Sequence determinants and evolution of splicing in budding yeast and related species

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

Abstract

mRNA splicing is one of the key processes in eukaryotic gene expression. Most Intron-containing genes are constitutively spliced, and hence must undergo splicing in order to produce a functional mature mRNA. Therefore, regulation of splicing efficiency greatly affects broader gene expression regulation. Here we use a large synthetic oligo library of ~25,000 variants to explore how different intronic sequence determinants affect splicing efficiency and mRNA expression levels in yeast. We found that the three splice sites (donor, acceptor, and branching point) differ in how deviations from the consensus sequence affect functionality. We also use intronic sequences from other yeast species with modified splicing machinery to show that intron architecture has co-evolved with the splicing machinery to adapt to the presence or absence of a specific splicing factor. Finally, we show that synthetic sequences containing two introns give rise to diverse RNA isoforms, which enables us to elucidate intronic features that control and enable alternative splicing. Our study reveals novel mechanisms by which introns are shaped in evolution to allow cells to regulate their transcriptome.

Introduction

mRNA splicing has a major role in eukaryotic gene expression. During splicing, introns are removed from a pre-mRNA molecule to create 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, potentially contributing to proteomic diversity (Andreadis et al., 1987; Graveley, 2001; Kalsotra and Cooper, 2011; Nilsen and Graveley, 2010). However, most introns or exons are constitutively spliced (Chen et al., 2006; Ding and Elowitz, 2019) and therefore their contribution to gene expression is not through increasing 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 in eukaryotes can be a mechanism to regulate gene expression (Gotic et al., 2016) and therefore 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. 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 proteins (Parenteau et al., 2011). Hence, yeast cells tightly regulate the splicing process and devote lots of cellular resources for its accurate execution.

Previous studies addressed constitutive intron splicing efficiencies by examining existing natural introns. Recent analyses of splicing efficiency of yeast genes were either done by analyzing RNA-seq data (Douglass et al., 2019; Oesterreich et al., 2016; Xia, 2019), by studying a library of a reporter gene that contains natural introns from the *S. cerevisiae* genome (Yofe et al., 2014), or by investigating intron sequence features and examining intron evolution and conservation (Hooks et al., 2014; Schwartz et al., 2008). Other studies have utilized large

synthetic libraries to study how sequence features affect alternative splicing decisions (Baeza-Centurion et al., 2019; Cheung et al., 2019; Mikl et al., 2019; Rosenberg et al., 2015). In this work, we aim to systematically study cis-regulatory features that affect splicing efficiency of an intron by using a large synthetic oligo library . 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 etc (Kosuri et al., 2013; Levy et al., 2017; Maricque et al., 2018; Shalem et al., 2015; Sharon et al., 2012; Weingarten-Gabbay et al., 2016). Here we designed and synthesized (Agilent Technologies, CA) a library of more than 25,000 oligos, each consisting of a unique intronic sequence aimed to explore a range of sequence determinants that may affect splicing efficiency. We then measured splicing efficiency of each designed oligo using targeted RNA sequencing. Using this OL we cover 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). We find that the three sites have significantly different specificity in terms of regulating splicing efficiency. While the 3'SS is tolerant to mutations, and all the sequence versions found in the genome have a similar effect on splicing, for the BS the dominant consensus sequence from the genome is the only sequence that allows efficient splicing.

In addition, to synthetic modifications we tested natural introns from 11 different yeast species. This enables us to examine if introns that have been evolved to be spliced by their host splicing apparatus are functional in a different species, and to check how intron architecture co-evolves with changes in the splicing machinery. We show that *S. cerevisiae* has a tendency to misidentify 3' splice sites and produce cryptic splice isoforms.We suggest that this is related to the loss of a specific splicing factor in some yeast species.

Finally, we examine the potential of yeast to feature alternative splicing. In yeast, a vast majority of intron-containing genes have a single intron, and there are only a handful of examples of regulated alternative splicing in *S. cerevisiae* (Grund et al., 2008; Hossain et al., 2011; Juneau et al., 2009; Meyer et al., 2011). Hence, we ask how frequently will the splicing machinery produce multiple splice isoforms given a two-intron gene. Meaning, how easy it is for *S. cerevisiae* to produce alternative splicing if the right information is embedded within genes.
We designed a subset of the library to harbor a range of two tandem introns designs that can result in up to five different mature RNA isoforms. By studying these designs, and experimentally observing the alternative variants they actually produced, we obtained significant insights into the basic properties that support alternative splicing in *S cerevisiae* and possibly in a broader scope.

Results

High-throughput splicing efficiency measurements of thousands of synthetic introns

To explore how the intron architecture affects splicing efficiency, we designed a synthetic oligonucleotide library (LeProust et al., 2010; Sharon et al., 2012) of 23,418 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 cloned the library into a non-coding gene to avoid any differences between variants that might result from differences in translation.
Each oligo consists of fixed primers for amplification and cloning, a unique 12nt barcode, and a 158nt oligo with a unique intron sequence 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 gene 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 dubious open reading frame in *S. cerevisiae* genome using a high-throughput Cre-Lox based method (Levy et al., 2015) (Fig. 1A).

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 4 different subsets: the first subset was created in a combinatorial design by introducing different splice site sequences with their exact sequence as observed in the genome, on the background of the *MUD1* derived gene, at different positions along the variable region to create synthetic intron sequences. The second subset was based on perturbations to the former subset by introducing modifications to the genome-observed splice site sequences, or to the sequence adjacent to the splice sites, to alter the secondary structures at the splice sites. A third subset of the library was composed of introns that naturally occur in *S. cerevisiae* and in other yeast species. And finally, the fourth subset of variants was composed of designs with two short introns, one next to the other, separated by a short spacer sequence. This last subset enables us to study the potential of the *S. cerevisiae* splicing machinery to process genes with multiple intron and to create alternative splice variants.

Splicing efficiency of each variant in the library was measured using targeted RNA sequencing of the library's variable region. The sequence amplicon of the library 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 intorn design. Shortly, each variant was identified by its unique barcode then relative abundances of the spliced and unspliced isoforms were determined by aligning exon-intron and exon-exon junction sequences against the RNAseq reads. Splicing efficiency was defined as the ratio between the spliced isoform abundance and total RNA abundance (see Materials and

Methods). In addition, if for a specific RNA read, neither an unspliced isoform nor the spliced isoform of the designed intorn were identified, we searched for an mRNA isoform that might have been created by a novel splicing event. This was achieved by identifying uninterrupted gaps in the alignment of the full sequence to the RNAseq read.

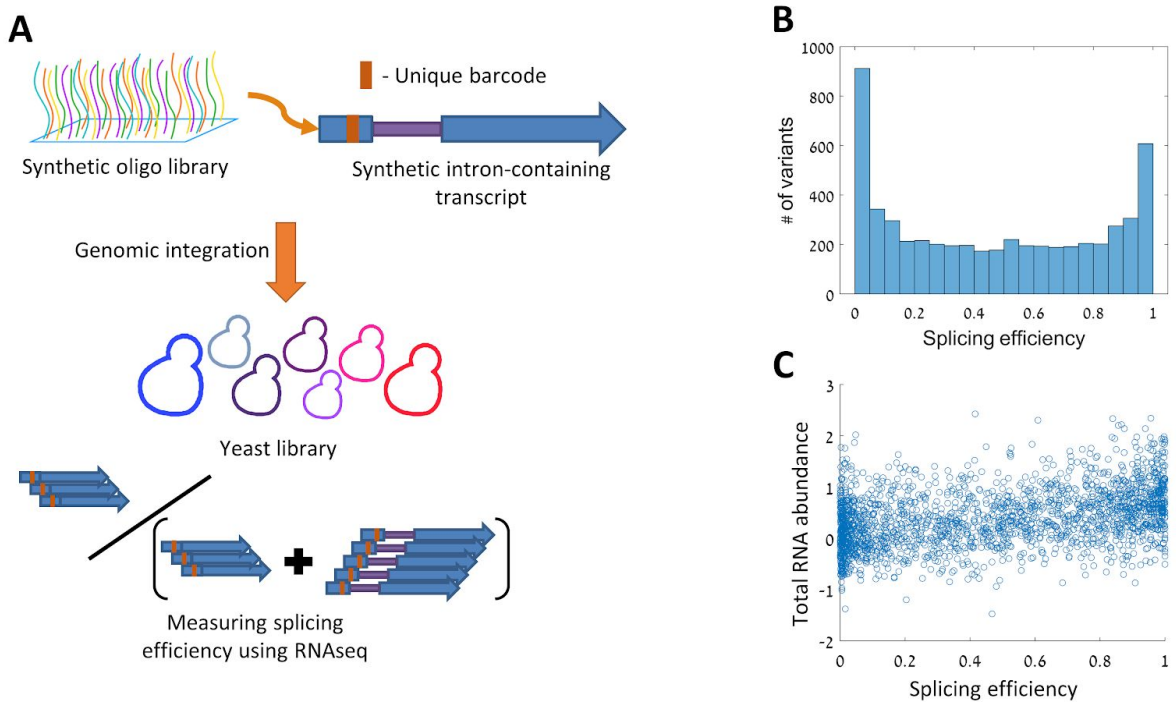

Figure 1 - A designed synthetic intron library in budding yeast features a wide range of splicing efficiency values

A. 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 using a high throughput Cre-Lox recombination method, to create a pooled yeast library. Then splicing efficiency was measured using targeted RNAseq of the intron's region, identification of RNAseq reads according to the barcode, and spliced isoform identification using alignment of exon-intron and exon-exon junctions. **B.** Distribution of splicing efficiency values for 5,383 variants that are observed with a spliced isoform. Splicing efficiency is defined, for each variant, as the number of reads for the spliced isoform, divided by the total RNA reads . **C.** Scatter plot of the total RNA abundance and splicing efficiency for all the variants with splicing efficiency >0 shows a significant positive correlation between RNA level and splicing efficiency (Pearson r=0.39 p-value<10^{-75}).

Synthetic introns are successfully spliced within the genomic construct

17,727 variants in the library were designed with a single intron (The remaining 5,691 are negative control variants or two-intron variants). For 34.5% of these single intron variants, we observed the designed splice isoform with median splicing efficiency of 0.435. We further see that splicing efficiency values have a bi-modal distribution, where ~10% of the variants show splicing of 100% of the RNA molecules, ~17% have close to, but higher than, zero splicing efficiency, and the other 73% are close to uniformly distributed with splicing efficiency values ranging between 0.05 to 0.95 (Fig 1B). For comparison, 4,224 variants were designed as negative controls. That is, variants which are not expected to be spliced, designed by introducing random sequences instead of splice sites. In this negative control set, only 1.5% yield a spliced isoform with a median splicing efficiency of 0.03. In addition, when examining the natural introns of *S. cerevisiae* included in our library, we see that 84% of them yielded spliced isoforms, with median splicing efficiency of 0.675.

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 for a randomly chosen set of 715 designs. When comparing the variance in splicing efficiency of these quartets (considering only designs with non-zero splicing efficiency) to a set of randomly chosen quartets, we see that the mean variance of the multiple barcode set is much lower than the mean variance of each of 10,000 randomly shuffled variants quartets from the same set (Fig S1A), indicating that the barcode choice exerts at most a low effect on splicing efficiency.

Splicing efficiency is positively correlated with RNA abundance

When examining total RNA abundance, we see a positive correlation between splicing efficiency and total RNA abundance (i.e. summed level of unspliced and spliced isoforms), for the set of

variants that are successfully spliced (Fig 1C). Since the calculation for splicing efficiency is dependent on total RNA abundance, we compare this result to a random set taken from the same distribution which does not present any correlation (Fig. S1B). This observation is interesting since in *S. cerevisiae* most genes do not contain an intron, so one might expect that the act of splicing would not affect the levels of total RNA. Moreover, this experiment was done on synthetic genes that were not selected in evolution to regulate their gene expression through splicing, which suggests that this correlation is explained by common molecular mechanisms. Nonetheless the 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.

<u>The three splice sites have a dramatically different effect on splicing efficiency</u>

The first set of features of intron sequence architecture which are considered essential for successful splicing are the 5' donor site (5'SS), the branch site (BS), and the 3' acceptor site (3'SS). As opposed to other eukaryotes that have degenerate splice site motifs, in most yeast species the 5'SS and the BS have a highly invariable consensus sequence, while in the 3'SS two possible sequences are prevalent in endogenous introns (Bon et al., 2003; Schwartz et al., 2008). Our library includes intron variants with all the splice site sequences that exist among natural *S. cerevisiae* introns. Using this set of variants we wish to check whether rare splice site sequences lead to reduced splicing efficiencies. We do so by splitting the library into groups according to each of the splice site sequences separately (5'SS, BS, and 3'SS) while keeping the other two sites with consensus sequences.

In the 5'SS we notice that the dominant consensus sequence (GUAUGU) leads to more efficient splicing than other sequences, but all other sequences found in the genome present a wide

range of splicing efficiencies, with a non-zero mean splicing efficiency for all of them (Fig. 2A). These splicing performances of non-consensus 5'SS are not in accordance with their low abundance in the genome. On the other hand, when examining different BS sequences we see a different phenomena, when only the consensus sequence (UACUAAC) allows efficient splicing, and most of the other possible sequences cause a failure in splicing (Fig. 2B). Finally, considering 3'SS sequences, the three possible sequences from *S. cerevisiae* genome are spliced with the same efficiencies, although in the genome two of them are much more prevalent (Fig. 2C).

In conclusion, we see different patterns of correspondence between abundance in the genome, and splicing performances in our synthetic intron library. The natural 5'SS sequences present a varying range of splicing efficiencies, with the consensus sequence performing the best. On the other hand for the BS and 3'SS we see two opposing patterns, while in the BS only the consensus sequence results in efficient splicing, in the case of 3'SS sequences all the natural site sequences are spliced with the same efficiencies. This raises a question, what drove the selection of a dual consensus for the genomic 3'SS.

To further investigate the effect of possible variations in the splice site sequences which are not present in the genome we designed a set of synthetic intron variants in which random mutations were introduced to the consensus splice sites' sequences (Fig. 2D-E, S2A). To check the effect of each nucleotide position in the splice sites separately, we examined all three possible mutations at a given position, and measured the proportion of successfully spliced variants for each of the three mutated nucleotides (Fig. 2D). First, we notice that this analysis supports the observation that the BS is sensitive to any deviation from its consensus site. Second, we see that in the 3'SS almost any single nucleotide mutation successfully produces spliced isoforms. Third, looking at 5'SS mutations we see that in positions 4 and 6, U→A, and U→C mutations

produce a spliced isoform, while all other mutations prevent splicing. This result suggests that the 3'SS is tolerant to a variety of mutations and that in contrast the 5'SS has only five possible sequences that leads to efficient splicing. Interestingly, these five sites are exactly the 5 sites that are present in the yeast genome.

In addition, we asked how mRNA folding energy around the splice site affects splicing efficiency. To this end, we perturbed the three splice sites by mutating adjacent nucleotides to increase mRNA folding free energy around the splice site, potentially embedding the splice site inside a stem within a stem-loop structure. As opposed to the direct splice site mutations, this secondary structure perturbation resulted in higher tolerance to perturbations in the 5'SS, while the 3'SS still presents some tolerance, and the BS is the least tolerant to tighter RNA secondary structure (Fig. S2B). This analysis suggests that reduced accessibility of each of the splice sites through RNA secondary structure, reduces the splicing efficiency, and that accessibility of the BS is more important for efficient splicing.

Finally, we examined positional interactions between the different splice sites. Since we observed that the BS is sensitive to any single nucleotide mutation, we focused here only on analyzing positions in 5'SS and 3'SS. We wished to check if the ratio of read counts of spliced variants of a double mutant can be predicted by the two ratios of the corresponding single nucleotide mutant sets of variants using a multiplicative model. For each pair of positions, we calculated an interaction coefficient which is defined as:

$$interaction(i,j) \;=\; PS_{i,j} - PS_i \cdot PS_j \qquad (Eq.\ 1)$$

$PS_{i,j}$ stands for the proportion of spliced variants of the double mutation variants at the respective position, and PS_i stands for the proportion of spliced variants of the single mutation variants (Fig. 2E). In this analysis, each value that is significantly different than zero means that there is a genetic interaction between the two positions in terms of its effect on splicing

efficiency that cannot be explained by a naive multiplicative model. A positive value indicates that the double mutant has a higher splicing efficiency than expected from the multiplication of efficiencies of its corresponding single mutants, and a negative value indicates the opposite. We notice that positions 4 & 6 in the 5'SS interact negatively with each other, and positively with position 1 in the 3'SS. In contrast, all three 3'SS positions interact positively with one another.

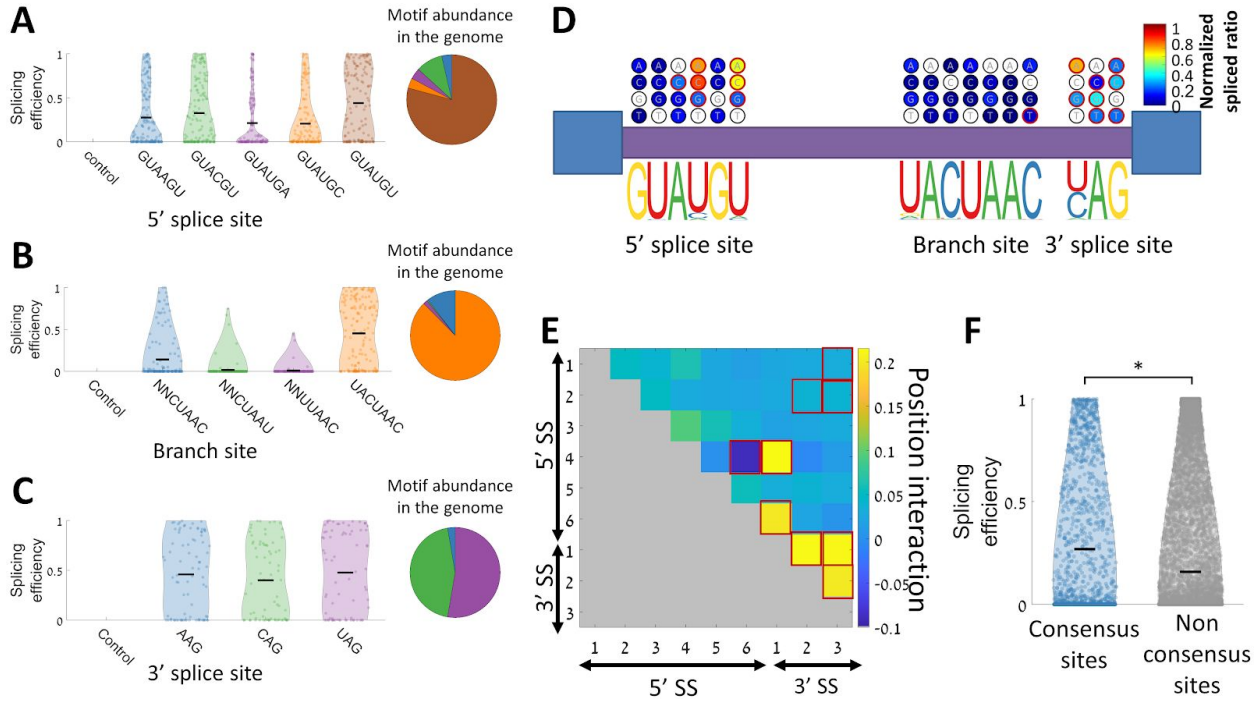

Figure 2 - The three different splice sites affect splicing efficiency differently

A-C. Splice site sequence effect of the three functional splice sites,(A) donor 5' splice site, (B) branch site, and (C) acceptor 3' splice site. In each sub-figure, the left panel violin plot represents the distribution of splicing efficiency values of all variants with the indicated site's sequence while the other splice sites hold a consensus sequences. The right panel pie-chart represents the relative abundance in the genome for each splice site sequence. **D.** 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. For each mutation, we present the normalized proportion of spliced variants, by dividing by the proportion of spliced variants for all the variants with consensus splice site sequences. Values significantly higher than zero are marked with a red circle (proportion z-test, with Bonferroni correction). Color bar represents the proportion of spliced isoforms in each group normalized by the same proportion for the consensus sites variants group. **E.** Double mutations interaction analysis. For each pair of positions in the 5' and 3' splice sites (numbered 1-6 and 1-3 respectively) we calculated an interaction coefficient as described in (Eq. 1). This interaction coefficient is presented using the color bar on the right. Positive values indicate that the double mutation variants are spliced better than expected by

the two single mutation variants, and negative values the other way around. Values for which zero is outside of the 99.9% confidence interval (computed using bootstrapping) are considered significant and marked with a red square. **F.** Distribution of splicing efficiency for all the variants with consensus splice sites at all three sites (left), compared to the distribution for non-consensus splice sites (right) (Wilcoxon ranksum test p-value=0.03).

Intronic length and nucleotide composition affect splicing efficiency

In the previous section, we described how splice site sequence content affects splicing efficiency, but even when analyzing a group of ~1,500 intron variants with consensus sequences at all three splice sites we see that only ~61% of them are spliced, and for those variants, splicing efficiency distribution spans across the entire range of zero to one (Fig. 2F). This observation raises questions about which other features in intron architecture affect splicing efficiency.

Previous studies have associated the following intronic features with splicing efficiency: intron length (Dewey et al., 2006; Wieringa et al., 1984), the distance between the branch site and the 3' end of the intron (BS-to-3' length) (Hooks et al., 2014; Luukkonen and Séraphin, 1997; Neuvéglise et al., 2011), and intronic GC content (Neuvéglise et al., 2011; Wong et al., 2013; Yofe et al., 2014). In addition, in higher eukaryotes, the 3' end of the intron is characterized by a polypyrimidine stretch just upstream of the 3'SS (Coolidge et al., 1997), however, in *S. cerevisiae* the polypyrimidine feature is much weaker, and is characterized by short uridine-rich sequence instead of pyrimidine (U or C) (Patterson and Guthrie, 1991; Schwartz et al., 2008). Comparing splicing efficiencies of intron groups split according to intron length, we noticed that up to an intron length of 121 nucleotides the mean splicing efficiency is increasing with intron length, but longer introns with lengths in the range 121-148 decreased the mean splicing efficiency compared to shorter introns (Fig. 3A). Hence, this analysis suggests that the *S. cerevisiae* splicing machinery has a preferred intron length. Interestingly, many of the ribosomal proteins in *S. cerevisiae* which are highly expressed, contain introns and all of them are

significantly longer than 121 nucleotides (Neuvéglise et al., 2011). Due to technical limitations of the oligonucleotide library synthesis, we could not include these ribosomal introns in our library, so we could not compare the performances of longer introns. We hypothesize that the ribosomal proteins' introns have a different architecture that supports efficient splicing of longer introns.

Next, we examined the contribution of the BS-to-3'SS distance to splicing efficiency (Fig. 3B). Notably, the group of intron variants with a 20-30 distance showed higher mean splicing efficiency, however longer BS-to-3'SS distances only slightly reduce the splicing efficiency. This result matches the fact that *S. cerevisiae* introns have a wide range of BS-to-3'SS distances.

We then turned to investigate the effect of intronic nucleotide composition on splicing efficiency. We observed a negative correlation between intronic GC content and splicing efficiency (Fig. 3C), as was previously observed in mammalian cells (Wong et al., 2013). Additionally, we examined the effect of U-rich sequence elements at the 3' end of the intron. In other eukaryotes introns are usually characterized by a polypyrimidine tract (PPT) at their 3' end, however, it has been previously reported that in *S. cerevisiae* and other yeast, the PPT contains specifically uridines (U), and not generally pyrimidines (Y, i.e. U or C) (Schwartz et al., 2008). We wish to examine whether U-rich elements have a stronger effect than Y-rich elements in *S. cerevisiae*. For that purpose, in every intron, we calculate the U content, and the Y content at a window of 20 nucleotides upstream, to the 3'SS, and define this measure as the U or Y enrichment score accordingly. We notice that the U enrichment score is highly correlated with splicing efficiency (Fig. 3D), while Y enrichment score has no significant correlation (Fig S3), suggesting that indeed the function of the PPT signal in *S. cerevisiae* is different than other eukaryotes and mediated specifically by a poly uridine tract.

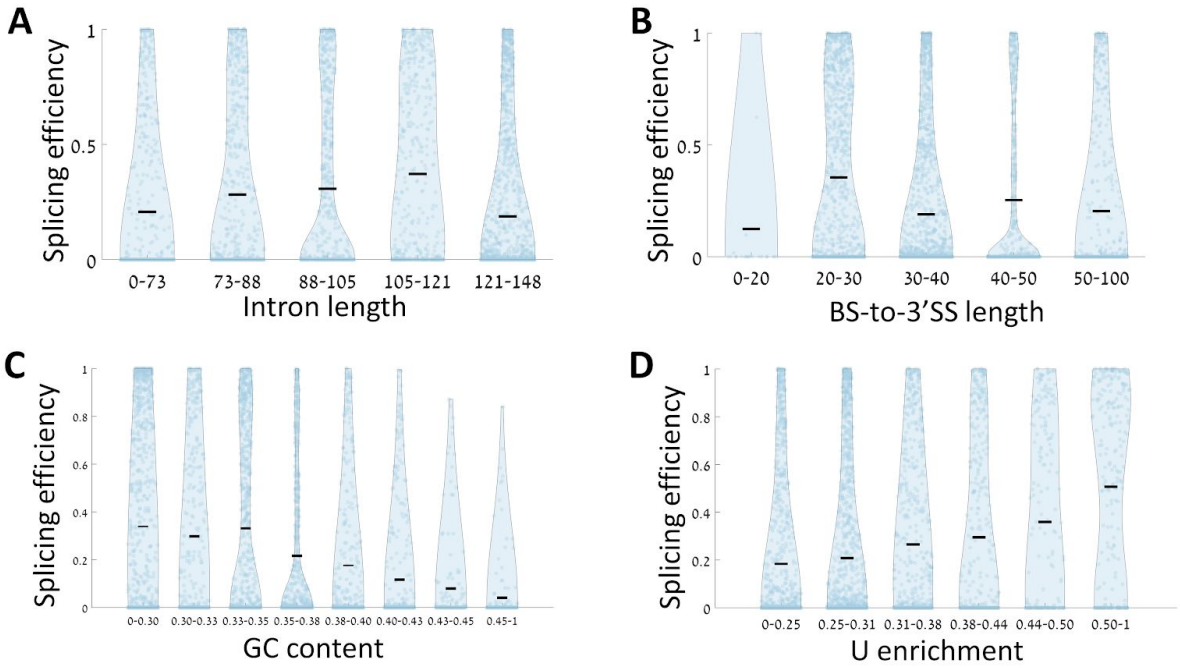

Figure 3 - Intron's parameters that affect splicing efficiency
A-B. Splicing efficiency distribution of spliced variants for (A) different intron lengths, or (B) different BS-to-3'SS length. **C.** Splicing efficiency distribution binned according to intronic GC content (Pearson r=-0.83 p-value=0.01). **D.** Splicing efficiency distribution binned according to poly uridine tract strength, 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 S3) (Pearson r=0.95 p-value=$3\cdot10^{-3}$).

Cryptic splicing events drive intron evolution

Up to this point, we have focused on designed splicing events. That is, successful splicing of the designed intron of each variant in the library. However, our designs might result in cryptic splicing isoforms, different than the ones we designed. To identify such splicing events, for each variant, we looked for cryptic spliced isoform 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 23.5% of the variants designed to have a single intron. This is a significant portion of the library, but the distribution of splicing efficiency values for the cryptic isoforms shows that most of them are spliced at low efficiency (Fig. 4A), with a median splicing efficiency of 0.04 for the cryptic splice isoforms, compared to the median splicing efficiency of 0.435 for the designed splice isoforms .

We then studied the location of the cryptic intron ends relative to the intended intron ends, we see that 87% of them have the same 5'SS as the designed intron, while only 1% of them have the same 3'SS (Fig. 4B). This observation suggests that a vast majority of cryptic splicing events are a result of misidentification of the 3'SS. This is in line with the observation that the 3'SS is tolerant to mutations (Fig. 2D,E, S2A). Previous work has also found that there are a significant number of alternative 3' splice site identification events in *S. cerevisiae* cells 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 observation, when analyzing our synthetic introns data, we considered the distribution of RNA free energy (ΔG) at a window of 30 nucleotides around the intended 3'SS. We notice that spliced variants with no cryptic splicing have more open 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 chosen alternative 3'SS we see that 50% of them are spliced at one of the three sequence motifs found in the genome (i.e. [T/C/A]AG) and that 66% 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. 4D).

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 appearance of a premature stop codon. 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. 4E).

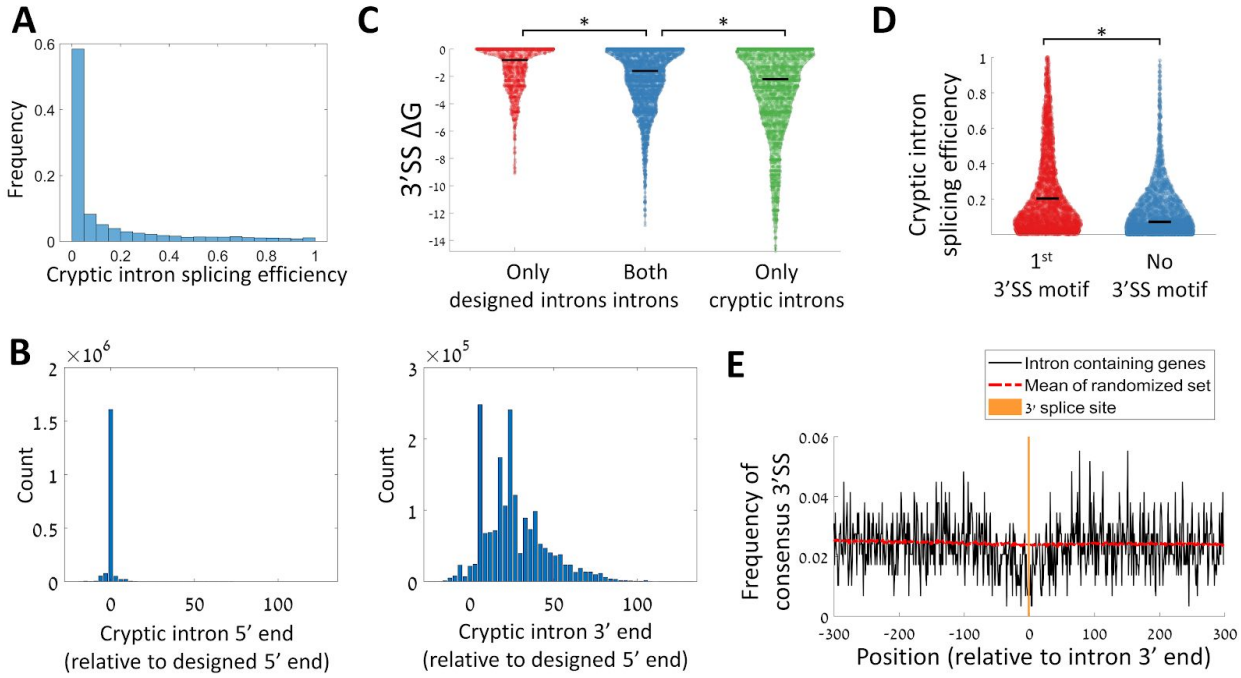

Figure 4 - Cryptic splice isoforms are produced due to 3'SS misidentification
A. Distribution of splicing efficiency values for cryptic introns. **B.** Relative splice site position distribution for cryptic introns (relative to the designed splice site position), for the 5' splice site (left), and the 3' splice site (right). **C.** Distribution of ΔG values at a window of 30 nucleotides around the designed 3'SS for vraints 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.** 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}). **E.** 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 TAG/CAG 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 the *S. cerevisiae* genome that can fit in our oligonucleotide design length constraint, (i.e. length limitation of 158 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. Next, for each natural *S. cerevisiae* intron, we included in our library introns from ortholog genes of the *S. cerevisiae* intron-containing genes from a set of 10 other yeast species, with orthology identified by (Hooks et al., 2014). We found that most *S. cerevisiae* endogenous introns are spliced in our system (85.5%), yet, introns from most of the other species are typically spliced at similar efficiencies (Fig. 5A). Furthermore, for each species we compared each one of its introns to the intron of its ortholog gene in *S. cerevisiae* and define ΔSE as the difference in splicing efficiency for introns of the same gene. Surprisingly, we found that many of the non *S. cerevisiae* introns in each species are spliced better than their *S. cerevisiae* orthologs (Fig. 5B), suggesting that *S. cerevisiae* introns are not specifically optimized for 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 much 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. One such 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). Interestingly, this gene is missing in 6 out of 11 of the yeast species we analyze here and in additional species (*T. blatae*) it is highly mutated and probably non-functional (Hooks et al., 2014). Indeed, 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). Furthermore, when comparing the distribution of splicing efficiencies between introns from species with or without a copy of U2AF1, we observed that introns that come from species lacking U2AF1 are better spliced in our *S. cerevisiae* system which also lacks U2AF1 (Fig. 5C) (*T. blattae* is excluded from this analysis). Hence we suggest that those introns were better optimized to splicing machinery lacking this factor, while introns that were adapted to machinery that use this factor 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, which leads to a selection against 3'SS motifs near introns 3' end. Interestingly, when performing the same analysis for the other 10 yeast species we see that only some of them demonstrate a similar signal to *S. cerevisiae* (Fig. S5). We then checked if there is a significant 3'SS motif avoidance in each of the 11 yeast species (Fig. 5D) and found that all but four species present *S. cerevisiae*-like avoidance signals. Strikingly, the 7 species that show the avoidance signal are those lacking U2AF1, and the one species with highly mutated copy of this factor. This result suggests 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, possibly due to tight constraints on the BS-to-3'SS distance.

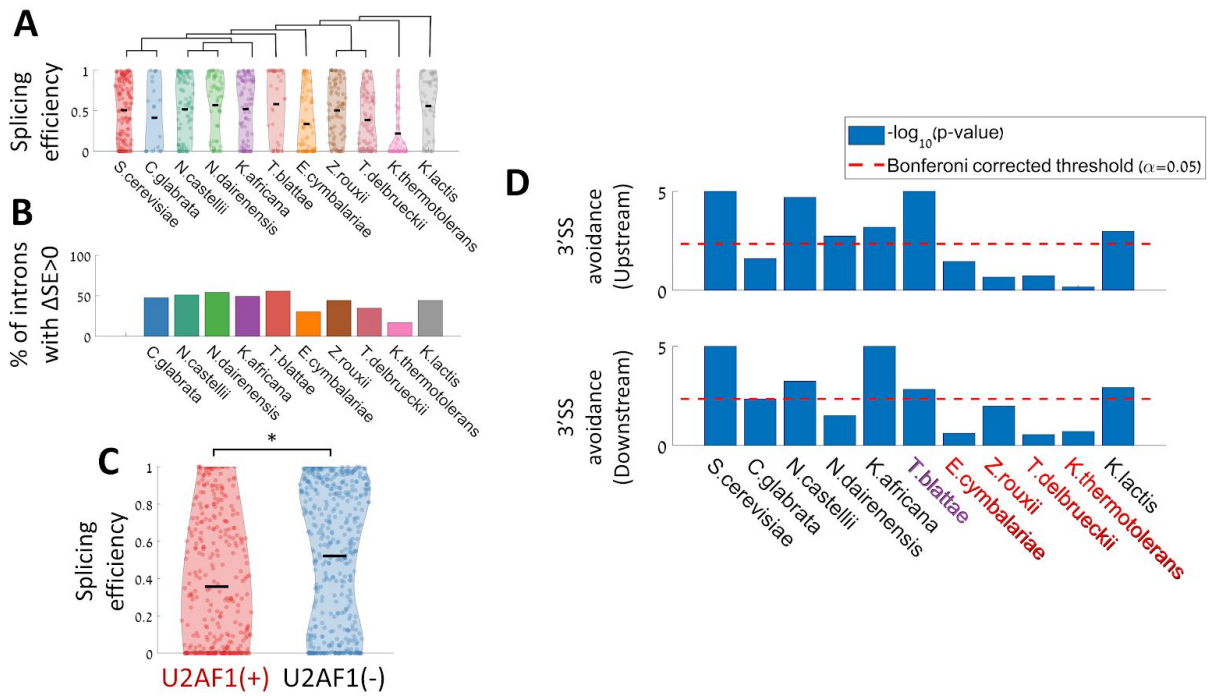

Figure 5 - Analysis of ortholog introns from other yeast species reveals intron architecture evolution

A. Splicing efficiency distribution of spliced variants for introns from orthologs of *S. cerevisiae* intron-containing genes. A species phylogenetic tree (created according to (Feng et al., 2017)) is presented below the yeast species names. **B.** Percent of introns that are spliced better than their *S. cerevisiae* ortholog's intron for each species. **C.** Splicing efficiency distribution for introns that come from species that have a copy of U2AF1 splicing factor (left), and introns that come from species without U2AF1 splicing factor (two sample 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. The P-value is given by the ratio of sets giving lower mean motif frequency than the tested introns set. 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.

<u>A computational model elucidates important features that govern splicing efficiency</u>

Using our synthetic library, 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. 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 75% of the single-intron library variants, excluding negative control variants that were designed with random splice sites (N=11,637). 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 at table S1). The model predictions were tested on the remaining 25% of the library (N=3,879). Predicted splicing efficiency values for the test set are well correlated with the measured splicing efficiency values (Pearson r=0.73, 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, moreover, in Figure 6B we present the individual feature contribution for each observation (i.e. library variant) of the 8 most important features according to this analysis, the distribution of Shapley values for each feature, and it's correspondence with the feature's values. We notice that the most important features are the sequences of the BS and 5'SS, yet, the 3'SS sequence is only the 8th ranked feature, in line with our observation that it can tolerate mutations and still maintain high splicing efficiency. In addition, in terms of local secondary structure, we see that only the local structure around the 3'SS is among the top 8 features, where we observe that open structure (high ΔG) around the 3'SS contributes positively to splicing efficiency, while the local structures around the 5'SS and BS are not found in the top 8 features contributing to splicing efficiency. Notably, the remaining features namely, intronic GC content, intron length, BS-to-3'SS distance, and the strength of the 3' U-rich element were observed as important also in analyses described above.

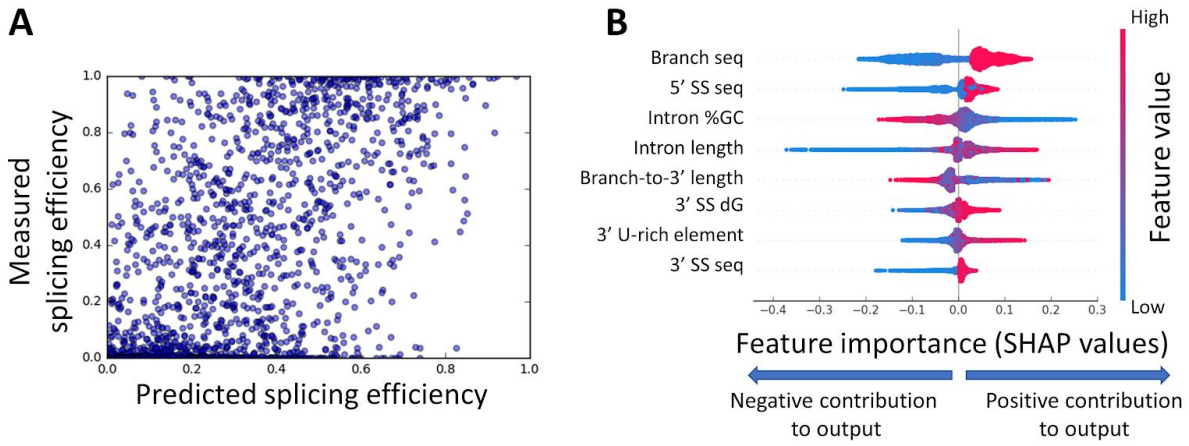

Figure 6 - Sequence features contribution to successful splicing are derived from a computational model

A. Measured splicing efficiency values versus predicted splicing efficiency values of the test set variants (N=3,879) as predicted by a gradient boosting model using 5-fold averaging cross-validation technique (Pearson r=0.73 p-value<10^{-100}). **B.** Distribution of Shapley values for the top 8 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, the higher its contribution to the model predictions. Positive values mean that the feature is predicted to improve splicing efficiency, and negative values means 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.

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

1,467 variants of the library included two short introns in tandem, one next to the other. In the *S. cerevisiae* genome, there are 10 known genes with two tandem introns (Clark et al., 2002), however they are not known to be alternatively spliced. Generally, since *S. cerevisiae* does not have many intron-containing genes, alternative splicing is not considered to have a major role in gene expression regulation in this species, and the known cases of alternative splicing are either a result of intron retention (Juneau et al., 2009) or alternative 3'SS (Meyer et al., 2011). Here, we tested whether *S. cerevisiae* has the potential to alternatively splice, and produce multiple spliced isoforms when given a two-intron gene.

When two introns are positioned next to each other the arrangement 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 observe all 4 spliced isoforms in our data (Fig 7B), the most abundant of which is the exon-skipping isoform.

When comparing the proportion of spliced variants for each intron sequence between variants in which it was placed as the first intron, and variants in which it was placed second, we see that those two measurements are in high agreement (Fig 7C), indicating that splicing efficiency of each intron is dependent mainly on its sequence, and less on the surrounding genetic context, or location relative to the other intron.

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 although *S. cerevisiae* lacks some of the auxiliary factors involved in alternative splicing, it still has the potential capacity to alternatively splice when more than a single intron is present in the pre-mRNA. To decipher which intron properties contribute to multiple spliced isoforms, we analyzed all the possible 100 pairs assembled from natural *S. cerevisiae* introns. We ranked each intron sequence according to the number of variants in which it was spliced in this subset. Multiple isoforms are observed mainly when both introns are ranked high. A single isoform is observed when one is ranked low and the other high. When a pre-mRNA consists of two introns that are ranked low, splicing is hardly observed (Fig 7D).

To further elucidate which features allow production of several isoforms from the same sequence, we use a computational model as was used for the single intron variants. In this case, the input features for the model are the same 39 features as the ones for the single intron but twice, as we took the set of features for each of the introns, in addition, GC content and intron length were calculated also for the exon skipping isoform, which yields a total of 80

features. Similarly to the single intron case, we train a gradient boosting model, but in this case, we use it to predict a multi-class classification problem, where each combination of isoforms is considered as a different class, and a variant is assigned to a class according to the isoforms observed for it, ignoring their relative frequency (we exclude the “Both introns” isoform from this analysis as it wasn’t observed in enough samples). We use 5-fold cross-validation technique, where each time we take 20% of the variants set as a validation set and train the model on the remaining 80%, the model performances are inferred by examining all the variants predictions when each of them was in the validation set. The model manages to classify correctly in 68% of the cases, with a weighted average F1 statistic of 0.656 . We infer the most important features for each class using a Shapley value analysis as was done for the single-intron model (Fig. 6B). In figure 7D we summarize this analysis by looking at 6 features that had a significant contribution in at least one of the classes. First, we can see that a sequence is predicted to be unspliced mainly if its first 5’SS and second BS contribute negatively, thus preventing successful splicing of any of the 3 introns. Second, we notice that in order to get only “Intron 1”, “Intron 2”, or “Exon skipping” isoform, the features for the spliced intron should contribute positively, but also the features unique for the other introns should contribute negatively. And finally, for a combination of spliced isoforms we see that all features should contribute positively, but specifically, the short intron length contributes the most. So, in order to get one of the short introns together with the “Exon skipping” intron, the short intron should be as long as it can, so it will have a chance to compete with the longer “Exon skipping” intron.

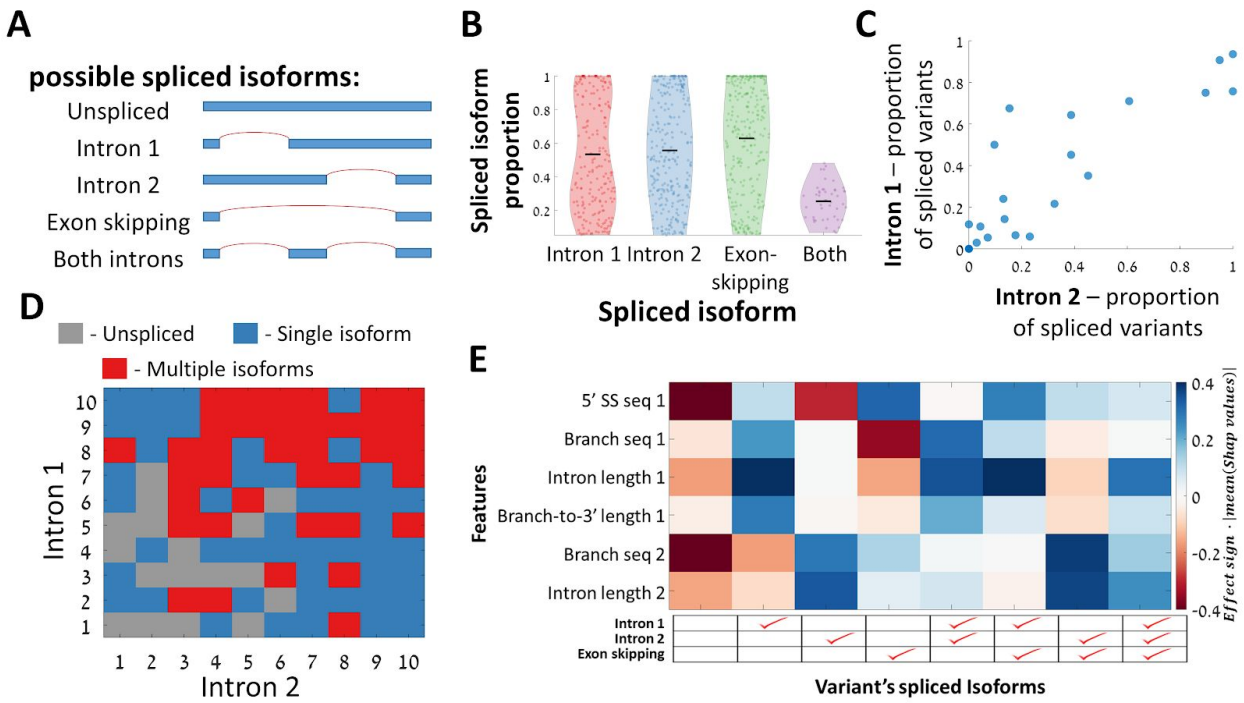

Figure 7 - Tandem two-intron designs demonstrate a capacity of *S. cerevisiae* to alternatively splice two introns.

A. Five possible isoforms can be observed for a two-introns 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 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 their proportion of spliced variants from low to high. **E.** Feature importance (Shapley values) derived from a gradient boosting model trained on a multi-class classification problem aiming to predict which spliced isoforms are observed for each variant. A set of Shapley values is calculated for each class (i.e. set of isoforms), and we present features that have a mean absolute value >3 standard deviations in at least one of the classes. For each feature and each class, we present the mean absolute value multiplied by an "effect sign" to indicate if the feature contributes positively/negatively to produce this isoform (for details on how we derive the "effect sign", see Materials and Methods)

Discussion

In this study, we used a designed, large synthetic oligonucleotide library to study the regulation of constitutive 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. 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 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.

Experimental evidence for intronic features contribution to splicing efficiency

We evaluated the effect of splice site sequences on splicing efficiency. Examining the highly conserved splice sites' sequences in *hemiascomycetous* fungi (Bon et al., 2003), one might assume that the consensus splice site sequences will be superior to any other sequence. However, our experimental data show that this is the case only for the branch site. Yet, for the 5' splice site we observe that the consensus site is spliced more efficiently, but other sequences with low representation in the genome still result in significant splicing efficiency. Moreover, in the case of the 3' splice site we observe that all three versions found in the genomes have identical effect on splicing efficiency. The 3'SS case raises the question, what caused the consensus sequences to be selected in evolution if just in terms of splicing efficiency they behave the same as non-consensus sites.

Another 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 more complicated, multicellular eukaryotes,, in *S. cerevisiae*, there is only a uridine enrichment (Schwartz et al., 2008). Here, we provide experimental evidence that indeed U-rich elements contribute to splicing efficiency, while general Y-rich elements do not. We suggest that this

might reflect evolutionary changes in the yeast splicing machinery that cause it to interact specifically with uridines.

Co-evolution of intron architecture and splicing machinery

By introducing into our library natural introns from 11 yeast species we could test if introns that come from different hosts which 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. 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 which lack U2AF1 splicing factor the splicing machinery misidentify the intended 3'SS due to similar sequence motifs near that site, which in turns creates a selective pressure to avoid such cryptic 3'SS motifs near the intended 3'SS. This hypothesis is supported by the fact that many cases of alternative 3'SS isoforms are observed both in this current study (Fig. 4A,B) and in previous findings in *S. cerevisiae* (Meyer et al., 2011; Plass et al., 2012).

Potential for regulated alternative splicing in yeast

Regulated alternative splicing is the focus of many studies on splicing because of its effect on increasing the proteomic diversity of the genome (Baeza-Centurion et al., 2019; Mikl et al., 2019; Rosenberg et al., 2015). In yeast, most intron-containing genes have a single intron, and there are very few known examples of functional alternative splicing (Grund et al., 2008; Hossain et al., 2011; Juneau et al., 2009; Meyer et al., 2011). However, using the relatively simplicity of the yeast splicing machinery, we aimed to decipher novel ways by which

alternatively 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 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 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.

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 3 sub-libraries that are separated in the initial amplification stage using different homology sequences.

All the experiments and data reported in this paper are based on one sublibrary with 21,714 variants (termed SplicingLib1), and another set of 1,704 intron-containing variants from a second sublibrary (termed SplicingLib2).

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 (TACTAAC), and three template sequences with two random

nucleotides at the first two positions, since non-consensus BS sequences differ greatly in these positions (NNCTAAC, NNCTAAT, NNTTAAC). 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 (ATTTTTAA, TTTAA, TAA). 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.

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* genes in a similar manner, and the remaining 7 based on random sequences. Another subset of 1,704 synthetic introns was created by randomly choosing sequences from non-intronic genic regions in the *S. cerevisiae* genome, and incorporating within them consensus splice sites with fixed length parameters (this subset was part of a different sublibrary SplicingLib2).

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). In addition, we also incorporated into the library a set of 200 randomly chosen introns from *S. pombe* that have *S. cerevisiae*-like splice sites. For the *S. cerevisiae* and *S. pombe* sets, 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' of the variable region. All possible pairs of intronic sequences were created and introduced to the library. In addition, for each two-intron variant, a corresponding variant of a joined long intron was created by removing the BS and 3'SS of the first intron, and the 5'SS of the second intron. These sequences were replaced by the corresponding sequence in the *MUD1* background sequence.

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 - two sublibraries

The data presented and analyzed in this work is derived from two sublibraries. The first, SplicingLib1, includes 24,510 variants, and the second SplicingLib2 includes 1,704 variants.

The sublibraries were amplified separately from the original pool using dedicated homology sequences. All the downstream assays including library cloning and transformation, RNA and DNA extraction, and next-generation sequencing were done separately on each of the sublibraries. The processed splicing data from both libraries were then combined and all the analysis was done on the combined data set.

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 sublibrary 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 (prDS45-48 for SplicingLib2, prDS55-58 for SplicingLib1)
- 1.5 ul - 10uM reverse primer (prDS49 for SplicingLib2, prDS59 for SplicingLib1)
- 200 pM of DNA oligo library
- 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 15 cycles
6. 72ºC 1 min

After amplification, the PCR product was cut from 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-Nage NucleoBon 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 yeast strain for which the library was inserted was based on SHA185, *S. cerevisiae* strain with Cre-Lox landing pad (derived from BY4709, ura3Δ ybr209w::GalCre-KanMX-1/2URA3 -lox66) kindly supplied to us by Sasha F. Levy's lab. 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, and $0.9 \cdot 10^5$ CFUs for SplicingLib2 which are ~60 and ~50 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 and grown for an additional 6 hours until they reached OD_{600} of 0.5, such 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:

7. 95ºC 3 min
8. 98ºC 20 sec
9. 58ºC 15 sec
10. 72ºC 15 sec
11. Repeat steps 2-4 for 20 cycles
12. 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, and 1.6, 2 million reads for the two RNA samples of SplicingLib2, and 0.25, 0.2 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 spliced if it was aligned to the *exon-exon* junction, and not aligned to the two *exon-intron* junctions. 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* junction. 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 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 S1). We took a set of 15,516 variants that includes all the designed single intron variants, excluding negative controls and the

1,704 variants set with identical intron features. 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.

<u>A computational model for predicting two-intron spliced isoforms</u>

We used gradient boosting classification to predict which set of isoforms will be produced from a two-intron variant. The gradient boosting implementation is based on LightGbm library for Python, and the feature importance inference is based on SHAP library for python.

We took a set of 1,414 two-intron variants, and each variant was characterized by the 39 features of each of its two introns (the same features as for the single intron variants, see table S1), in addition, the GC content and intron length of the exon-skipping intron were also added as features to create a set of 80 features (39 features for each of the two introns, plus 2 extra features for the “exon-skipping” intron). We then classified the variants according to the splice

isoforms seen for each variant in the data (excluding the 'both introns spliced' isoform because there were not enough cases for it). Each class is characterized by a set of isoforms that appear together for the same variant, the isoforms are namely, intron 1, intron 2, and exon-skipping.
In this classification task, we are interested specifically in the feature importance, so we trained a model on the entire set using 5-fold averaging cross-validation technique as described above. The Shapley values analysis outputs a set of Shapley values for each feature, for each of the 8 classes. We are looking for features that have a high contribution in at least one of the classes. So, we decided to look only at features that have an absolute mean Shapley value greater than 3 times the standard deviation of all Shapley values, in at least one of the features. We find 6 such features (Fig 7D). Then for each of these features in each of the classes, we present the absolute mean Shapley value which presents the total importance of this feature for this class, multiplied by the feature importance sign. A positive sign means that a variant that has a value in this feature which contributes positively to splicing efficiency, is more likely to be found in this class, while a negative sign, means that a variant that has a value in this feature that contributes negatively to splicing efficiency is likely to be found in this class. Feature importance sign is calculated by computing the Pearson correlation between the feature values and the Shapley values both for the single-intron set described above, and the two-intron set for each class described here. If the two correlations have the same sign (meaning, the feature has the same effect as in the single intron case), the feature importance sign is set to '1', otherwise, if the two correlations have opposing signs, the feature importance sign is set to '-1'.

Bibliography

Agilent Technologies (2016). G7555-90000 SureGuide Custom CRISPR Guide Library - Guidelines for Amplification and Cloning Assembly (Agilent Technologies).

Agilent Technologies (2017). G7556-90000 SureVector CRISPR Library Cloning Kit - Protocol (Agilent Technologies).

Andreadis, A., Gallego, M.E., and 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, J.M., Valcárcel, J., and Lehner, B. (2019). Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. Cell *176*, 549-563.e23.

Bond, S.R., and Naus, C.C. (2012). RF-Cloning.org: an online tool for the design of restriction-free cloning projects. Nucleic Acids Res. *40*, W209-13.

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuvéglise, C., Munsterkotter, M., Guldener, U., Mewes, H.-W., Van Helden, J., Dujon, B., et al. (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. *31*, 1121–1135.

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

Cheung, R., Insigne, K.D., Yao, D., Burghard, C.P., Wang, J., Hsiao, Y.-H.E., Jones, E.M., Goodman, D.B., Xiao, X., and 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, T.A., Sugnet, C.W., and Ares, M. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907–910.

Coolidge, C.J., Seely, R.J., and Patton, J.G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Res. *25*, 888–896.

Dewey, C.N., Rogozin, I.B., and Koonin, E.V. (2006). Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC Genomics *7*, 311.

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

Douglass, S., Leung, C.S., and Johnson, T.L. (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., and Tang, J. (2017). Reconstructing Yeasts Phylogenies and Ancestors from Whole Genome Data. Sci. Rep. *7*, 15209.

Friedman, J.H. (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., and 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., and Schibler, U. (2016). Temperature regulates splicing efficiency of the cold-inducible RNA-binding protein gene Cirbp. Genes Dev. *30*, 2005–2017.

Graveley, B.R. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. *17*, 100–107.

Grund, S.E., Fischer, T., Cabal, G.G., Antúnez, O., Pérez-Ortín, J.E., and 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.

Hooks, K.B., Delneri, D., and Griffiths-Jones, S. (2014). Intron evolution in Saccharomycetaceae. Genome Biol. Evol. *6*, 2543–2556.

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

Hossain, M.A., Rodriguez, C.M., and Johnson, T.L. (2011). Key features of the two-intron Saccharomyces cerevisiae gene SUS1 contribute to its alternative splicing. Nucleic Acids Res. *39*, 8612–8627.

Juneau, K., Nislow, C., and Davis, R.W. (2009). Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. Genetics *183*, 185–194.

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

Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat. Rev. Genet. *12*, 715–729.

Ke, G., Meng, Q., Finley, T., Wang, T., and Chen, W. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 3146.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc Natl Acad Sci USA *110*, 14024–14029.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., 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, E.M., Peck, B.J., Spirin, K., McCuen, H.B., Moore, B., Namsaraev, E., and Caruthers, M.H. (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., and 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, S.F., Blundell, J.R., Venkataram, S., Petrov, D.A., Fisher, D.S., and Sherlock, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature *519*, 181–186.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell.

Luukkonen, B.G., and 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.

Maricque, B.B., Chaudhari, H.G., and Cohen, B.A. (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, I.J., Pierce, D.A., and Lowe, H.W. (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., and 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., and Segal, E. (2019). Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. Nat. Commun. *10*, 4572.

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

Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature *463*, 457–463.

Oesterreich, F.C., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016).

Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372–381.

Palmiter, R.D., Sandgren, E.P., Avarbock, M.R., Allen, D.D., and Brinster, R.L. (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, R.J., Chabot, B., and Elela, S.A. (2011). Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. Cell *147*, 320–331.

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

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

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

Rosenberg, A.B., Patwardhan, R.P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. Cell *163*, 698–711.

Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and 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., and 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., and 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., and Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res. *24*, 1698–1706.

Stajich, J.E., Dietrich, F.S., and Roy, S.W. (2007). Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biol. *8*, R223.

Wang, H.-F., Feng, L., and 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, A.A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., and Segal, E. (2016). Comparative genetics. Systematic discovery of

cap-independent translation sequences in human and viral genomes. Science *351*.

Wieringa, B., Hofer, E., and 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.

Wong, J.J.-L., Ritchie, W., Ebner, O.A., Selbach, M., Wong, J.W.H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–595.

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., and Ben-Yehezkel, T. (2014). Accurate, model-based tuning of synthetic gene expression using introns in S. cerevisiae. PLoS Genet. *10*, e1004407.

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

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

Supplementary figures

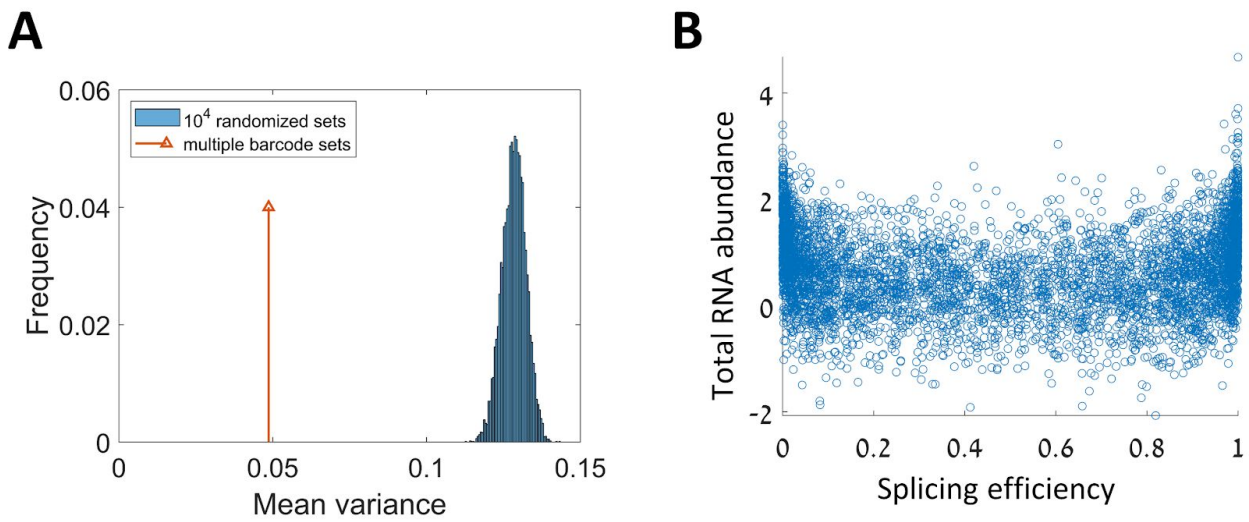

Figure S1 - Multiple barcodes, and randomized control for RNA abundance and splicing efficiency

A. 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. **B.** 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 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).

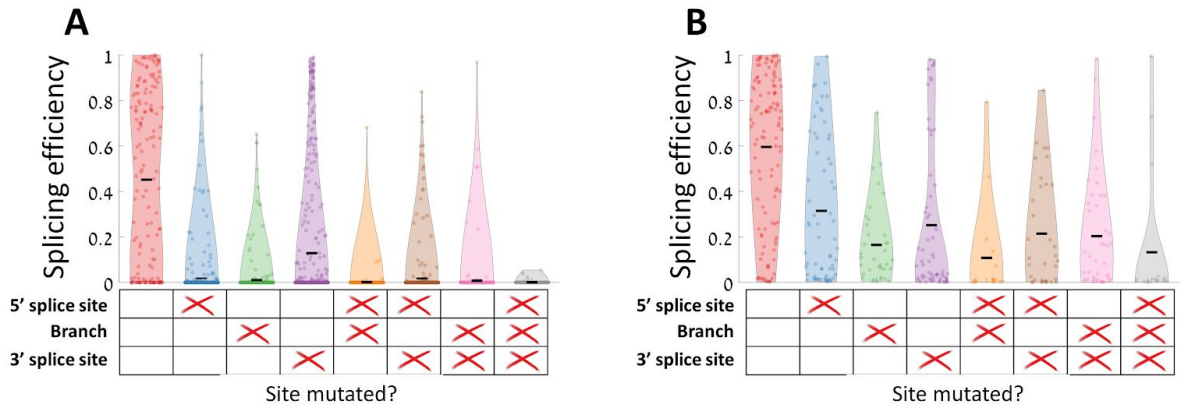

Figure S2 - Systematic mutagenesis to splice sites sequence, or splice site adjacent sequences
A. Distribution of splicing efficiency values for groups of variants with mutated splice sites. Variants are binned according to the identity of the mutated site as shown in the bottom table.
B. Similar to (A), only here each site is perturbed by mutating the surrounding region to create a stem-loop structure at the splice site, the bottom table indicates the identity of sites within stem-loop structure.

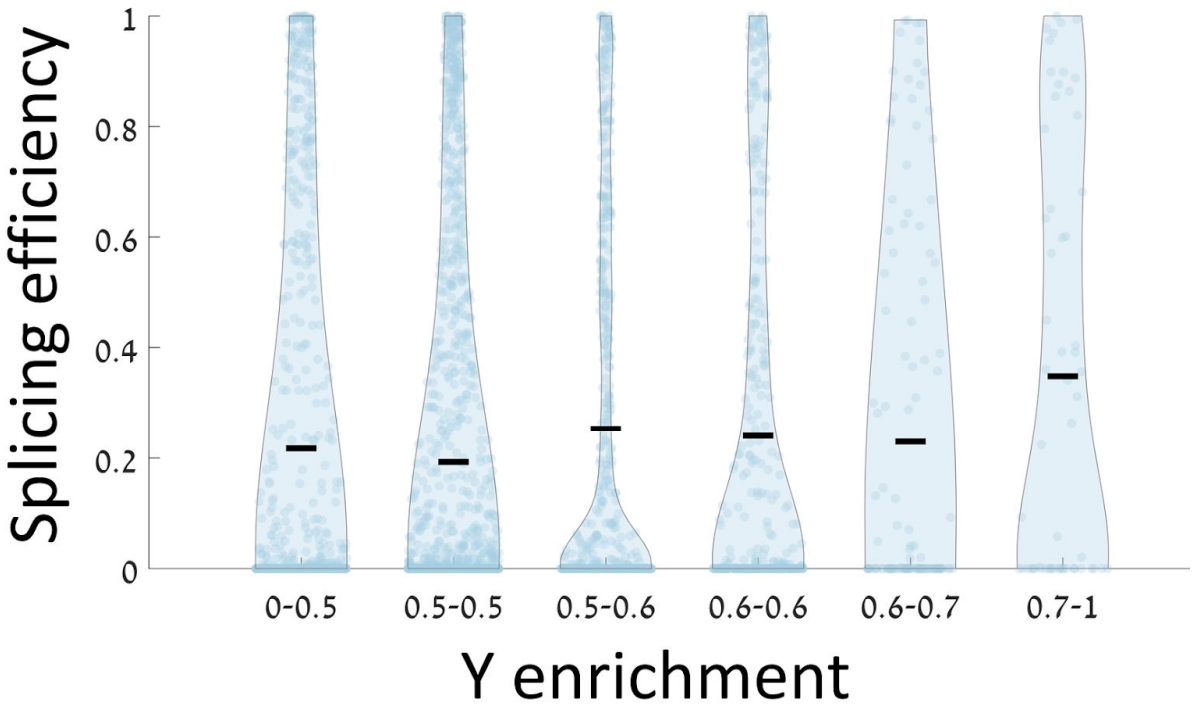

Figure S3 - Splicing efficiency is not correlated with Y-enrichment at intron's 3' end
Splicing efficiency distribution is binned according to poly-pyrimidine tract strength , which is calculated as the Y (i.e. C or U) content at a window of 20 nucleotides upstream to the 3'SS. In order to specifically 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=0.16), compared to the highly significant correlation for U-rich elements (Fig 3D).

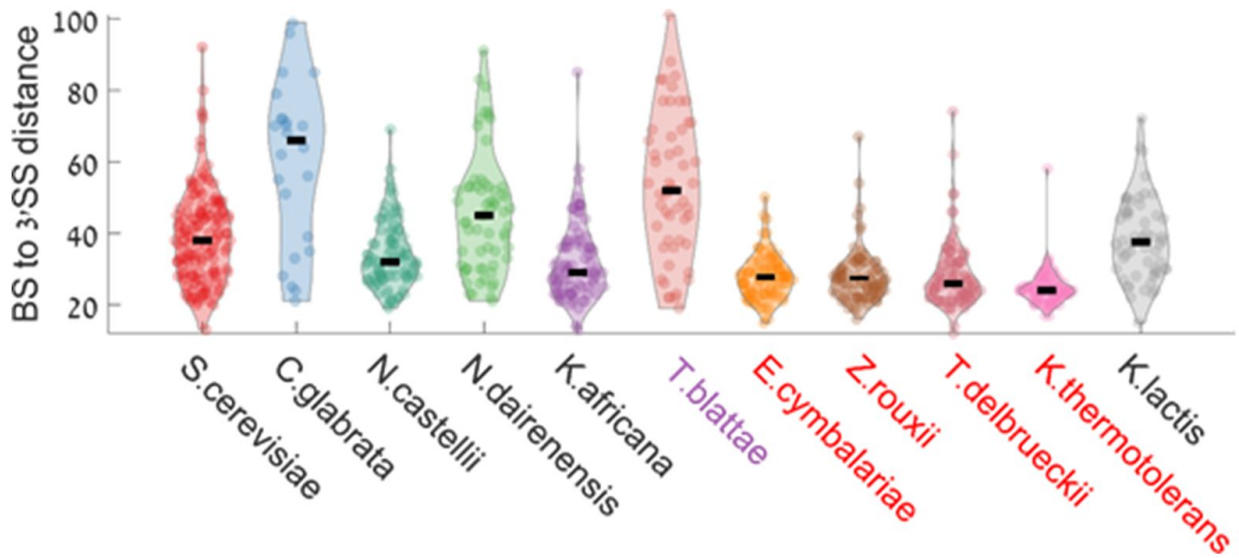

Figure S4 - BS-to-3'SS distance distribution is associated with existence U2AF1 splicing factor
Distribution of BS-to-3'SS distance in each of the 11 species. Species with no copy of the gene coding for the splicing factor U2AF1 are marked in black, one species with a malfunctioned copy of U2AF1 is marked in purple, and the ones with a copy of U2AF1 are marked in red.

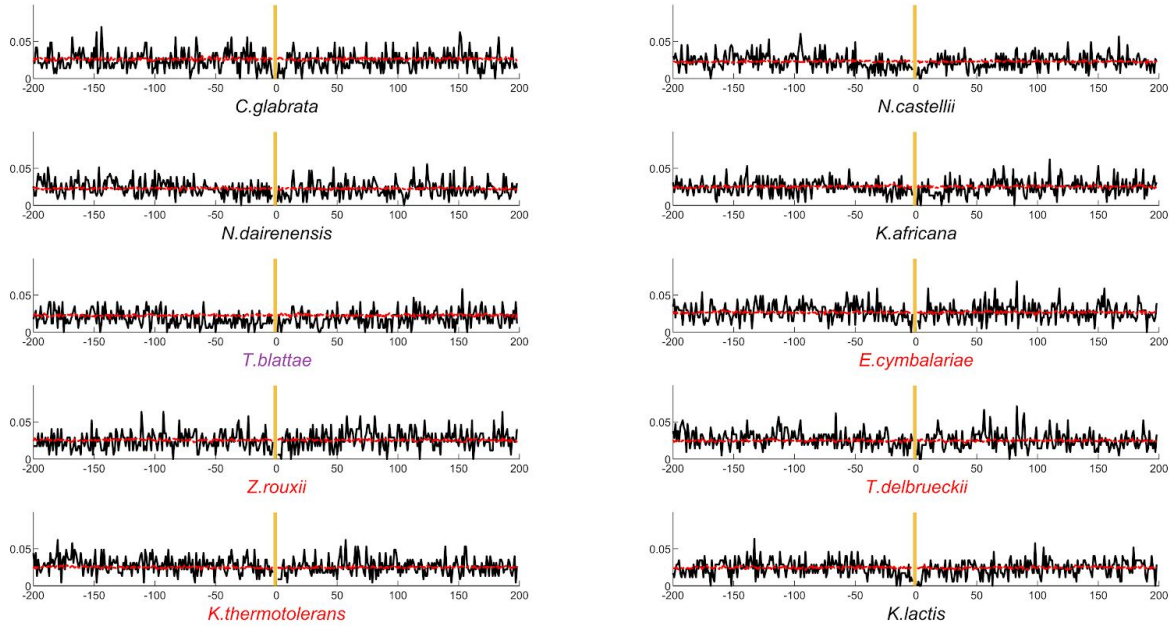

Figure S5 - 3'SS motif avoidance signal for 10 yeast species

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 4E.

Supplementary tables

Table S1 - 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 S2 - 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

prDS45	Amplification & cloning SplicingLib2 index1	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGAACATCTAAACGCACGACAAA CATCAGG
prDS46	Amplification & cloning SplicingLib2 index2	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGTGTTGGGAAACGCACGACAAA CATCAGG
prDS47	Amplification & cloning SplicingLib2 index3	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGAAGCCATGAACGCACGACAAA CATCAGG
prDS48	Amplification & cloning SplicingLib2 index4	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGGCTAAAGAAACGCACGACAAA CATCAGG
prDS49	Amplification & cloning SplicingLib2 R	ATTGTGGGGAGTGGAACGCAGTCAC ATTGATCGCCTAAACGAAGAAACCG
prDS55	Amplification & cloning SplicingLib1 index1	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGAACATCTAAATACGAGGCACT TACTCCG
prDS56	Amplification & cloning SplicingLib1 index2	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGTGTTGGGAAATACGAGGCACT TACTCCG
prDS57	Amplification & cloning SplicingLib1 index3	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGAAGCCATGAATACGAGGCACT TACTCCG
prDS58	Amplification & cloning SplicingLib1 index4	AAAAGTGGAAGTCAGGGTGTTGGTG TAAAGGCTAAAGAAATACGAGGCACT TACTCCG
prDS59	Amplification & cloning SplicingLib1 R	ATTGTGGGGAGTGGAACGCAGTCAC ATTGATAGGAATAGCGAACTCCAGG
prDS62	Linearize pDS102	CTTTACACCAACACCCTGAC
prDS63	Linearize pDS102	TCAATGTGACTGCGTTCCAC
prDS137	NGS library preparation shift0	ACGACGCTCTTCCGATCTGTCAGGGT GTTGGTGTAAAG
prDS138	NGS library preparation shift1	ACGACGCTCTTCCGATCTAGTCAGG GTGTTGGTGTAAAG
prDS139	NGS library preparation shift2	ACGACGCTCTTCCGATCTTCGTCAGG GTGTTGGTGTAAAG
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	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
prDS145	NGS library preparation PCR2 with index	CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT