different splicing efficiency values between variants that differ only in their barcode, suggesting

that there is some effect for the barcode - a random exonic 12 nucleotides sequence, located slightly upstream to the 5' end of the intron. Yet, This result indicates that the barcode choice exerts at most a low effect on splicing efficiency. All together, these results establish the validity of our system as an *in vivo* quantitative splicing assay.

Splicing efficiency is positively correlated with RNA abundance

When examining total RNA abundance, we see a significant positive correlation between total splicing efficiency (i.e. a sum over all spliced isoforms for each variant) and total RNA abundance (i.e. summed level of unspliced and spliced isoforms)(Fig 2A). Since the calculation for splicing efficiency is dependent on total RNA abundance per design, we compare this result to a random set taken from the same distribution. No correlation was observed in the random set (Fig. S1D). This observed correlation between splicing efficiency and total RNA abundance per design is intriguing. We note that since this correlation is obtained with synthetic transcripts that were not selected in evolution to regulate their gene expression through splicing, it suggests a molecular mechanism that may be at work. This observed positive correlation might be explained by effects of splicing on nuclear export (Zhou et al. 2000), or on RNA stability (Wang et al. 2007), that can both enhance steady-state RNA levels per gene. Additionally, it was shown that splicing and transcription occur simultaneously (Oesterreich et al. 2016)

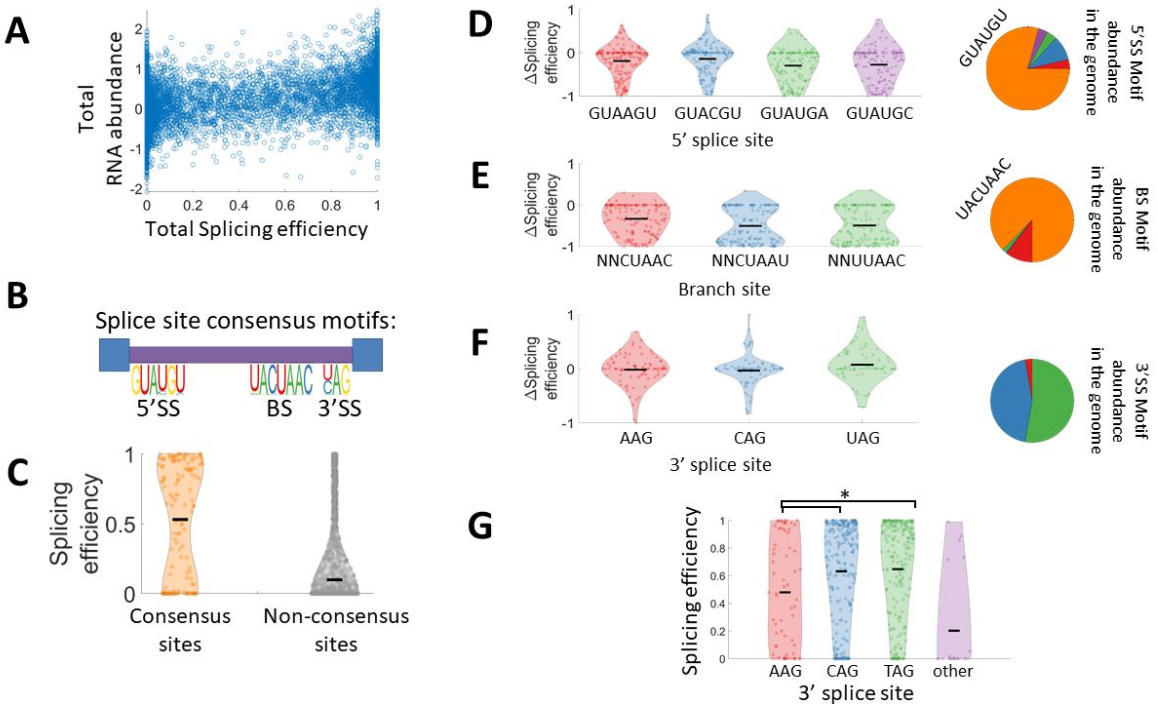


Figure 2 - A set of combinatorially designed synthetic introns elucidate splice sites' variants contribution to splicing efficiency

A. Scatter plot of the total RNA abundance and total splicing efficiency shows a significant positive correlation between RNA level and splicing efficiency (Pearson r=0.47 p-value<10^{-100}). **B.** Splice site motifs for *S. cerevisiae* introns as determined by their frequency in *S. cerevisiae* natural introns, one can notice a dominant consensus sequence for the 5'SS and BS, and two consensus sequences for the 3'SS. **C.** Distribution of splicing efficiency for synthetic intron variants with consensus splice sites (orange), is significantly higher than splicing efficiency distribution for non-consensus splice sites (grey) (two sample t-test, p-value<10^{-80}). Black horizontal line represents mean splicing efficiency. **D-F.** Effect of non-consensus splice site sequences on splicing efficiency. Violin plots represent the distribution of the difference in splicing efficiency between a variant with a single non-consensus splice site, to a corresponding consensus sites variant which is identical in any other parameter. For the 3'SS since there are two consensus sequences, AAG variants were compared against the average of the two consensus variants, and CAG/UAG variants were compared against the other consensus variant. Pie charts show the relative abundance of each splice site in *S. cerevisiae* genome (orange slice represents the consensus site, and other colors correspond to the colors in the violin plots). In the case of the BS, NNCUAAC non-consensus variants represent all sequences that fit this template but different from the consensus sequence UACUAAC. note that for the BS pie chart in (E) the blue portion representing NNCUAAU variants is too small to be visible. **G.** Splicing efficiency distribution only for the natural introns set considering introns with consensus 5'SS and BS, and binned according to different 3'SS (AAG introns' splicing efficiency are significantly lower than CAG/TAG introns, t-test p-value<0.003)

Combinatorial design of introns elucidates features contributing to splicing efficiency

We next analyzed the set of synthetic introns created by combinatorial design of different splice site sequences and length properties. As expected we noticed that introns that contain the consensus splice site sequence in all three splice sites are, as a population, better spliced than introns with at least one non-consensus splice site (Fig. 2B,C). Next, for each of the non-consensus splice site variants we examined how it affects splicing efficiency by analyzing variants with a single non-consensus splice site, and comparing their splicing efficiency to the corresponding design with consensus splice sites and otherwise identical sequence (Fig. 2D-F). We notice that almost all non-consensus branch site sequences result in much lower splicing efficiency, although they all contain the catalytic A residue at position 6. Indicating that out of all three functional sites, the branch site is most crucial for efficient splicing (Fig. 2E). On the other hand, in the 5'SS, while on average the non-consensus variants are spliced less efficiently we do observe a substantial number of variants that are spliced better than the corresponding variant with consensus site (Fig. 2D), we also notice that for two of the splice site variants, lower splicing efficiency is observed only for longer introns (Fig S2A). For the 3'SS we see that there is no measurable difference in splicing efficiency between the three variants found in the genome, although two of them are significantly more abundant than the third (Fig. 2F). The fact that the AAG 3'SS variants are spliced as well as the two YAG 3'SS variants is surprising due to the fact that ~95% of introns in all eukaryotes use a YAG 3'SS (Wilkinson et al. 2020). However, when considering the set of natural introns with consensus splice sites at their 5'SS and BS we notice that introns that utilize AAG as their 3'SS are spliced less efficiently, suggesting that in natural intron sequences there is embedded information that disfavors AAG as a 3'SS (Fig. 2G). Additionally, we used a set of variants with random mutations in their splice sites (with fixed length properties), to analyze the effect on splicing efficiency of all possible single nucleotide

mutations in the three splice sites, and this analysis replicated the results observed for the splice sites variants found in the genome (Fig. S2C). We have also analysed the effect of predicted secondary structure on splicing efficiency but found only borderline effects.

Next, we examined how other intron features can affect splicing efficiency. In addition to the splice sites, an intron is characterized by a poly-pyrimidine tract upstream to the 3'SS (Coolidge et al. 1997), However, it was noticed that in yeast a weaker feature is observed, and it is characterized by short uracil-rich sequence instead of pyrimidine (U or C) (Patterson and Guthrie 1991; Schwartz et al. 2008). Using our library we examine the effect of uracil rich sequences upstream to the 3'SS, by binning all the variants that utilize consensus splice site sequences, according to their U content in a 20nt window upstream to the 3'SS, and then comparing the splicing efficiency distribution in each bin (Fig. 3A). A striking pattern of correlation between higher U content in this window and increased splicing efficiency emerges. To demonstrate that the observed effect is specific to uracil enrichment and not to pyrimidine enrichment, we repeat the same analysis by binning according to Y content, considering only variants with well balanced U and C composition. We observe no correlation between Y enrichment and splicing efficiency in this setup (Fig S3A). This result is the first experimental evidence that *S. cerevisiae* splicing machinery is specifically affected by a poly uracil tract, as opposed to other eukaryotes (Coolidge et al. 1997).

Another feature that was identified as important for efficient splicing is the distance between the BS and 3'SS (Luukkonen and Séraphin 1997; Neuvéglise et al. 2011; Hooks et al. 2014). This feature is also highlighted by the recent high resolution cryo-EM structural analysis of the spliceosome when it was demonstrated that at the transition between branching conformation to the exon-ligation conformation (C to C*) the BS is removed from the spliceosome catalytic core, to allow space in the active site for the docking of the 3'SS. This transition requires a minimal

distance between the BS and the 3'SS (Wilkinson et al. 2020; Fica et al. 2017; Yan et al. 2017; Horowitz 2012). By binning all the variants that utilize consensus splice site sequences according to their BS-to-3'SS length we observe that a distance of 20-30 nt is optimal in terms of splicing efficiency (Fig. 3B). This result suggests a slightly longer optimal distance than the 13-22 nt distance that was observed in a previous work that used a single intron with varying 3'SS positions (Luukkonen and Séraphin 1997).

Previous studies have associated other intronic features with splicing efficiency such as, intron length (Wieringa et al. 1984; Dewey et al. 2006), and intronic GC content (Neuvéglise et al. 2011; Wong et al. 2013; Yofe et al. 2014; Mordstein et al. 2020). The data from the current library significantly supports the effect of intronic GC content. Specifically, we observed that splicing efficiency decreases with increasing GC content (Fig S3B). As for intron length, we do not observe a specific length that is spliced more efficiently (Fig S3C). We note though that introns taken for this library were bounded by a length of 158nt and the distribution of intron lengths represented in this library represent the length distribution of introns from non-ribosomal genes in *S. cerevisiae*. Introns from ribosomal genes are longer (mean intron length of ~400 nt). These lengths are not represented in this work. Thus we cannot exclude the possibility that intron length does affect splicing efficiency, if such dependence affects only longer introns.

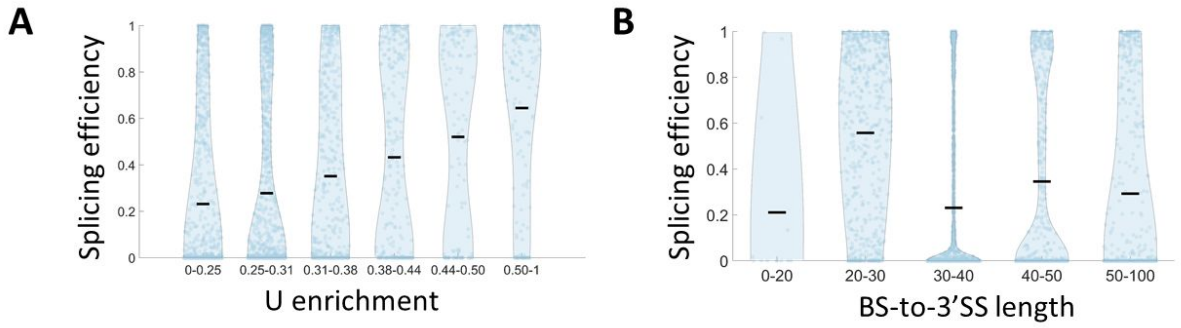


Figure 3 - BS-to-3'SS region effects on splicing efficiency
A. Splicing efficiency distribution for library variants utilizing consensus splice sites' sequences, is binned according to poly uracil tract enrichment, which is calculated as the U content at a window of 20 nucleotides upstream to the 3'SS. We see a significant positive correlation,

compared to a non-significant correlation for Y-rich elements (Fig S3A) (Pearson r=0.95 p-value=$4 \cdot 10^{-3}$). **B.** Splicing efficiency distribution for library variants utilizing consensus splice sites' sequences, is binned according to BS-to-3'SS length. We notice that splicing efficiency is significantly higher for length of 20-30 nt compared to all other bins (t-test, p-value<10^{-4}).

Cryptic splicing events drive intron evolution

Up to this point, we have focused on "intended" splicing events. That is, successful splicing of the designed intron of each variant in the library. However, our designed constructs might also result in cryptic splicing isoforms, different from the intended ones. To identify such splicing events, for each variant, we looked for cryptic spliced isoforms by aligning the RNAseq reads to the full unspliced sequence, allowing large gaps in the alignment. A long uninterrupted gap in the RNA read is potentially a spliced intron, and if at least one of its ends is not found at the designed ends of this variant we label it as a "cryptic intron".

We found cryptic splice isoforms in 25.2% of the variants all the library variants designed to have a single intended intron with a median splicing efficiency of 0.038 for the cryptic splice isoforms. Note that this is significantly lower than a median splicing efficiency of 0.428 for the designed splice isoforms. We then studied the location of the cryptic intron ends relative to the intended intron ends, and found that 87% of them have the same 5'SS as the designed intron, while only 1% of them have the same 3'SS (Fig. 4A). Further, most cryptic splicing occurs through 3'SS sites that are downstream, but not upstream to the designated 3'SS (Fig. 4A). This observation suggests that a vast majority of cryptic splicing events are a result of utilization of the canonical 5'SS with an alternative 3'SS during the splicing process. We acknowledge the possibility that due to our amplicon sequencing based method, there is a lower chance to detect putative upstream alternative 5'SS selection, and that alternative 5'SS might possibly result in unfinished splicing intermediant product (Harigaya and Parker 2012), which also wouldn't be detected by our method.

Since we observe cryptic isoforms as a result of alternative 3'SS utilization we inspect how the designed 3'SS affects the levels of cryptic isoforms. We consider only variants with consensus sites in their 5'SS and BS. As was shown for the designed isoforms (Fig. 2F), we don't observe a difference in cryptic isoforms levels between different 3'SS sequence variants, for synthetic introns (Fig. S2B). Importantly, however, we do observe a significant increase in cryptic isoforms levels for natural introns utilizing AAG 3'SS compared to the two common 3'SS sites (Fig. 4B). This suggests that AAG sites are only seldomly found in the yeast genome because they may lead to higher unintended alternative 3'SS utilization.

Previous work has also found alternative 3' splice site usage events in *S. cerevisiae* and suggested that the 3'SS choice can be explained by local RNA secondary structure at the original 3'SS (Plass et al. 2012). To examine this suggestion, using our synthetic introns data, we considered the distribution of the predicted RNA free energy (ΔG) at a window of 30 nucleotides around the designed 3'SS. We found that spliced variants with no cryptic splicing have more open predicted structures at their 3'SS compared to spliced variants with cryptic splicing, and to a greater extent than unspliced variants with cryptic introns (Fig. 4C). When studying the active alternative 3'SS we see that 70% of the alternative isoforms are spliced at one of the three sequence motifs found in the genome ([U/C/A]AG) (Fig. 4D) and that 68.5% of them are spliced at the first downstream occurrence of this 3'SS motif after the designed 3'SS. Those isoforms that are spliced at the first downstream 3'SS motif, are more efficiently spliced than other cryptic splice isoforms (Fig. 4E).

The observation that the *S. cerevisiae* splicing machinery can easily misidentify 3'SS leads us to hypothesize that mechanisms to avoid such cryptic splicing events must exist, as these events can result in frameshifts and in premature stop codon occurrences. Hence, we checked if we observe a selection against 3'SS motifs near the 3' end of natural introns in the genome. We

registered all *S. cerevisiae* introns at their 3' end and calculated the frequency of the two dominant 3'SS motifs ([C/T]AG) around introns' end. Indeed we found a depletion of these motifs at a window of -50 to +30 around introns' end compared to a null model based on 1000 random sets of genomic loci (Fig. 4F).

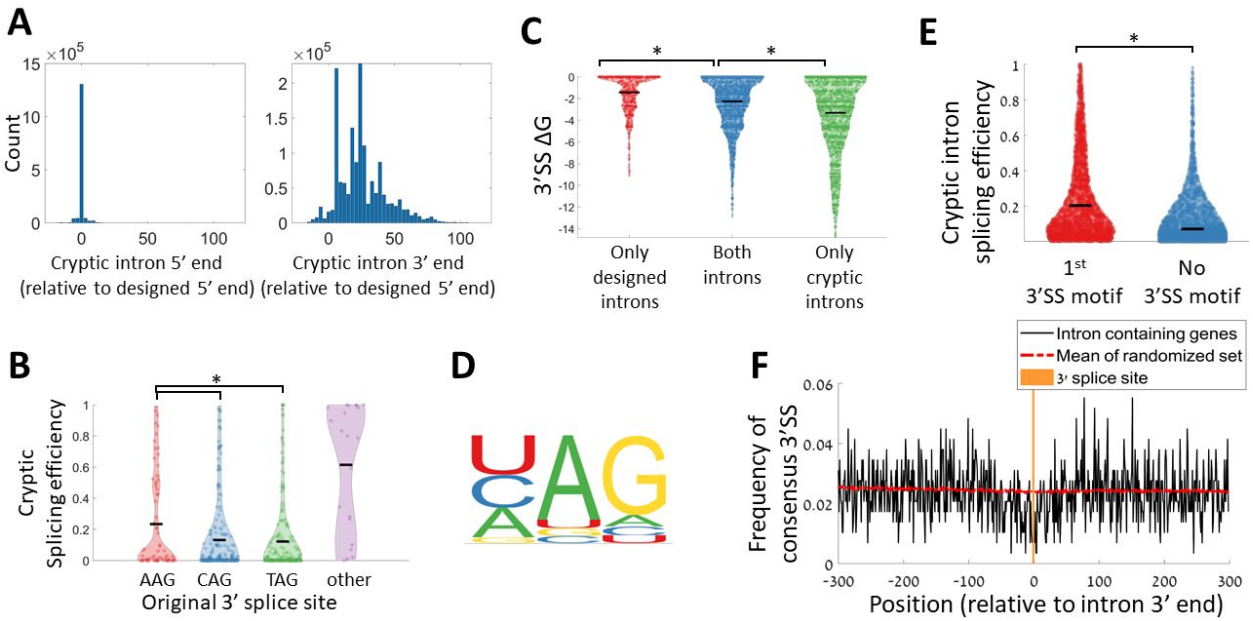


Figure 4 - Cryptic splice isoforms are produced due to selection of alternative 3'SS
A. Relative splice site position distribution for cryptic introns (relative to the designed splice site position), for the 5' splice site (left), and the 3' splice site (right). **B.** Distribution of cryptic isoforms splicing efficiency for the set of natural introns, binned according to the original intron's 3'SS. Introns utilizing AAG 3'SS significantly result in higher cryptic splicing efficiency values (t-test p-value<0.005) **C.** Distribution of predicted ΔG values at a window of 30 nucleotides around the designed 3'SS for variants with designed splice isoform and no cryptic splice isoform (red), variants with both designed and cryptic splice isoform (blue), and variants with only cryptic splice isoform (green). The difference between all three distributions is significant (t-test, p-value < 10^{-18}). **D.** Sequence motif of the 3'SS for all cryptic intron isoforms detected. **E.** The distribution of splicing efficiency for cryptic splice isoforms, binned according to isoforms in which the cryptic intron is spliced at the first appearance of a 3'SS motif (red), or isoforms for which the last 3 nucleotides of the introns are not a 3'SS motif. (t-test, p-value<10^{-100}). **F.** 3'SS motif avoidance pattern around introns' 3' end in the *S. cerevisiae* genome. All *S. cerevisiae* intron-containing genes were registered according to the 3' end of their intron. The black line presents the frequency of TAG/CAG motif for each position. The red dashed line presents the expected frequency by averaging the motif frequencies over 1000 sets of sequences registered according to random positions inside coding genes.

Co-evolution of the splicing machinery and intron architecture across yeast species

Our system allows us to introduce any short intron sequence into the *S. cerevisiae* genome. This gives us the opportunity to study the evolution of intron architecture by introducing introns from other yeast species and observing how well they are spliced in our system. We first introduced all the naturally occurring introns from *S. cerevisiae* genome that can fit in our oligonucleotide design length constraint. For each intron we inserted the full length of the intron flanked by 5 exonic nucleotides from each end. This sequence was inserted on the background of the standard *MUD1* derived background sequence of the library, at its 5' end. Hence, the length limit for an intron was 148 nucleotides, amounting to 149 introns out of 299 in this species. It should be noted that this limit on intron length forces us to use only introns from non-ribosomal genes in our library, as all the introns in ribosomal genes in *S. cerevisiae* are significantly longer (mean length of ~400 nucleotides).

Next, for each natural *S. cerevisiae* intron, we included in our library introns from ortholog genes from a set of 10 other yeast species, according to orthology identified by (Hooks et al. 2014). We found that most *S. cerevisiae* endogenous introns are spliced in our system (85.5%). Interestingly, introns from most of the other species are, typically, also spliced at similar efficiencies (Fig. 5A). Furthermore, we compare the splicing efficiency of each intron to its corresponding ortholog intron in *S. cerevisiae*, and define ΔSE as the difference in splicing efficiency for ortholog introns of the same gene compared to *S. cerevisiae* . We found that many of the non *S. cerevisiae* introns are spliced better than their *S. cerevisiae* orthologs (Fig. 5B), suggesting that *S. cerevisiae* introns are not specifically optimized for high splicing efficiency by their own splicing machinery.

Although we did not see a specific preference for the natural introns of *S. cerevisiae*, we still observe that introns from some species like *E. cymbalariae* or *K. thermotolerans,* are spliced in

lower efficiency compared to introns from other species. We further note that these two species do not stand out phylogenetically from others (Fig. 5A,B). Hence we hypothesized that introns from these species might have been optimized to evolutionary changes in the splicing machinery in their original species. One such molecular candidate could be the gene U2AF1, which is a splicing factor that is associated with the location of the branch site relative to the 3' end of the intron (Neuvéglise et al. 2011). This gene is missing in 6 out of 11 of the yeast species we analyze here including *S. cerevisiae*. In additional species (e.g *T. blatae*) it is highly mutated and probably non-functional (Hooks et al. 2014). Indeed, the introns from the 11 yeast species we used here show a different distribution of BS-to-3'SS distances, that is concordant with the absence or presence of U2AF1 (Fig. S4A), while other properties are not significantly different between the two groups (Fig. S4B-E) (intron length distribution does seem to be different for the two groups, but this difference is solely ascribed to the BS-to-3'SS distance, as can be seen by the lack of difference in 5'SS-to-BS distances (Fig. S4B,C)). When comparing the distribution of splicing efficiencies between introns from species with or without U2AF1, we observed that introns that come from species lacking U2AF1 are better spliced in our *S. cerevisiae* system, which, as noted, lacks U2AF1 (Fig. 5C). Hence we suggest that introns that were adapted to a splicing machinery that uses U2AF1 diverged in evolution and are less suitable to *S. cerevisiae* splicing machinery.

In the previous section, we demonstrated that *S. cerevisiae* has a tendency to splice cryptic introns at alternative 3'SS downstream of the original site, leading to a selection against 3'SS motifs near natural introns 3' end. Interestingly, when performing the same analysis for the other 10 yeast species we found, in all but four species, a similar significant *S. cerevisiae*-like depletion of 3'SS sequence motifs near their introns 3'SS. This suggests an active evolution driven avoidance of the motif in these regions. Strikingly, the 7 species that show this depletion

signal are those lacking U2AF1, and the one species with highly mutated copy of this factor (Fig. 5D, Fig. S5). Put together, our results suggest that loss of the U2AF1 gene results in a flexible recognition of the 3'SS, which in turn generates a selective pressure to avoid 3'SS motifs near the intended 3'SS in order to avoid cryptic splicing events. On the other hand, splicing machinery that includes U2AF1 results in a more stringent 3'SS recognition mechanism, possibly due to tight constraints on the BS-to-3'SS distance.

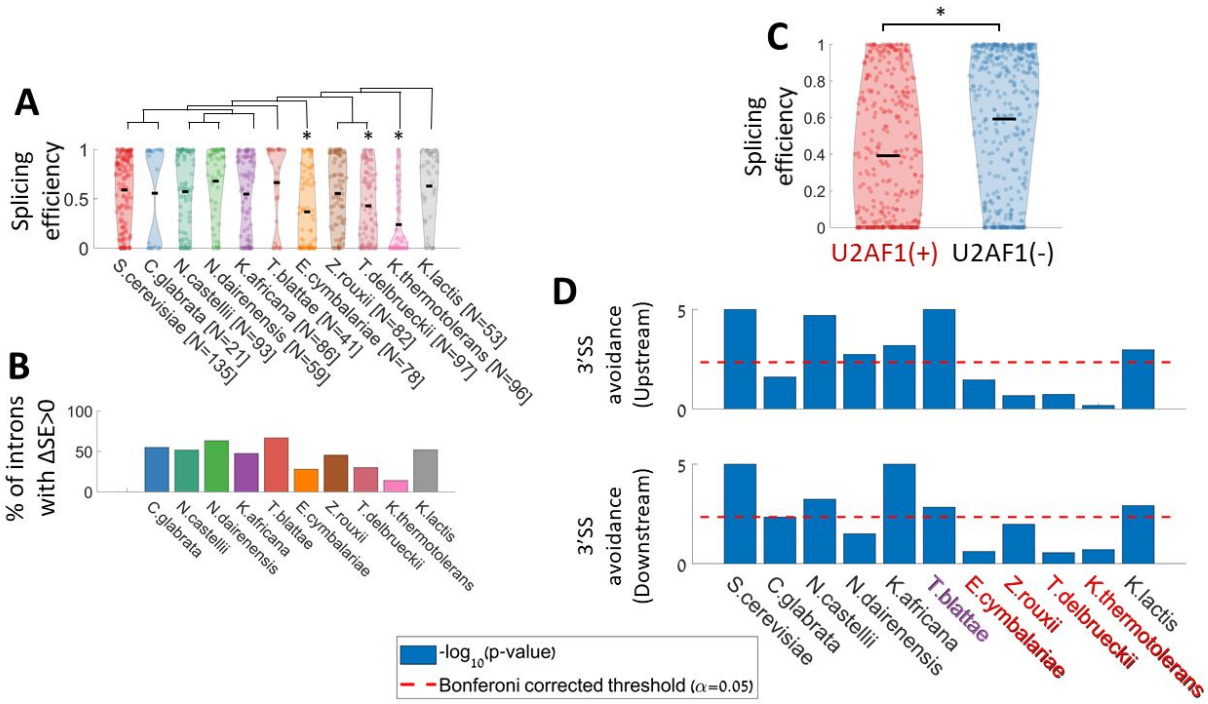


Figure 5 - Analysis of ortholog introns from other yeast species reveals intron architecture evolution

A. Splicing efficiency distribution of spliced variants of introns from orthologs of *S. cerevisiae* intron-containing genes. A species phylogenetic tree (created according to (Feng et al. 2017)) is presented above the corresponding violins. The number of introns included in the library from each species are indicated after each species name. Asterisks mark origin species with splicing efficiency distribution significantly lower than that of *S. cerevisiae*'s introns (t-test, p value<0.001) **B.** Percent of introns that are spliced better than their *S. cerevisiae* ortholog intron for each species. **C.** Splicing efficiency distribution for introns that come from species that have a functional copy of U2AF1 splicing factor (left), and introns that come from species without U2AF1 splicing factor (t-test, p-value<10^{-9}). **D.** Hypothesis test for the 3'SS motif avoidance for each of the 11 species upstream (top) or downstream (bottom) to the 3'SS. P-value was calculated by comparing the mean frequency of the 3'SS motif at a 30nt window upstream/downstream to the 3'SS against 10^5 sets of sequences each composed of coding

genes sequences registered according to randomly chosen positions. Species with a copy of U2AF1 are marked in red, species with malfunctioned U2AF1 are marked in purple, and species without any copy of U2AF1 are marked in black.

A computational model elucidates important features that govern splicing efficiency

In this work we created a large collection of single intron variants, with a systematic exploration of different intron design features. This wide collection of variants allows us to train a computational model that predicts splicing efficiency values from sequence features. For the purpose of this model we used the set of all single-intron variants including both synthetic and natural introns from all species examined in our library, and excluding negative control variants (N=12,667). We trained a gradient boosting model (Friedman 2001; Ke et al. 2017) using a 5-fold averaging cross-validation technique (Jung and Hu 2015) on randomly chosen 75% of the variants set (N=9,500). As an input to the model, we used a set of 39 features, comprising the splice site sequences (as a categorical feature), intron length parameters, GC content, 3' u-rich element, and local secondary structure predictions at each splice site (see a full list of parameters in table S2). The model predictions were tested on the remaining 25% of the set of single-intron variants used for this model (N=3,167). Predicted splicing efficiency values for the test set are reasonably well correlated with the measured splicing efficiency values (Pearson r=0.77, Fig. 6A).

The predictive model enables us to examine the contribution of each feature to a successful prediction of splicing efficiency. We used Shapley values (Lundberg et al. 2020) to infer individual features' importance. Meaning, we analyzed the global contribution of each feature to the predicted splicing efficiency value across all observations. Figure 6B presents the individual feature contribution for each observation (i.e. library variant) of the 7 most important features according to this analysis, the distribution of Shapley values for each feature, and its correspondence with the feature's values. We found that the most important feature is the

sequence of the BS which corresponds with the large difference in splicing efficiency we observed for non-consensus BS variants (Fig 2E). Next, we notice that intronic GC content has high contribution, as low GC content contributes to higher splicing efficiency, which is in agreement with previous findings (Wong et al. 2013; Yofe et al. 2014). The 5'SS sequence also has a high contribution to efficient splicing. Interestingly, while the 3'SS sequence is considered one of the defining features of introns, it is only ranked 7th in terms of importance for the model predictions.

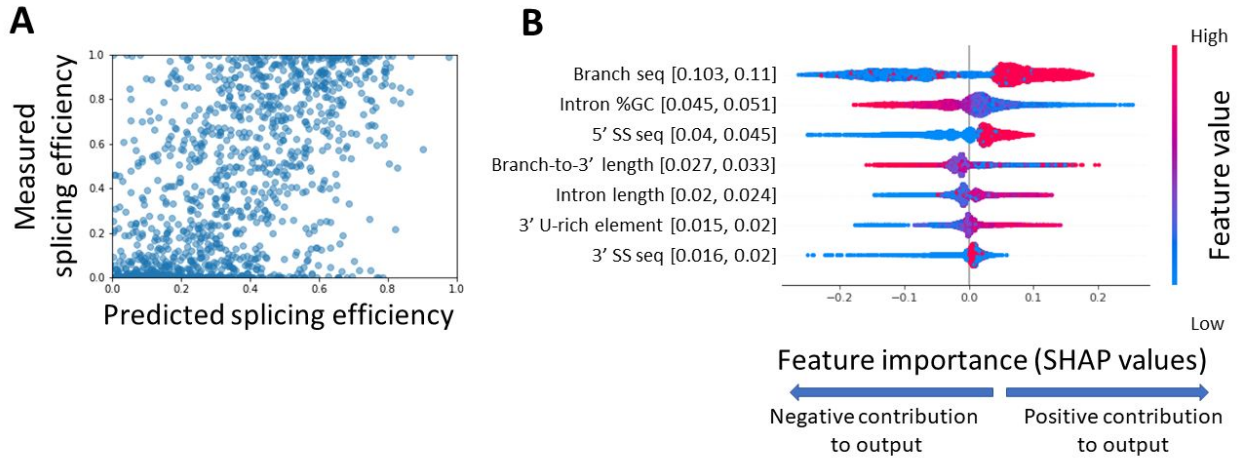


Figure 6 - Sequence features contribution to successful splicing are derived from a computational model

A. Measured splicing efficiency values versus predicted splicing efficiency values of the test set variants (N=3,187) as predicted by a gradient boosting model using 5-fold averaging cross-validation technique (Pearson r=0.77 p-value<10^{-100} 95% CI = [0.745, 0.774]. **B.** Distribution of Shapley values for the top 7 features when ranked according to the mean absolute value of the Shapley values. The x-axis represents the Shapley values, the higher the absolute value of the mean of the distribution, the higher its contribution to the model predictions. Positive values mean that the feature is predicted to improve splicing efficiency, and negative values mean that the feature is predicted to reduce splicing efficiency. Sample points are colored according to their feature's value for numerical features, and for splice site sequences that are treated as categorical features, they are colored by the splice site relative abundance in the *S. cerevisiae* genome (high values represent abundant sequence variants). Numbers in parentheses represent the 95% confidence interval of the mean absolute value of each feature, we notice that except for the bottom two features, the confidence intervals of features do not overlap, indicating stable ranking of the feature contributions. Confidence interval values were calculated by resampling training and test sets 1000 times.

S. cerevisiae has the capacity to alternatively splice two tandem introns, thus generating alternative splice variants from the same RNA

Alternative splicing is not considered to have a major role in gene expression regulation in *S. cerevisiae*. There are 10 known genes with two tandem introns in the *S. cerevisiae* genome (Clark et al. 2002), and most of them are not known to be alternatively spliced. Previous works have examined alternative splicing of a two intron gene in *S. cerevisiae* by studying the spliced isoforms of the two genes that are known to be alternatively spliced (i.e. *DYN2* and *SUS1*) (Howe et al. 2003; Hossain et al. 2011; Li et al. 2019). In these works, the regulation of alternative splicing of a specific gene was studied through chemical or genetic perturbations (Howe et al. 2003) or changing environmental conditions (Hossain et al. 2011). Here we use our library to provide a novel unbiased examination of the potential of *S. cerevisiae* splicing machinery to alternatively splice a synthetic two intron gene. This system allows us to examine intronic features that may facilitate alternative splicing.

We created a subset of the library with two short introns separated by an exon. For this set we chose 25 short introns (<76 nucleotides), 10 of them are the 10 shortest natural introns in *S. cerevisiae*, additional 10 were randomly chosen from all the natural *S. pombe* introns that fit to the length limits of the library and utilize splice sites that are found in *S. cerevisiae*. Lastly, we created 5 synthetic introns with *S. cerevisiae* consensus splice sites, at a length of 56 nucleotides, and BS-to-3'SS distance of 20 nt. Using these 25 short introns, we created a set of variants composed of all possible pairings of two introns, where the first intron was inserted at the 5' end of the variable region, and the second intron at the 3' end of the variable region, separated by an exon, the exon sequence was taken from the *MUD1* based background sequence used for other parts of the library, which resulted in a variable length of the exon depending on the length of the two introns.

Using this set of two-intron variants, we tested whether *S. cerevisiae* has the potential to alternatively splice, and produce multiple spliced isoforms when given a two-intron gene. Such two-intron designs can result in 5 possible isoforms (Fig. 7A). For each variant, we measured the relative frequency of each of the isoforms by aligning its predicted exon-exon junctions to the RNAseq reads. We observed all 4 spliced isoforms in our data (Fig. 6B). Interestingly, out of 614 variants that are spliced, 130 variants have more than one spliced isoform observed for the same pre-mRNA sequence. This observation suggests that the *S. cerevisiae* splicing machinery has the capacity to alternatively splice many possible two-intron sequences, and not only the two natural two-intron genes that are known to be alternatively spliced in this species (Howe et al. 2003; Hossain et al. 2011; Li et al. 2019). This potential for alternative splicing is notable since this species lacks some of the auxiliary factors involved in alternative splicing in other species (Plass et al. 2008; Schwartz et al. 2008).

We compared the relative abundance of the alternative splice variants. There were two distinct hypotheses we could test, either that in each pair of introns one of them will be better spliced than the other regardless to its location in the intron, or that the location of the two introns in the gene will dictate, so that either the upstream or downstream introns will be better spliced. Our results show that the splicing efficiency of each intron is dependent mainly on its sequence, and less on the relative locations of the two introns in the gene. Isoforms with only the upstream or only downstream intron spliced, appear in similar numbers and have similar splicing efficiency distributions (Fig 7B). Furtner, when comparing the proportion of spliced variants for each intron sequence between variants in which it was placed as the upstream intron, and variants in which it was placed as the downstream intron, we see that those two measurements are in high agreement (Fig 6C).

To decipher which intron properties contribute to multiple spliced isoforms, we analyzed all the possible 100 pairs assembled from natural *S. cerevisiae* introns. For each of the 10 introns in this analysis we counted the number of variants for which we observed an isoform in which this intron was spliced out. Then we ranked the 10 introns according to the number variants in which each of them was spliced (regardless of its position as the 1st or 2nd intron). We noticed that multiple isoforms are observed mainly when both introns are ranked high. A single isoform is observed when one intron is ranked low and the other high, and when a pre-mRNA consists of two introns that are ranked low, splicing is hardly observed (Fig. 7D).

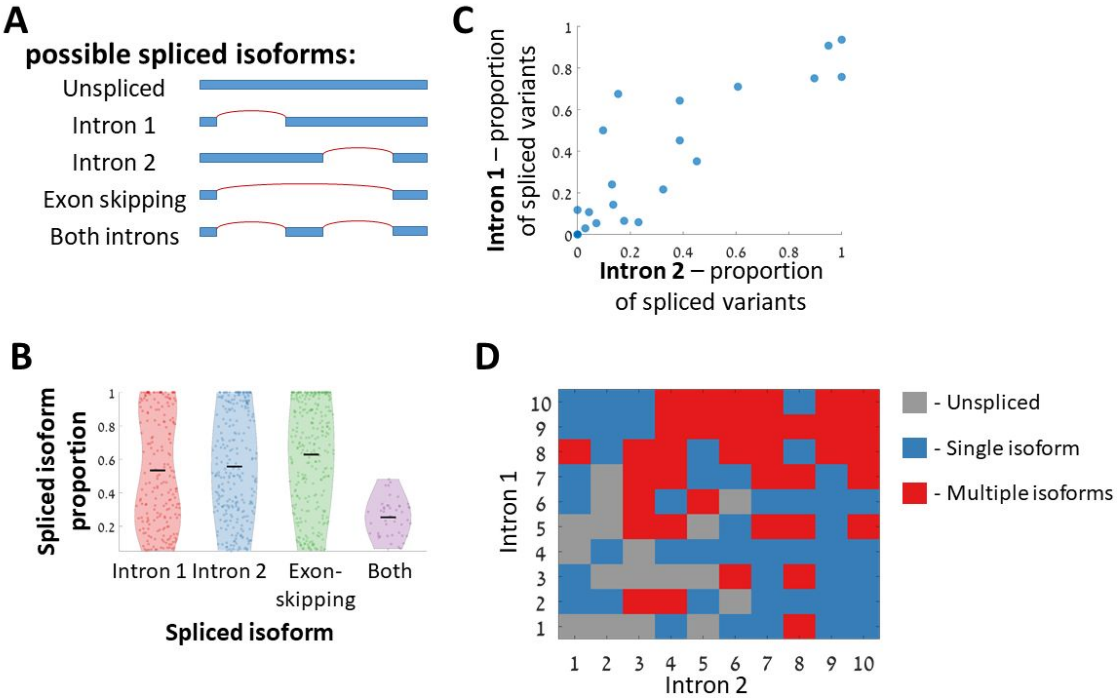


Figure 7 - Tandem two-intron designs demonstrate a capacity of *S. cerevisiae* to alternatively splice two introns.
A. Five possible isoforms can be observed for a two-intron design. **B.** Distribution of isoform's relative frequency for the 4 possible spliced isoforms. **C.** Intron performances when placed as the first intron vs. placed as the second intron. Each dot represents a single intron sequence, the x-axis represents the proportion of variants with this intron spliced as the first intron, and y-axis axis represents the same for the second intron variants (Spearman correlation r=0.88 p-value=10^{-8}). **D.** The number of observed isoforms for each of the natural *S. cerevisiae* intron pair variants. Each number represents a different intron sequence, and they are ordered according to the number of variants in which the intron was spliced.

Discussion

In this study we used a designed, large synthetic oligonucleotide library to study the regulation of constitutive and alternative mRNA splicing. We chose for this study the budding yeast *S. cerevisiae* because a vast majority of introns are constitutively spliced in this species, making it an ideal experimental system to study molecular mechanisms of intron splicing. Yet we could use our system to expose a potential for alternative splicing as well. First, we observed a positive correlation between splicing efficiency and total RNA abundance. This result is in line with previous findings that incorporation of an intron often improves heterologous gene expression in *S. cerevisiae* (Hoshida et al. 2017) and in higher eukaryotes (Mascarenhas et al. 1990; Palmiter et al. 1991). The result is nonetheless surprising to some extent in yeast since most yeast genes do not include an intron, so we would expect the gene expression machinery to not be dependent on splicing. The fact that we observe this correlation for synthetic intron-containing genes suggests there is a molecular mechanism that mediates the correlation, as opposed to an evolutionary indirect effect. Several possible molecular mechanisms could mediate a positive effect of splicing on RNA expression levels. These include, potential effect of splicing on nuclear export (Zhou et al. 2000), on RNA stability (Wang et al. 2007), or on transcription rate (Oesterreich et al. 2016; Ding and Elowitz 2019). Conversely, we note the possibility that splicing efficiency could be increased as a result of another mechanism that in parallel increases gene expression levels. For example in (Ding and Elowitz 2019) it is shown that splicing rates are increased due to spatial clustering of highly expressed transcripts, which in turn increases splicing efficiency through an “economy of scale” principle. However, in our experimental system we presume that is not the case since all library variants are expressed using the same promoter at the same genomic region.

Using combinatorial design approach we compared the effect on splicing efficiency for all naturally occurring splice site variants. Interestingly, in a set of synthetic introns we see no difference between the three possible 3'SS variants. However, when comparing full naturally occuring intron sequences we notice lower splicing efficiencies for AAG 3'SS variants. This result suggests that natural introns contain within them information that disfavors AAG 3'SS utilization. Our results further suggest that such avoidance of the AAG site might be due to the potential of this motif to increase the probability of alternative 3'SS usage. This might explain the fact that AAG 3'SS are rarely observed in fungi genomes, despite the fact that as we demonstrate here, mechanistically it can be spliced as well as the two other variants.

An intronic feature that is unique to *S. cerevisiae* and other fungi is a U-rich element at the 3' end of introns. While a pyrimidine-rich element at introns' end is very common in other eukaryotes, in *S. cerevisiae*, there is only a uracil enrichment (Schwartz et al. 2008). Here, we provide the first experimental evidence that indeed U-rich elements contribute to splicing efficiency, while general Y-rich elements do not (Fig. 3A). We suggest that this might reflect evolutionary changes in the yeast splicing machinery that cause it to interact specifically with uracils. Additionally we provide experimental evidence that *S. cerevisiae* splicing machinery has a preferred BS-to-3'SS length (Fig. 3B), which is supported by the recent high resolution cryo-EM structural analysis of the spliceosome. These optimal lengths might indicate a linkage between the removal of the BS from the active site and the docking of the 3'SS there, although other explanations cannot be ruled out.

We used our library to study the evolution of introns architecture by introducing into it natural introns from 11 yeast species. We tested if introns that come from different hosts that use modified splicing machineries are spliced differently when processed by the *S. cerevisiae* machinery. We suggest that a loss of a splicing factor (U2AF1) that occurred in some of the

yeast species examined here, affects intron architecture through the distance between the branch site and the intron's 3' end. We then observe that introns coming from hosts including this factor are spliced less efficiently in our system lacking the U2AF1 factor (Fig. 5C). We suggest that introns evolved to adapt to the presence or absence of this splicing factor. Additionally, we find that the loss of this splicing factor is accompanied by a genomic depletion of the 3'SS motifs near introns' 3' ends. We hypothesize that in species that lack U2AF1 splicing factor the splicing machinery misidentifies the intended 3'SS due to similar sequence motifs near that site.. This hypothesis is supported by the fact that many cases of alternative 3'SS isoforms are observed both in this current study (Fig. 3A,B) and in previous findings in *S. cerevisiae* (Meyer et al. 2011; Plass et al. 2012). To conclude, the 3'SS in *S. cerevisiae* is recognized through a site with a low information content, and since this organism's splicing machinery lacks the structural rigidity provided by U2AF1, it has the flexibility to recognize several possible splice sites downstream of a BS. On one hand this flexibility creates an opportunity for the organism to adapt through alternative spliced isoforms (Meyer et al. 2011), but on the other hand it creates a risk of producing unwanted spliced isoforms. We believe that this observation presents a general tradeoff seen in molecular biology, when a low information motif results both in flexibility which enables adaptation, and a risk for deleterious effects which drives selection to avoid cryptic motifs. This tradeoff as was also demonstrated recently for a different system, the bacterial promoters (Yona et al. 2018)

Interestingly, in humans, a mutation in U2AF1 was associated with hematopoietic stem cell disorders that can progress into acute myeloid leukemia, and this mutation was shown to cause missplicing events as a result to alterations in preferred 3'SS, which are presumed to be related with the disease (Graubert et al. 2011; Przychodzen et al. 2013; Ilagan et al. 2015).

Lastly, regulated alternative splicing is the focus of many studies on splicing because of its effect on increasing the proteomic diversity of the genome (Rosenberg et al. 2015; Baeza-Centurion et al. 2019; Mikl et al. 2019). In *S. cerevisiae*, there are very few known examples of functional alternative splicing (Howe et al. 2003; Grund et al. 2008; Juneau et al. 2009; Hossain et al. 2011; Meyer et al. 2011). However, using the relative simplicity of the yeast splicing machinery, we aimed to decipher novel ways by which alternative splicing is carried out in biological systems. The hallmark of such effort will be to create a synthetic gene that is alternatively spliced to different isoforms in reaction to different environmental conditions. Here we provide the first step towards this aim by creating a set of multiple two-intron designs and by characterizing which of them give rise to multiple RNA isoforms. We find that the potential for creating multiple isoforms exist in this organism and is mainly dictated by the individual splicing capacity of each of the two introns irrespective of their relative location. If both introns are designed for efficient splicing, their corresponding two-intron design will have the capacity to create multiple isoforms. This result suggests that alternative splicing might be an easily evolvable trait whose evolution might have not necessarily required sophisticated evolution in *trans* as an initial evolutionary step.

Acknowledgments

We wish to thank Sasha F. Levy for kindly providing us the yeast strain and plasmids for the Cre-Lox library integration method, and for helpful discussions on high throughput library experiments. We thank Ruth Sperling for helpful discussions on structural properties of the splicing reaction. We thank Carsten P. Carstens and Ben Borgo from Agilent Technologies for help with the oligo library design and cloning protocols. We thank Leon Anavy and Idan Frumkin

for helpful discussions on large library design, and Idan Frumkin, and Martin Mikl for critical reading of this manuscript.

We thank the Minerva Foundation for grant support. YP is a Kimmel Investigator at the Weizmann Institute.

Data and code availability

All relevant raw files sequencing data are available in the Sequence Read Archive. Accession number PRJNA631112. All the scripts used for data analysis and for producing the figures for this manuscript can be found in https://github.com/DvirSchirman/SplicingLib

Materials & Methods

Synthetic library - general design notes

We used Agilent's oligo library synthesis technology (LeProust et al. 2010) (Agilent Technologies) to produce a pool of 45,000 designed single-stranded DNA oligos at a length of 230 nucleotides. Each oligo includes two 30 nucleotides fixed homology regions at their 5' and 3' end for amplification and cloning, and a 12 nucleotide unique barcode downstream to the 5' homology. This leaves an effective variable region of 158 nucleotides for each variant. The entire synthesized library was composed of several sub-libraries aimed for different projects, these libraries were separated in the initial amplification stage using different homology sequences.

All the experiments and data reported in this paper are based on one sub-library with 21,714 variants (termed SplicingLib1).

Barcodes were chosen such that the minimal edit distance between any two barcodes will be greater than 3 to allow for single error correction for all types of errors including insertion/deletion which are the common error types in oligo synthesis.

The library was designed as a non-coding RNA library in order to avoid possible differences between variants that result from translation. Hence, for each variant, any occurrence of ATG triplet at any frame was mutated to avoid occurrences of a start codon. Except for cases where a 5'SS includes an ATG triplet, in which case, a stop codon was introduced 2 codons downstream of the ATG.

Synthetic library - variants design

The synthetic introns library is composed of several sets of variants. A first set is based on a combinatorial assembly of intron features. Six features were chosen to represent an intron, and all the possible combinations of features were combined to create a set of 5,331 synthetic introns. The features used for this set are: the three splice sites, 5'SS, BS, 3'SS, intron length, BS-to-3'SS length, and a 3' U-enriched sequence element. For each feature, a set of few values was chosen, the 5'SS and 3'SS sets included all the splice sites variants that are found in *S. cerevisiae* genome (5 sequence variants for the 5'SS, and 3 for the 3'SS). The BS included the consensus BS sequence (UACTAAC), and three template sequences with two random nucleotides at the first two positions, since non-consensus BS sequences differ greatly in these positions (NNCUAAC, NNCUAAU, NNUUAAC). For the intron length feature, 5 representing lengths were chosen (73, 89, 105, 121, and 137 nucleotides), and for the BS-to-3'SS length 4 representing lengths were chosen (20, 30, 40, and 50 nucleotides). For the 3' U-enriched sequence element, 3 sequences at different lengths were used (AUUUUUAA, UUUAA, UAA). In addition, For each of the splice sites, a random control sequence was created and a set of control variants was created by assembling the three control sites with all combinations of the other features (see table S1, for summary of the synthetic combinatorial design subset).

Full oligo sequences were based on a background sequence that was derived from the intron-containing region of *MUD1* gene from *S. cerevisiae* genome (positions 4-161 in *MUD1*

open reading frame), followed by randomization of its three splice sites. Each oligo sequence was created by placing a 5'SS 5 nucleotides downstream of the effective variable region instead of the background sequence in this position. Then a BS, U-rich element, and 3'SS sequences were placed in a similar manner according to the chosen length parameters of each variant. In addition, a set of 2,094 variants was created by taking only the consensus splice site sequences at different lengths and incorporating them within 9 additional background sequences, the first two from *UBC9*, and *SNC1* (positions 34-191 and 98-255 in the gens' ORF respectively) genes in a similar manner, and the remaining 7 based on random sequences.

A second set is based on mutating consensus sites' variants from the previous set. 3,607 variants were created by introducing random mutations to splice site sequences. 1,344 additional variants were created by mutating positions adjacent to splice sites with the aim to create a stem-loop RNA structure at the splice sites. This aim was achieved by introducing random mutations and selections *in-silico* of variants for which RNA secondary structure tool (Lorenz et al. 2011) predicts that the splice site will be base-paired within a stem-loop structure.

A third set was based on 1,297 naturally occurring introns from 11 yeast species. We first took all the endogenous intron sequences from *S. cerevisiae* (Clark et al. 2002) that fit into our 158 nucleotides effective variable region (149 introns). Each intron was inserted with a flanking region of 5 nucleotides from each side on the background of the *MUD1* derived background sequence described above. Next, we took intron sequences from orthologs of these intron-containing genes from a set of 10 other yeast species and added them to the library in the same manner. Intron annotations were taken from (Hooks et al. 2014). For the *S. cerevisiae* introns, we also created a set of 3,151 variants with random mutations in introns' splice sites.

Finally, a fourth set of 1,467 variants was created by combining two intronic sequences to create synthetic two-intron variants. For this set, we chose all the introns from *S. cerevisiae*

genome shorter than 76 nucleotides (10 introns), plus 10 randomly chosen short introns from *S. pombe* genome and an additional 5 synthetic introns based on combining consensus splice sites on the background of a random sequence. Each variant sequence was created by placing two introns on the background of the *MUD1* sequence, the first intron at the 5' end of the variable region, and the second at the 3' end of the variable region. All possible pairs of intronic sequences were created and introduced to the library.

Construction of master plasmid

In order to integrate the library into *S. cerevisiae* genome, we used a Cre-Lox based method (Levy et al. 2015). We built a master plasmid to clone the library into, which is compatible with this method. The master plasmid was based on pBAR3 (Levy et al. 2015). A Lox71 site was cloned into pBAR3 to allow Cre-Lox recombination using restriction-free cloning method (Bond and Naus 2012) (primers prDS20, prDS21) to create pDS101. Then we cloned into the plasmid a background sequence that will serve as the library's non-coding gene. A non-coding sequence was designed by taking the sequence of *MUD1* intron-containing gene from its transcription start site to its 3' UTR(-70 to 1106, relative to the start codon), excluding a region around the intron into which the oligo library would be cloned (-45 to 211, relative to start codon). The background sequence was then mutated at any occurrence of ATG to avoid start codons, and additional 27 sites were mutated to reduce homology to the endogenous copy of *MUD1* in the genome. In the cloning site of the oligo library two 20 nucleotides sequences were added, to be used as homology sequences during library cloning. Upstream to the background gene we added a synthetic promoter taken from a published promoter library that was chosen for its high expression level and low noise (Promoter id #2659, from Supp table 3 in (Sharon et al. 2012)). Downstream to the background gene we added *ADH1* terminator sequence. The entire promoter+background gene+terminator construct (total length of 1,397 nucleotides) was

synthesized as a Gene Fragment (Twist Bioscience). The synthesized background gene was cloned into pDS101 using NEBuilder HiFi DNA Assembly (New England Biolabs) to create pDS102 (primers prDS22, prDS23).

Synthetic library - cloning and amplification of plasmid library

Synthetic oligos were first amplified according to Agilent's recommendations (Agilent Technologies 2016). Library oligos were amplified using sub-library specific homology plus 4 different 8 nucleotide sequences that were inserted to serve as an index for control purposes, such that every unique variant could be measured independently 4 times, and a homology sequence to the master plasmid for cloning.

The library was amplified in 4 PCR reactions, Each PCR reaction included:

- 25 ul - KAPA HiFi HotStart ReadyMix (Roche)
- 1.5 ul - 10uM forward primer prDS55-58 for SplicingLib1)
- 1.5 ul - 10uM reverse primer (prDS59 for SplicingLib1)
- 200 pM of DNA oligo library
- H_2O to complete volume to 50ul

PCR program:

1. 95°C 3 min
2. 98°C 20 sec
3. 58°C 15 sec
4. 72°C 15 sec
5. Repeat steps 2-4 for 15 cycles
6. 72°C 1 min

After amplification, the PCR product was cut from agarose gel, purified using Wizard SV Gel and PCR Clean-Up System (Promega), and all 4 reactions were pooled together. The master plasmid pDS102 was linearized using PCR reaction (primers prDS62, prDS63). Then a plasmid library was assembled using 4 independent reactions of NEBuilder HiFi DNA Assembly (New England Biolabs) to avoid biases in assembly that might affect the library's distribution.

From this stage, we followed Agilent's library cloning kit protocol (Agilent Technologies 2017) steps 2-7. In short, the plasmid library was purified using AMPure XP beads (Beckman Coulter),

then inserted to electrocompetent *E.coli* cells (ElectroTen-Blue, Agilent Technologies) using electroporation. Then bacterial cells were inoculated into two 1 liter low gelling agarose LB bottles, in order to grow isolated colonies in 1-liter volume. After 48 hours of growth in 37°C bacterial cells were harvested using centrifugation, and cells were grown overnight on liquid LB media in 37°C. Finally, the amplified plasmid library was extracted from bacterial cells using 4 reactions of Midiprep kit (Macherey-Nagel NucleoBond Xtra Midi Plus).

Growth media

Growth media used in this work:

1. YPG - 10g/L yeast extract, 20 g/L peptone, 20 g/L galactose
2. YPD - 10g/L yeast extract, 20 g/L peptone, 20 g/L glucose
3. SC complete - 6.7 g/L nitrogen base without aminoacids, 20 g/L glucose, 1.5 g/L amino acid mix
4. SC -URA - 6.7 g/L nitrogen base without aminoacids, 20 g/L glucose, 1.5 g/L drop-out mix lacking Uracil

Synthetic library - integration into yeast genome

The library was inserted to *S. cerevisiae* strain yDS101 (ura3Δ ybr209w::GalCre-KanMX-1/2URA3 -lox66 HOΔ::TEF2-mCherry::pCUP1-YiFP::NAT). This strain was based on SHA185 strain, that contains Cre-Lox landing pad and is derived from BY4709 strain, kindly supplied to us by Sasha F. Levy's lab (Levy et al. 2015). Transformation of the plasmid library to yeast cells was done using a Cre-Lox based high throughput genomic integration method (Levy et al. 2015), that inserts the plasmid sequence into the YBR209W dubious open reading frame. yDS101 yeast cells were transformed with 500ug plasmid library and grown overnight in YPG media to induce Cre expression. Then cells were plated on selective media (SC-Ura) approximately 50 plates per transformation. We counted the number

of colony-forming units by plating diluted samples and got $1.5 \cdot 10^6$ CFUs for SPlicingLib1 which are ~60 times the number of unique variants in the library accordingly.

RNA extraction, cDNA synthesis, and genomic DNA extraction

Total RNA of the library cells was extracted in two independent repeats. Library cells were grown overnight in SC complete media, and then diluted to a fresh media by 1:100 factor and grown for an additional 6 hours until they reached OD_{600} of 0.5, such that cells are harvested in mid-log phase. The cell culture was centrifuged for 45 seconds at 4,000g and the pellet was immediately frozen in liquid nitrogen.

RNA was extracted using MasterPure Yeast RNA Purification Kit (Lucigen), and treated with TURBO DNase (ThermoFischer) to remove any residues of genomic DNA. We then synthesized cDNA using reverse transcription with random primers using qScript Flex cDNA Synthesis Kit (QuantaBio).

In order to normalize RNA levels by the relative frequency of each variant in the sample, we extracted genomic DNA from the same samples used for RNA extraction. Cells were harvested and frozen at mid-log the same as for RNA extraction. DNA was extracted from all samples using MasterPure Yeast DNA Purification Kit (Lucigen).

Next-generation sequencing - library preparation

Both cDNA and genomic DNA samples were prepared for sequencing in the same manner. We used a two-step PCR protocol to amplify the library's variable region and link it to Illumina's adaptors with indexes.

The first PCR reaction was used to amplify the variable region and link homology sequences to Illumina's adaptors, we performed 8 parallel reactions to each sample to reduce PCR biases. We used 6 different forward primers each with one extra nucleotide, to create shifts of the amplicon sequence in order to avoid low complexity library.

Each reaction included:

- 25 ul - KAPA HiFi HotStart ReadyMix (Roche)
- 1.5 ul - 10uM forward primer (prDS137-142)
- 1.5 ul - 10uM reverse primer (prDS143)
- 100ng DNA
- H_2O to complete volume to 50ul

PCR program:
1. 95ºC 3 min
2. 98ºC 20 sec
3. 58ºC 15 sec
4. 72ºC 15 sec
5. Repeat steps 2-4 for 20 cycles
6. 72ºC 1 min

Next, we pooled all 8 reactions for each sample and purified the PCR product using AMPure XP beads (Beckman Coulter). The second PCR was used to link specific indexes to each sample so we can multiplex several samples in a single sequencing run.

Each reaction included:
- 25 ul - Phusion High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs)
- 2.5 ul - 10uM forward primer (prDS144)
- 2.5 ul - 10uM reverse primer (prDS145)
- 1-5ng DNA
- H_2O to complete volume to 50ul

PCR program:
1. 98ºC 30 sec
2. 98ºC 10 sec
3. 62ºC 20 sec
4. 72ºC 15 sec
5. Repeat steps 2-4 for 15 cycles
6. 72ºC 5 min

Next, we purified the PCR product using AMPure XP beads (Beckman Coulter), quantified final concentration using Qubit dsDNA HS (ThermoFischer), diluted all samples to 4nM, and pooled together all the samples. NGS library was sequenced in Illumina NextSeq 500 system, using 150x2 paired-end sequencing. We obtained a total of 13.9, 12 million reads for the two RNA samples of SplicingLib1, and 1.9, 1.9 million reads for the two corresponding DNA samples.

Mapping sequencing reads to the library's variants

Sequencing reads from all samples were processed the following way: We first merged paired-end reads using PEAR (Zhang et al. 2014), next we trimmed homology sequences and demultiplexed the reads according to the 4 control indexes using Cutadapt (Martin 2011), then we clustered all unique reads using 'vsearch --derep_prefix' (Rognes et al. 2016).

All the unique reads were mapped to a library variant according to the first 12 nucleotides in the read, which are the designed barcode. A read was mapped to one of the library's variants by searching the barcode with minimal edit distance to the read's barcode. If this minimal distance was <3, and only a single library barcode is found in this distance, the read was aligned to this variant.

Data analysis

All data analysis except the gradient boosting model were done in Matlab (R2018b). Gradient boosting modeling was done in Python 3.7.

Computing splicing efficiencies

For each variant, the mapped reads obtained from the RNA sequencing were first classified into three possible types: unspliced, intended spliced isoform, and undetermined. A read was classified into one of these types using an alignment of 40 nucleotide sequences representing *exon-intron*, and *exon-exon* junction sequences. We aligned each read to the reference junction sequences using local Smith-Waterman alignment (swalign function in Matlab), and a normalized alignment score was defined the following:

$$Junction\ alignment\ score\ = \frac{SW(junction,read)}{SW(junction,junction)}$$

If the normalized score was >0.8 we infer the junction is positively aligned to the RNA read.

A read was classified as unspliced if it was aligned to the two *exon-intron* junctions and not aligned to the *exon-exon* junction. A read was classified as 'intended spliced isoform' if it was

aligned to the *exon-exon* junction and not to the two reference *exon-intron* junctions. All other reads were classified as undetermined.

Intended splicing efficiency for each variant was then calculated for each index according to:

$$SE = \frac{spliced\ isoform\ abundance}{total\ RNA\ abundance}$$

A final splicing efficiency value for each variant was then set by taking the median between indexes in each repeat and then taking the mean between the two repeats.

The undetermined reads were further analyzed to search for cryptic spliced isoforms, meaning, isoforms that result from splicing of an intron different than the designed intron, hence no *exon-exon* reference junction could be defined. Each read was aligned against the full reference design with the following parameters to the Smith-Waterman algorithm (Gapopen = 100, ExtendGap=1) in order to allow for alignment with long uninterrupted gaps, if the normalized alignment score was <0.7 and the number of mismatches in the alignment was <6, a read was set as cryptic spliced isoform. Then the 5' and 3' end of the intron were set according to the ends of the uninterrupted gap.

For each variant, cryptic splice isoforms were clustered according to their 3' and 5' intron ends, and for each cluster, we calculated the splicing efficiency as described above for the intended spliced isoforms. cryptic spliced isoforms were counted only for isoforms with splicing efficiency higher than 0.01.

<u>Computing two-intron spliced isoforms ratio</u>

For the set of two-intron variants, we needed to classify each read to one of five possible isoforms: unspliced, intron 1, intron 2, exon-skipping, or 'both introns spliced'. Reads were classified into one of these isoforms according to junctions alignment as described above. A read was classified to an isoform according to the following conditions:

- Intron 1 - positive alignment to the *exon1-exon2* junction, and negative to the *exon1-intron1* and *intron1-exon2* junctions.

- Intron 2 - positive alignment to the *exon2-exon3* junction, and negative to the *exon2-intron2* and *intron2-exon3* junctions.
- Exon-skipping - positive alignment to the *exon1-exon3* junction, and negative to the *exon1-intron1* and *intron2-exon3* junctions.
- Both introns - positive alignment to the concatenated *exon1-exon2-exon3* junction, and negative to all the 4 *exon-intron* junctions.
- Unspliced - negative alignment to both *exon1-exon2* and *exon2-exon3* junctions, and positive alignment to all 4 *exon-intron* junctions.

Then the splicing ratio of each isoform was determined by the ratio of its spliced isoform abundance and the total RNA abundance.

<u>Total RNA abundance, and data filtering</u>

Genomic DNA levels were used to determine total RNA abundance, and to filter outlier spliced isoforms.

To determine total RNA abundance we wish to normalize by the variant's frequency in the population. Hence, Total RNA abundance of variant x was determined according to:

$$Total\ RNA\ abundance(x) = log_{10}\left(\frac{RNA\ frequency(x)}{DNA\ frequency(x)}\right)$$

Some RNA read alignments might be inferred as spliced isoforms due to errors in synthesis, or systematic errors in alignment. Therefore, the splicing efficiency calculation was done also on the DNA samples, and if a variant had an intended or cryptic splicing efficiency higher than 0.05 in the DNA samples the corresponding value was set to zero.

<u>3'SS avoidance calculation</u>

For each of the 11 yeast species, we examine the frequency of the 3'SS sequence motif, to check if it is avoided near introns' 3' end. First, we calculate the frequency of the two major 3'SS sequences ([C/T]AG) at positions relative to the introns' 3' end. For that purpose, in each species, we register the sequences of all the intron-containing genes at their intron's 3' end and set the end of the intron as position 0. Then, at every position downstream or upstream to the

intron end, we count the number of occurrences of the two 3'SS sequences and divide it by the number of introns in each species.

Then we test if there is a statistically significant depletion of this motif at a window of 30 nucleotides upstream or downstream of the intron end. We perform the statistical analysis using sampling of random control sets. Each control set includes N random positions from coding regions in the same genome, where N is the number of introns in a species. Those positions are set as the reference positions at which we register N sequences, and measure the 3'SS motif frequency around them. We randomly sample 10^5 such control sets, and then we count the number of sets for which the mean frequency within a window of 30 nucleotides is lower than the mean frequency in the true introns set. $p-value$ is defined according to: ($f_{motif}^{introns}$ and $f_{motif}^{control}$ are the frequency of the 3'SS motif at the true introns set, or the control set accordingly)

$$p-value(upstream) = \#\left[\sum_{i=-32}^{-3} f_{motif}^{control} < \sum_{i=-32}^{-3} f_{motif}^{introns}\right] \cdot 10^{-5}$$

$$p-value(downstream) = \#\left[\sum_{i=1}^{30} f_{motif}^{control} < \sum_{i=1}^{30} f_{motif}^{introns}\right] \cdot 10^{-5}$$

<u>A computational model for predicting single-intron splicing efficiency</u>

We used a gradient boosting regression model to predict splicing efficiencies of library variants. The gradient boosting implementation is based on LightGbm (Ke et al. 2017) library for Python, and the feature importance inference is based on SHAP (Lundberg et al. 2020) library for python.

Each variant is characterized by a set of 39 features (see table S2). We took a set of 12,745 variants that includes all the designed single intron variants, excluding negative controls. This set was randomly divided into a training set composed of 75% of the variants and a test set with the remaining 25%. We then trained the model on the training set using 5-fold averaging

cross-validation technique (Jung and Hu 2015), meaning, we divided the training set to 5 subsets, each time training the model on 4 of them, using the fifth as a validation set, and predicting the splicing efficiency value for the test set. Thus, creating 5 different predictions for the test set, which we next averaged to create a single prediction.

The parameters given to the model are the following:

- Number of leaves - 50
- Learning rate - 0.1
- Feature fraction - 0.8
- Bagging fraction - 0.8
- Bagging frequency - 5
- Number of boost rounds - 500
- Number of early stopping rounds - 5

Feature importance was inferred by running Shapley value analysis (Lundberg et al. 2020) on the training set for each of the 5 k-fold iterations, followed by averaging the Shapley values over the 5 iterations. Confidence interval of the mean absolute values of Shapley values was calculated by resampling the training and test set in 1000 repeats and taking the 2.5% and 97.5% quantiles of mean absolute value for each feature.

References

Agilent Technologies. 2016. *G7555-90000 SureGuide Custom CRISPR Guide Library - Guidelines for Amplification and Cloning Assembly*. Agilent Technologies https://www.agilent.com/cs/library/usermanuals/public/G7555-90000.pdf (Accessed March 19, 2020).

Agilent Technologies. 2017. *G7556-90000 SureVector CRISPR Library Cloning Kit - Protocol*. Agilent Technologies https://www.agilent.com/cs/library/usermanuals/public/G7556-90000.pdf (Accessed March 20, 2020).

Andreadis A, Gallego ME, Nadal-Ginard B. 1987. Generation of protein isoform diversity by alternative splicing: mechanistic and biological implications. *Annu Rev Cell Biol* **3**: 207–242.

Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. 2019. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176**: 549-563.e23.

Bond SR, Naus CC. 2012. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. *Nucleic Acids Res* **40**: W209-13.

Chen F-C, Wang S-S, Chen C-J, Li W-H, Chuang T-J. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol* **23**: 675–682.

Cheung R, Insigne KD, Yao D, Burghard CP, Wang J, Hsiao Y-HE, Jones EM, Goodman DB, Xiao X, Kosuri S. 2019. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. *Mol Cell* **73**: 183-194.e8.

Clark TA, Sugnet CW, Ares M. 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**: 907–910.

Coolidge CJ, Seely RJ, Patton JG. 1997. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res* **25**: 888–896.

Dewey CN, Rogozin IB, Koonin EV. 2006. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* **7**: 311.

Ding F, Elowitz MB. 2019. Constitutive splicing and economies of scale in gene expression. *Nat Struct Mol Biol* **26**: 424–432.

Douglass S, Leung CS, Johnson TL. 2019. Extensive splicing across the Saccharomyces cerevisiae genome. *BioRxiv*.

Feng B, Lin Y, Zhou L, Guo Y, Friedman R, Xia R, Hu F, Liu C, Tang J. 2017. Reconstructing Yeasts Phylogenies and Ancestors from Whole Genome Data. *Sci Rep* **7**: 15209.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Ann Statist* **29**: 1189–1232.

Frumkin I, Yofe I, Bar-Ziv R, Gurvich Y, Lu Y-Y, Voichek Y, Towers R, Schirman D, Krebber H, Pilpel Y. 2019. Evolution of intron splicing towards optimized gene expression is based on various Cis- and Trans-molecular mechanisms. *PLoS Biol* **17**: e3000423.

Gotic I, Omidi S, Fleury-Olela F, Molina N, Naef F, Schibler U. 2016. Temperature regulates splicing efficiency of the cold-inducible RNA-binding protein gene Cirbp. *Genes Dev* **30**: 2005–2017.

Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, et al. 2011. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet* **44**: 53–57.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**: 100–107.

Grund SE, Fischer T, Cabal GG, Antúnez O, Pérez-Ortín JE, Hurt E. 2008. The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. *J Cell Biol* **182**: 897–910.

Harigaya Y, Parker R. 2012. Global analysis of mRNA decay intermediates in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **109**: 11764–11769.

Hooks KB, Delneri D, Griffiths-Jones S. 2014. Intron evolution in Saccharomycetaceae. *Genome Biol Evol* **6**: 2543–2556.

Hoshida H, Kondo M, Kobayashi T, Yarimizu T, Akada R. 2017. 5´-UTR introns enhance protein expression in the yeast Saccharomyces cerevisiae. *Appl Microbiol Biotechnol* **101**: 241–251.

Hossain MA, Rodriguez CM, Johnson TL. 2011. Key features of the two-intron Saccharomyces cerevisiae gene SUS1 contribute to its alternative splicing. *Nucleic Acids Res* **39**: 8612–8627.

Howe KJ, Kane CM, Ares M. 2003. Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae. *RNA* **9**: 993–1006.

Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, Bradley RK. 2015. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res* **25**: 14–26.

Juneau K, Nislow C, Davis RW. 2009. Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics* **183**: 185–194.

Jung Y, Hu J. 2015. A K-fold Averaging Cross-validation Procedure. *J Nonparametr Stat* **27**: 167–179.

Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet* **12**: 715–729.

Ke G, Meng Q, Finley T, Wang T, Chen W. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 3146.

Kosuri S, Goodman DB, Cambray G, Mutalik VK, Gao Y, Arkin AP, Endy D, Church GM. 2013. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci USA* **110**: 14024–14029.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

LeProust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH. 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* **38**: 2522–2540.

Levy L, Anavy L, Solomon O, Cohen R, Brunwasser-Meirom M, Ohayon S, Atar O, Goldberg S, Yakhini Z, Amit R. 2017. A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial σ54 Promoters. *Cell Rep* **21**: 845–858.

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**: 181–186.

Li X, Liu S, Zhang L, Issaian A, Hill RC, Espinosa S, Shi S, Cui Y, Kappel K, Das R, et al. 2019. A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**: 375–380.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* **2**: 56–67.

Luukkonen BG, Séraphin B. 1997. The role of branchpoint-3' splice site spacing and interaction between intron terminal nucleotides in 3' splice site selection in Saccharomyces cerevisiae. *EMBO J* **16**: 779–792.

Madhani HD, Guthrie C. 1994. Dynamic RNA-RNA interactions in the spliceosome. *Annu Rev Genet* **28**: 1–26.

Maricque BB, Chaudhari HG, Cohen BA. 2018. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol*.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* **17**: 10.

Mascarenhas D, Mettler IJ, Pierce DA, Lowe HW. 1990. Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol Biol* **15**: 913–920.

Meyer M, Plass M, Pérez-Valle J, Eyras E, Vilardell J. 2011. Deciphering 3'ss selection in the yeast genome reveals an RNA thermosensor that mediates alternative splicing. *Mol Cell* **43**: 1033–1039.

Mikl M, Hamburg A, Pilpel Y, Segal E. 2019. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat Commun* **10**: 4572.

Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020. Codon Usage and Splicing Jointly Influence mRNA Localization. *Cell Syst*.

Neuvéglise C, Marck C, Gaillardin C. 2011. The intronome of budding yeasts. *C R Biol* **334**: 662–670.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457–463.

Oesterreich FC, Herzel L, Straube K, Hujer K, Howard J, Neugebauer KM. 2016. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**: 372–381.

Palmiter RD, Sandgren EP, Avarbock MR, Allen DD, Brinster RL. 1991. Heterologous introns can enhance expression of transgenes in mice. *Proc Natl Acad Sci USA* **88**: 478–482.

Parenteau J, Durand M, Morin G, Gagnon J, Lucier J-F, Wellinger RJ, Chabot B, Elela SA. 2011. Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. *Cell* **147**: 320–331.

Patterson B, Guthrie C. 1991. A U-rich tract enhances usage of an alternative 3' splice site in yeast. *Cell* **64**: 181–187.

Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet* **24**: 590–594.

Plass M, Codony-Servat C, Ferreira PG, Vilardell J, Eyras E. 2012. RNA secondary structure mediates alternative 3'ss selection in Saccharomyces cerevisiae. *RNA* **18**: 1103–1115.

Przychodzen B, Jerez A, Guinta K, Sekeres MA, Padgett R, Maciejewski JP, Makishima H. 2013. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood* **122**: 999–1006.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**: 698–711.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res* **18**: 88–103.

Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, Segal E. 2015. Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet* **11**: e1005147.

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**: 521–530.

Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res* **24**: 1698–1706.

Smith CW, Chu TT, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol* **13**: 4939–4952.

Stajich JE, Dietrich FS, Roy SW. 2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biol* **8**: R223.

Wang H-F, Feng L, Niu D-K. 2007. Relationship between mRNA stability and intron presence. *Biochem Biophys Res Commun* **354**: 203–208.

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E. 2016. Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**.

Wieringa B, Hofer E, Weissmann C. 1984. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell* **37**: 915–925.

Wilkinson ME, Charenton C, Nagai K. 2020. RNA splicing by the spliceosome. *Annu Rev*

Biochem **89**: 359–388.

Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583–595.

Wu S, Romfo CM, Nilsen TW, Green MR. 1999. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* **402**: 832–835.

Xia X. 2019. RNA-Seq approach for accurate characterization of splicing efficiency of yeast introns. *Methods*.

Yofe I, Zafrir Z, Blau R, Schuldiner M, Tuller T, Shapiro E, Ben-Yehezkel T. 2014. Accurate, model-based tuning of synthetic gene expression using introns in S. cerevisiae. *PLoS Genet* **10**: e1004407.

Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo promoters. *Nat Commun* **9**: 1530.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**: 614–620.

Zhou Z, Luo MJ, Straesser K, Katahira J, Hurt E, Reed R. 2000. The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407**: 401–405.

Supplementary figures

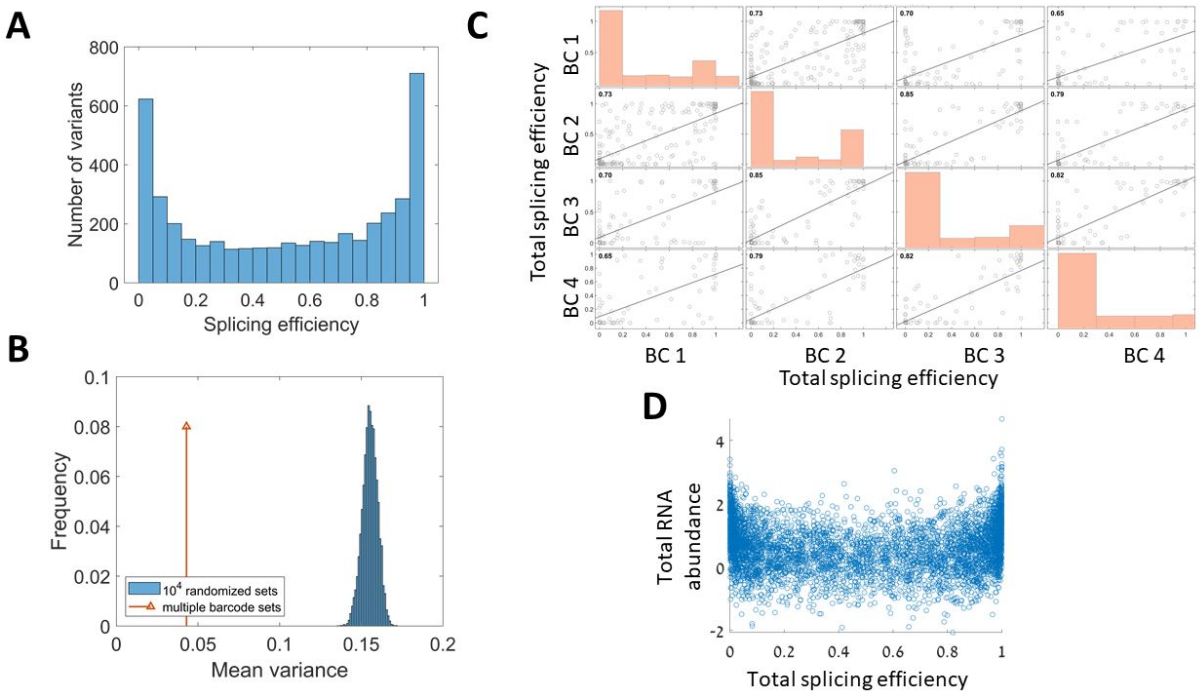


Figure S1 - Multiple barcodes, and randomized control for RNA abundance and splicing efficiency

A. Histogram of total splicing efficiency values. Only variants with splicing efficiency > 0 are presented. **B.** Splicing efficiency mean variance of quartets of the same sequence design, with different barcode sequences (Red arrow), compared to the distribution of mean variance of 10,000 sets of quartets, randomly chosen from the set of designs with multiple barcodes. **C.** Pairwise Pearson correlations of total splicing efficiency values (i.e. splicing efficiency of intended + cryptic spliced isoforms) between designs with identical sequence and different 12nt barcode sequence (mean Person correlation, r=0.76). **D.** To check if the correlation between RNA abundance and splicing efficiency results trivially from the dependence of splicing efficiency value on the total RNA abundance we ran the same analysis as in Figure 2A on a randomized dataset. For 5,000 mock variants we randomly assign unspliced RNA levels, and spliced RNA levels. Both values are randomly chosen from a log-normal distribution. We then calculate the splicing efficiency of each mock variant, and plot the scatter of total RNA abundance and splicing efficiency. No significant correlation is observed in this mock data (p-value = 0.45).

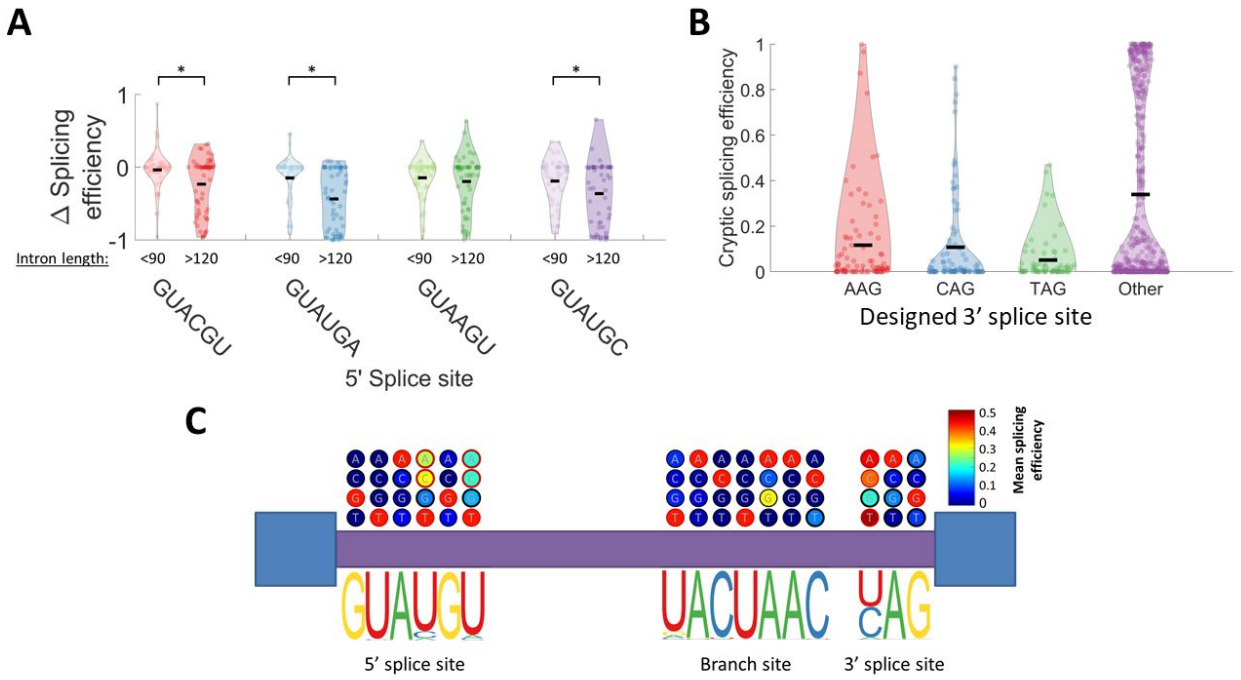


Figure S2 - 5'SS variants effect is dependent on intron length, and single nucleotide mutation analysis

A. For two of the 5'SS variants (i.e. GUACGU, GUAUGA), the difference in splicing efficiency as described in Fig. 2D, is significantly negative only for introns longer than 120 nucleotides. This figure presents the distribution of the difference in splicing efficiency for 5'SS variants after binning variants according to their intron length. **B**. Distribution of cryptic isoforms splicing efficiency for the set of synthetic introns, binned according to the original intron's 3'SS. **C.** Single mutation analysis of splice sites' positions. For each site, the bottom part presents the sequence logo of the consensus splice site, and above it a depiction of the effect of each mutation on the proportion of spliced variants. For each mutation, the other positions within splice sites are kept at the consensus sequence. Color bar represents the mean splicing efficiency of all variants with this mutation as a single mutation in the splice sites. Values significantly higher than zero are marked with a red circle (proportion z-test, with Bonferroni correction).

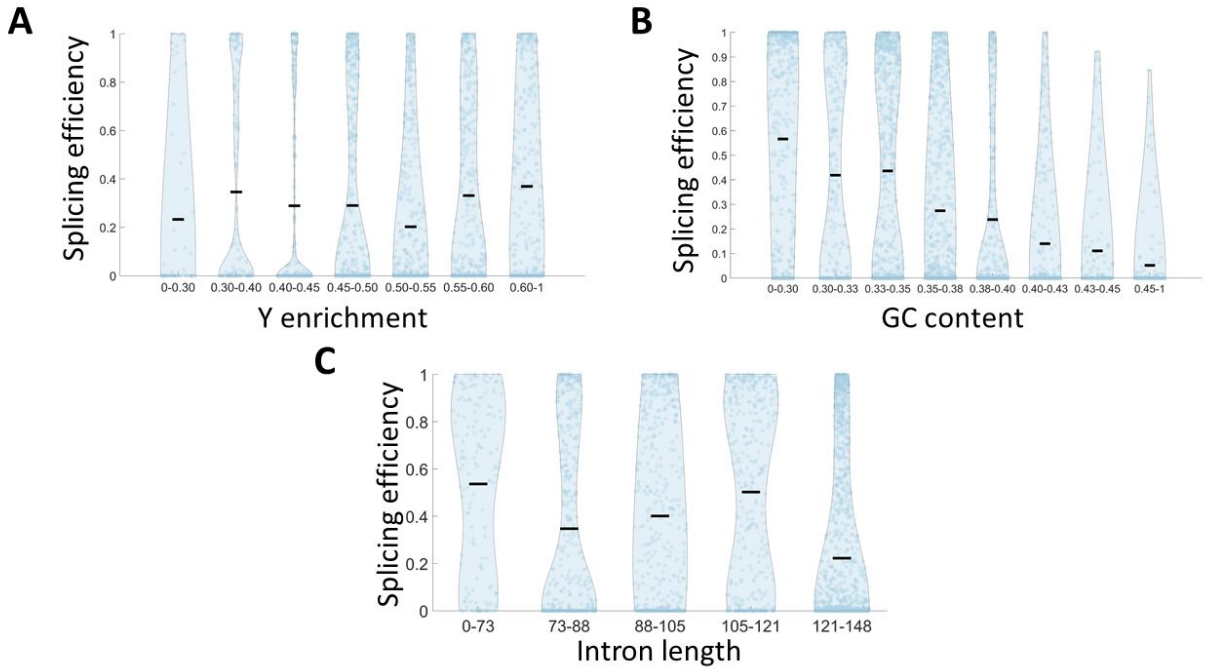


Figure S3 - Splicing efficiency is not correlated with Y-enrichment at intron's 3' end
A-C. All the panels of this figure present a comparison of splicing efficiency values for all the library variants utilizing consensus splice sites' sequences. **A.** Splicing efficiency distribution is binned according to poly-pyrimidine tract strength , which is calculated as the Y (i.e. C or U) content at a window of 20 nucleotides upstream to the 3'SS. In order to specifically check elements that are not U-rich, only elements with at least 30% C out of their Y content are taken into account. Correlation is not significant (p-value=0.79), compared to the highly significant correlation for U-rich elements (Fig 3A). **B.** Splicing efficiency distribution binned according to intronic GC content (Pearson r=-0.85 p-value<0.01). **C.** Splicing efficiency distribution of spliced variants for different intron lengths.

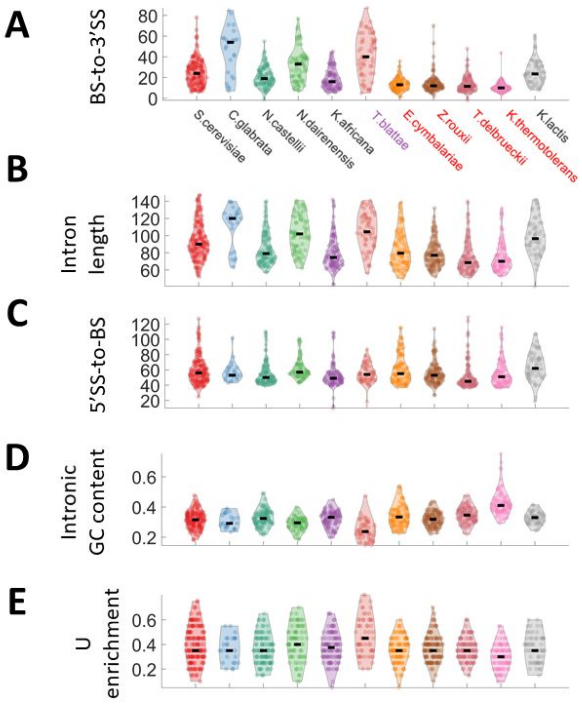


Figure S4 - BS-to-3'SS distance distribution is associated with existence U2AF1 splicing factor
A. Distribution of BS-to-3'SS distance in each of the 11 species. Species with no copy of the gene coding for the splicing factor U2AF1 are marked in black, one species with a malfunctioned copy of U2AF1 is marked in purple, and the ones with a functional copy of U2AF1 are marked in red. **B-E.** For the same species represented in (A), and in corresponding positions, distribution of intron length (B), 5'SS-to-BS distance (C), intronic GC content (D), and poly uracil enrichment as calculated in Fig. 3A. We notice that although intron length differs substantially between species with U2AF1 splicing factor, to species that lack it (B), this difference is ascribed solely to differences in BS-to-3'SS distance (A), as we see no difference in 5'SS-to-BS distance

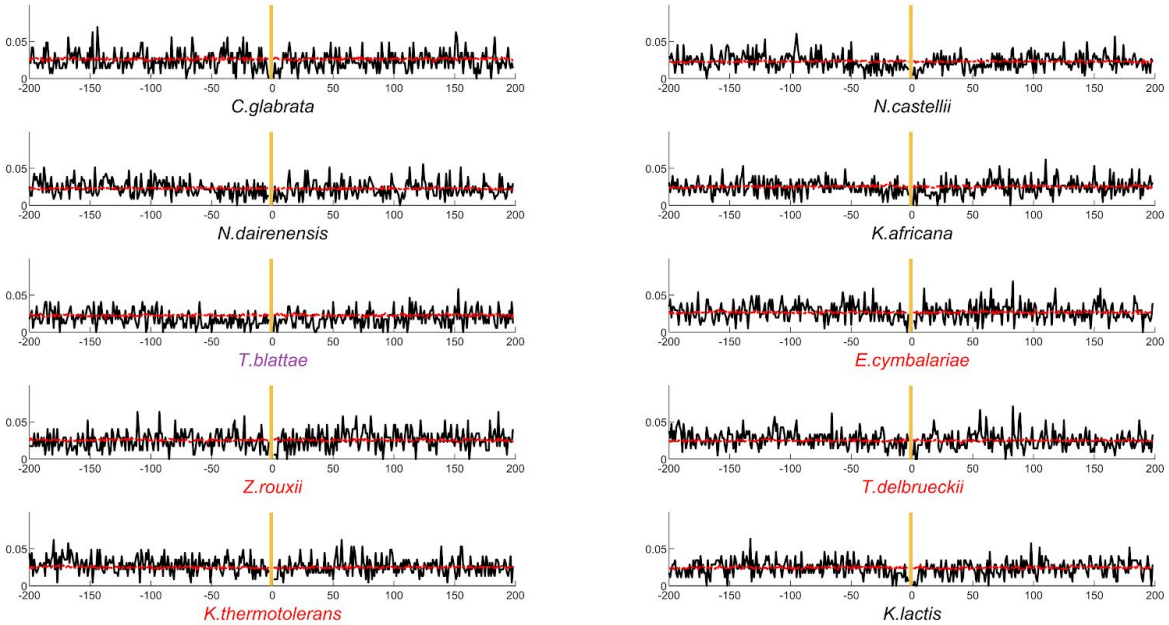


Figure S5 - 3'SS motif avoidance signal for 10 yeast species
3' splice site motif avoidance signal for each of the other 10 yeast species that contributed introns to the library. Each panel presents the motif avoidance signal as explained in Fig 4F.

Supplementary tables

Table S1 - Synthetic combinatorial design library subset features

Feature name	Possible values
5'SS sequence	'GUAUGU'; 'GUACGU'; 'GUAUGA' ;'GUAAGU'; 'GUAUGC'
Negative control 5'SS sequence	'AUUGUG'; 'CGAUGG'
BS sequences*	'UACUAAC'; 'NNCUAAC'; 'NNCUAAU'; 'NNUUAAC'
Negative control BS sequences	'CAUAUCA'; 'AUCGAGC'
3'SS sequences	'UAG'; 'CAG'; 'AAG'
Negative control 3'SS sequences	'AGU'; 'CAU'
Intron lengths [Nucleotides]	73, 89, 105, 121, 137
BS-to-3'SS lengths [Nucleotides]	20, 30, 40, 50
3' U-rich sequence element	'AUUUUUAA'; 'UUUAA'; 'UAA'

* - In BS sequence variants since in the genome there is variation in these positions for non-consensus variants, 'N' nucleotides were replaced with a randomly chosen nucleotide. For the 'NNCUAAC' variant, only variants different than 'UACUAAC' are considered (as this is the consensus sequence).

Table S2 - gradient boosting model features

Feature name	Feature type
5'SS sequence	categorical
BS sequence	categorical
3'SS sequence	categorical
Intron GC%	numeric
U-enrichment @ 3' end	numeric
Intron length	numeric
BS-to-3'SS length	numeric
5'SS ΔG (30nt window)	numeric
BS ΔG (30nt window)	numeric
3'SS ΔG (30nt window)	numeric
3'SS GC% (30nt window)	numeric
5'SS stem length	numeric

BS stem length	numeric
3'SS stem length	numeric
5'SS stem arm	numeric
BS stem arm	numeric
3'SS stem arm	numeric
5'SS fraction of nucleotides based paired	numeric
BS fraction of nucleotides based paired	numeric
3'SS fraction of nucleotides based paired	numeric
5'SS - is 1st nucleotide paired	categorical
5'SS - is 2nd nucleotide paired	categorical
5'SS - is 3rd nucleotide paired	categorical
5'SS - is 4th nucleotide paired	categorical
5'SS - is 5th nucleotide paired	categorical
5'SS - is 6th nucleotide paired	categorical
BS - is 1st nucleotide paired	categorical
BS - is 2nd nucleotide paired	categorical
BS - is 3rd nucleotide paired	categorical
BS - is 4th nucleotide paired	categorical
BS - is 5th nucleotide paired	categorical
BS - is 6th nucleotide paired	categorical
BS - is 7th nucleotide paired	categorical
3'SS - is 1st nucleotide paired	categorical
3'SS - is 2nd nucleotide paired	categorical
3'SS - is 3rd nucleotide paired	categorical

Table S3 - List of primers

Name	Used for	Sequence

prDS1	NEBuilder assembly of reporter cassette	CTCATAAGCAGCAATCAATTCTATCTATACTTTAAAATGCTTTCTGCATCTATATTACCCTGTTATCCC
prDS6	NEBuilder assembly of reporter cassette	GATCGGCTTACTAATATGGGGCCGTATACTTAC
prDS7	NEBuilder assembly of reporter cassette	ACGGCCCCATATTAGTAAGCCGATCCCATTAC
prDS8	NEBuilder assembly of reporter cassette	TCACCTTTAGACATTTTATGTGATGATTGATTGATTG
prDS9	NEBuilder assembly of reporter cassette	AATCATCACATAAAATGTCTAAAGGTGAAGAATTATTCACTGGTGT
prDS10	NEBuilder assembly of reporter cassette	CTGGTTGAAACAAATCAGTGCCGGTAACGCTTTTTGTATCTTGAGTCGACACTGGATGGCGGC
prDS20	RF cloning pBAR3	CCTTCGTTCTTCCTTCTGTTCGGAGGGGACCAGGTGCCGTAAG
prDS21	RF cloning pBAR3	CCGGGTGACCGATTCGGTAATCCCGGTAGAGGTGTGGTCAATAAG
prDS22	Linearize pDS101	TCCGAACAGAAGGAAGAAC
prDS23	Linearize pDS101	GATTACCGAATCGGTCAC
prDS55	Amplification & cloning SplicingLib1 index1	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAACATCTAAATACGAGGCACTTACTCCG
prDS56	Amplification & cloning SplicingLib1 index2	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGTGTTGGGAAATACGAGGCACTTACTCCG
prDS57	Amplification & cloning SplicingLib1 index3	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGAAGCCATGAATACGAGGCACTTACTCCG
prDS58	Amplification & cloning SplicingLib1 index4	AAAAGTGGAAGTCAGGGTGTTGGTGTAAAGGCTAAAGAAATACGAGGCACTTACTCCG
prDS59	Amplification & cloning SplicingLib1 R	ATTGTGGGGAGTGGAACGCAGTCACATTGATAGGAATAGCGAACTCCAGG
prDS62	Linearize pDS102	CTTTACACCAACACCCTGAC
prDS63	Linearize pDS102	TCAATGTGACTGCGTTCCAC
prDS137	NGS library preparation shift0	ACGACGCTCTTCCGATCTGTCAGGGTGTTGGTGTAAAG
prDS138	NGS library preparation shift1	ACGACGCTCTTCCGATCTAGTCAGGGTGTTGGTGTAAAG
prDS139	NGS library preparation shift2	ACGACGCTCTTCCGATCTTCGTCAGGGTGTTGGTGTAAAG

prDS140	NGS library preparation shift3	ACGACGCTCTTCCGATCTCATGTCAG GGTGTTGGTGTAAAG
prDS141	NGS library preparation shift4	ACGACGCTCTTCCGATCTACTAGTCA GGGTGTTGGTGTAAAG
prDS142	NGS library preparation shift5	ACGACGCTCTTCCGATCTTAGCCGTC AGGGTGTTGGTGTAAAG
prDS143	NGS library preparation R	AGACGTGTGCTCTTCCGATCTGTGGA ACGCAGTCACATTGA
prDS144	NGS library preparation PCR2	AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTC CGATCT
prDS145	NGS library preparation PCR2 with index	CAAGCAGAAGACGGCATACGAGAT [index]GTGACTGGAGTTCAGACGTGT GCTCTTCCGATCT