Nucleotide sequence composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi

ZOHAR ZAFRIR[1] and TAMIR TULLER[1,2]

[1]Department of Biomedical Engineering, Tel Aviv University, Tel Aviv 69978, Israel
[2]The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

RNA splicing is the central process of intron removal in eukaryotes known to regulate various cellular functions such as growth, development, and response to external signals. The canonical sequences indicating the splicing sites needed for intronic boundary recognition are well known. However, the roles and evolution of the local folding of intronic and exonic sequence features adjacent to splice sites has yet to be thoroughly studied. Here, focusing on four fungi (*Saccharomyces cerevisiae, Schizosaccharomyces pombe, Aspergillus nidulans,* and *Candida albicans*), we performed for the first time a comprehensive high-resolution study aimed at characterizing the encoding of intronic splicing efficiency in pre-mRNA transcripts and its effect on intron evolution. Our analysis supports the conjecture that pre-mRNA local folding strength at intronic boundaries is under selective pressure, as it significantly affects splicing efficiency. Specifically, we show that in the immediate region of 12–30 nucleotides (nt) surrounding the intronic donor site there is a preference for weak pre-mRNA folding; similarly, in the region of 15–33 nt surrounding the acceptor and branch sites there is a preference for weak pre-mRNA folding. We also show that in most cases there is a preference for strong pre-mRNA folding further away from intronic splice sites. In addition, we demonstrate that these signals are not associated with gene-specific functions, and they correlate with splicing efficiency measurements ($r = 0.77$, $P = 2.98 \times 10^{-21}$) and with expression levels of the corresponding genes ($P = 1.24 \times 10^{-19}$). We suggest that pre-mRNA folding strength in the above-mentioned regions has a direct effect on splicing efficiency by improving the recognition of intronic boundaries. These new discoveries are contributory steps toward a broader understanding of splicing regulation and intronic/transcript evolution.

Keywords: **pre-mRNA folding; splicing efficiency; gene expression; intron evolution; transcript evolution**

INTRODUCTION

RNA splicing is the process in which introns, intervening noncoding fragments within pre-mRNA transcripts, are removed, leaving retained coding regions termed exons and untranslated regions (UTRs). In eukaryotes spliceosomal introns are confined to the nucleus and the splicing process is executed by the spliceosome, one of the largest molecular complexes in the cell (Nilsen 2003; Hoskins and Moore 2012). Accurate processing of introns is a crucial regulatory step in determining the cell expression profile and is required before protein translation can be initiated. As such, splicing efficiency (SE; efficient recognition and proper splicing by the spliceosome) and specificity are used to regulate growth, development, and overall response to external signals (Le Hir et al. 2003; Nasim and Eperon 2006; Wang and Cooper 2007; Khodor et al. 2012). Extensive studies in the last two decades have revealed the core chemical reactions, several sequence determinants, and the major protein component interactions during intron splicing (Krämer 1996; McKee and Silver 2007; Toor et al. 2008; Wahl et al. 2009). Yet, the debate on intron origin and evolution has been ongoing intensively for decades (Rodríguez-Trelles et al. 2006; Rogozin et al. 2012).

The dynamic nature of spliceosome assembly and the complex interactions of both its proteins and RNA components with the pre-mRNA give rise to a range of intronic SEs among different genes and within the same gene that

Abbreviations: AA, amino acid; AS, alternative splicing; BS, branch site; CUB, codon usage bias; FDR, false discovery rate; GO, gene ontology; LFE, local folding energy; ORF, open reading frame; PPT, polypyrimidine tract; PSSM, position specific scoring matrix; SE, splicing efficiency; SNR, signal to noise ratio; SS, splice site; STD, standard deviation; UTR, untranslated region; YFP, yellow fluorescent protein

Corresponding author: tamirtul@post.ac.il

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.051268.115.

may facilitate the creation of different mature mRNA products from identical pre-mRNA transcripts (Maniatis and Tasic 2002). This phenomenon is termed alternative splicing (AS); in this process particular exons of a gene may be included within or excluded from (i.e., skipped) the final processed mRNA transcript produced from that gene. Additional AS events include intron retention and alternative splice sites selection, where the combination of two or more events can generate more complex mature mRNA products (Ast 2004; Ner-Gaon et al. 2004; Barbazuk et al. 2008; Kim et al. 2008; Keren et al. 2010; McManus and Graveley 2011). Therefore, AS is a chief constituent of the increased number of proteins encoded by the genomes of multicellular organisms, relative to their gene number (Black 2003).

The amount of intron-containing genes, intron density per gene or per kilobase pair, and intron length, varies from one eukaryote to another (Kriventseva and Gelfand 1999; Lim and Burge 2001; Yu et al. 2002; Alexander et al. 2011). In *Saccharomyces cerevisiae*, there are less than 300 intron-containing genes (5% of ∼6000) of which few are mediated by two introns or more (Spingola et al. 1999; Grate and Ares 2002). In fact, until now, only a single AS event has been identified in this organism (Mishra et al. 2011). Nevertheless, in *S. cerevisiae* the intronic genes are highly expressed and account for >70% of its proteome; in addition, several intron sequences are known to be duplicated within ribosomal protein paralogs (Ares et al. 1999; Juneau et al. 2006). Conversely, in *Schizosaccharomyces pombe* (an organism with ∼5000 genes) there are ∼5000 introns spreading over one-third of its genome, with up to 20 introns mediating a single gene (Wood et al. 2002) and with few known genes displaying AS (Habara et al. 1998; Okazaki and Niwa 2000; Marshall et al. 2013). Other fungi such as *Aspergillus nidulans* and *Candida albicans* show variations in intronic characteristics as well (Kupfer et al. 2004; Mitrovich et al. 2007). Nonetheless, even though eukaryotic evolution has been generally characterized by widespread intron gain and loss events (Jeffares et al. 2006; Roy and Gilbert 2006; Carmel et al. 2007; Stajich et al. 2007; Rogozin et al. 2012; Zhu and Niu 2013), the ubiquity of introns and the core of the spliceosome are conserved in all well-characterized eukaryotes (Nilsen 2003; Collins and Penny 2005). Moreover, *Hemiselmis andersenii* is currently the only known eukaryotic organism without any introns or spliceosome subunit genes (Lane et al. 2007).

Some previous studies have investigated the efficiency of splicing either for a specific intron or systematically for all introns, and under a variety of environmental conditions or in varying genetic backgrounds (Pleiss et al. 2007; Bergkessel et al. 2011; Pérez-Valle and Vilardell 2012). Such studies have demonstrated that different introns exhibit a large range of SEs under varying conditions and have diverse proteinaceous requirements (Clark et al. 2002).

It is known that at the intronic donor and acceptor splice sites (SS; the 5′SS and 3′SS, respectively), at the branch site (BS; the region surrounding the branch point), and at the polypyrimidine tract (PPT), there are canonical sequence elements which are essential for intron recognition and for splicing to occur. The factors that bind to these sequence motifs and the biochemical reactions which they perform are relatively well known due to the extensive research in this field. Systematic investigations show that this process is highly regulated: from spliceosome assembly, through pre-mRNA recognition and binding, to the splicing reaction and complex disassembly (Warf and Berglund 2010; McManus and Graveley 2011). Additionally, it has been suggested that in yeast, introns regulate ribosome biogenesis and functions and affect cell fitness under stress (Parenteau et al. 2011). Other small-scale studies in yeast and mammals have indicated that pre-mRNA secondary structure and nucleotide composition in the region between the BS location and 3′SS can affect 3′SS selection (Mougin et al. 1996; Gahura et al. 2011; Meyer et al. 2011; Plass et al. 2012).

However, it has not been shown that evolution shaped the pre-mRNA secondary structure near splice sites and what the exact regions under such a selection are; moreover, the effect of this evolutionary process on SE at the genomic level has not been estimated. Here we aim at providing answers to these questions at a genomic level.

RESULTS

In this study, we analyze the intronome of four fungi: *S. cerevisiae, S. pombe, A. nidulans,* and *C. albicans.* Specifically, the analyzed data include 280 introns from *S. cerevisiae*, 4747 introns from *S. pombe*, 2427 introns from *A. nidulans*, and 391 introns from *C. albicans* (for further details regarding these organisms see the Materials and Methods section). Our objective was to evaluate systematically how pre-mRNA folding in the proximity of both splice sites promotes regulation of SE, toward a better understanding of intron evolution. To this end, we defined four pre-mRNA exonic and intronic regions that are used throughout the paper: Exonic Donor, Intronic Donor, Intronic Acceptor, and Exonic Acceptor, as defined in Figure 1; the pre-mRNA exon–intron boundaries, consensus sequences, and the assigned domains are illustrated also. Here, we focused on the role of local pre-mRNA secondary structure strength in regulation of SE. Hence, Figure 1 also includes the randomized models used to provide evidence of selection and evaluate the preference level over the pre-mRNA folding adjacent to intronic splice sites (more details are provided in the following sections).

Indication for weak secondary structure and low GC content adjacent to intronic splice and branch sites

It has been shown previously that GC content and thermodynamic patterns affect exon–intron splice site recognition, and that increased pre-mRNA secondary structure promotes AS in multicellular higher eukaryotes including human (Shepard and Hertel 2008; Zhang et al. 2011; Amit et al. 2012; Nedelcheva-Veleva et al. 2013). Yet, evolutionary preference for weak or strong pre-mRNA secondary structure has never been studied systematically in fungi or other organisms. In addition, the possibility that local structure stability could influence SE may be underestimated; specifically, its effect on SE has not yet been evaluated at a genomic scale. Finally, the exact regions under selective pressure for secondary structure have yet to be inferred. Therefore, here we focused on the exon–intron boundaries, i.e., the regions surrounding the donor and acceptor splice sites, and aimed to systematically infer regions (at a single nucleotide resolution) that have a preference for weak/strong pre-mRNA secondary structure and/or low/high GC content. To this end, we used a sliding window scheme with varying window sizes (30–50 nt, the approximate region surrounding the exon–intron junction where the major splicing complexes interact [Le Hir et al. 2000; Reichert et al. 2002; Hoskins et al. 2011]) to analyze local pre-mRNA folding energy (LFE) and GC content; in most analyses, a profile was computed for each pre-mRNA transcript and then averaged over the entire intronome (see Materials and Methods for more details).

Figure 2 shows the mean assembled profiles over the intronome, aligned to the 5′SS (left), BS (middle), and 3′SS (right) using a sliding window size of 40 nt. As can be seen, for all analyzed organisms there is a clear ascent in the LFE (corresponding to weaker local folding) near the donor, acceptor, and branch sites of the introns, as well as a decrease in GC content near these regions; further away from the splice and branch sites toward the exonic/intronic domains, the LFE and GC content are generally lower and higher, respectively (distance from 5′SS/BS/3′SS is relative to the sliding window's center; the two clear GC content peaks located at ±20 nt from the splice sites are due to the donor and acceptor consensus sequences). In addition, on the intronic ends near both splice sites and surrounding the BS there is a notable peak in LFE that cannot be explained by

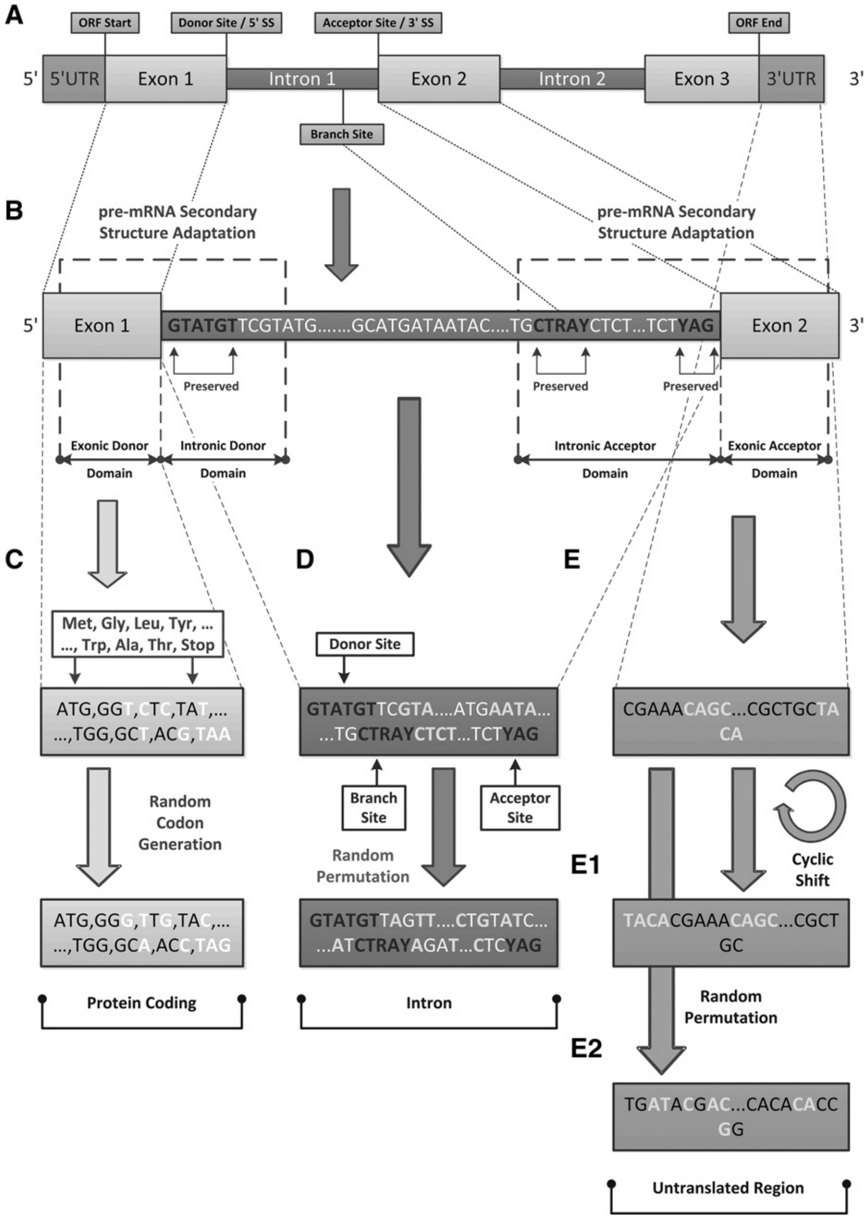

FIGURE 1. pre-mRNA exonic and intronic regions, basic definitions, and randomization models. (*A*) The analyzed fungal genes can be divided into three major regions: untranslated regions (UTRs), exons, and introns (we did not consider UTR introns since in the analyzed organisms few introns appear in the UTRs; e.g., <6% in the case of *S. cerevisiae*). (*B*) The introns include three canonical consensus sequences: the donor (or 5′SS; subsequence *GTATGT*) and acceptor (or 3′SS; subsequence *YAG*) that define the intronic boundaries, and the branch site (BS; subsequence *CTRAY*) that is required for the lariat formation; those sequences are preserved in all our randomization models. In our analyses, exons and introns were divided into four domains: Exonic Donor (the exonic sequence up to 200-nt upstream of the 5′SS), Intronic Donor (the intronic sequence up to 200-nt downstream from the 5′SS), Intronic Acceptor (the intronic sequence up to 200-nt upstream of the 3′SS), and Exonic Acceptor (the exonic sequence up to 200-nt downstream from the 3′SS); in cases of introns/exons shorter than 200 nt, we considered the entire intron/exon. The features surrounding the boundaries of these regions are studied here. (*C–E*) In order to demonstrate that the reported features are under selective pressure in endogenous transcripts, we compared the intronic sequences to the ones obtained by the following randomized models: (*C*) encoded protein information is maintained; synonymous codons are generated based on their whole-genome codon frequencies; (*D*) uniform permutation of intronic nucleotides; (*E*) UTR randomization maintains GC content using cyclic shift (*E1*) and uniform nucleotide permutation (*E2*); all the randomization models preserve these consensus sequences (5′SS/BS/3′SS) as well as additional exonic and intronic characteristics (see Materials and Methods for more details). The results provide evidence of selection for weak folding around the donor, acceptor, and branch sites, as well as a preference for strong folding in most of the exonic and intronic domains adjacent to the splice sites.

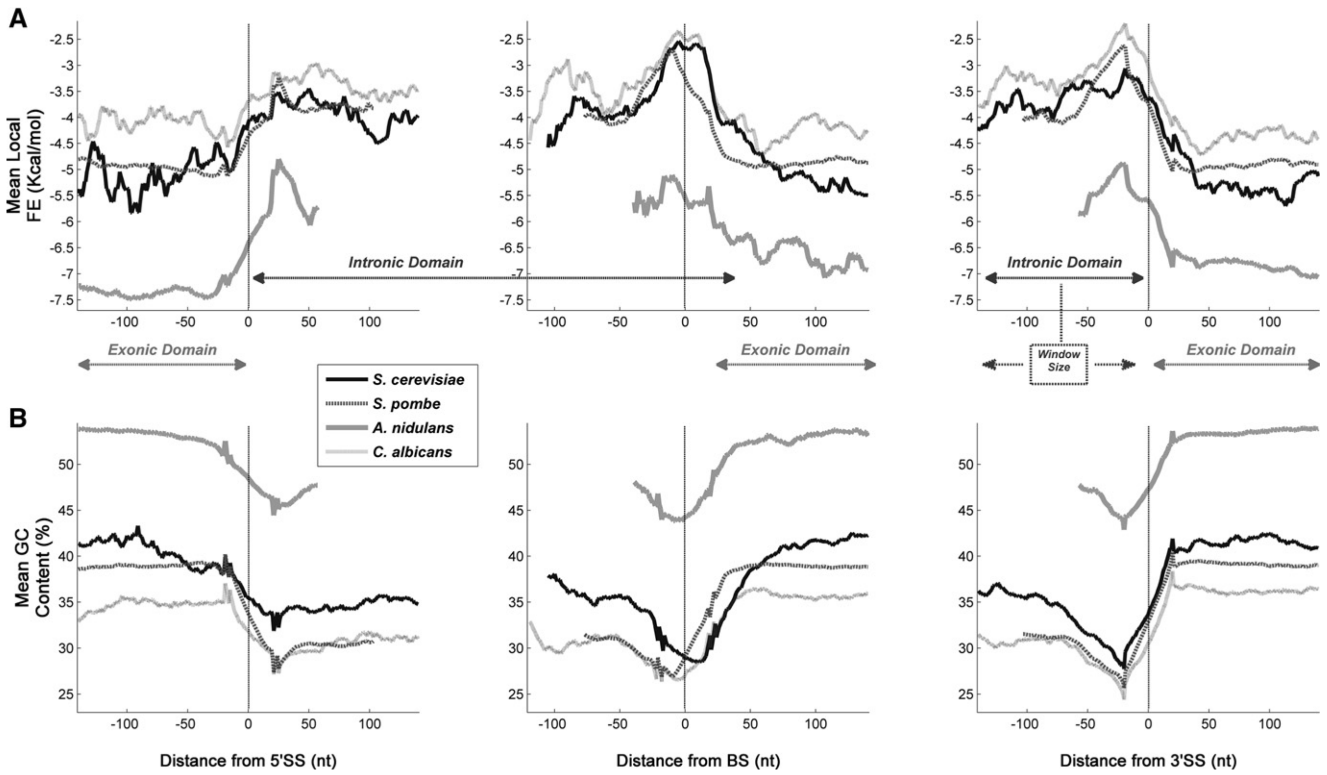

FIGURE 2. Intronome pre-mRNA profiles using a sliding window scheme. Analysis of the various fungi examined shows a clear ascent in local folding energy (LFE; corresponds to weaker folding) near intronic splice sites and branch site (BS), which cannot be fully explained by the GC content gradient, and a descent in LFE (corresponds to stronger folding) further away from the splice sites toward the exonic and intronic domains (further details in Materials and Methods). (*A*) Mean LFE profiles aligned around the donor site, BS, and acceptor site (*left*/*middle*/*right*, respectively) show weaker local folding around the splice sites and BS, and stronger local folding in the exonic/intronic domains further away. The distance from the BS to the beginning of the exonic domain (i.e., the 3′SS) varies between different introns; therefore, the exonic and intronic domains overlap (*middle*). (*B*) Mean GC content profiles aligned around the donor site, BS, and acceptor site (*left*/*middle*/*right*, respectively) show lower GC content surrounding the splice sites and BS. *A. nidulans* has significantly different (higher) GC content than the other fungi, resulting in lower LFE (i.e., stronger folding). UTRs and locations with <20% of the intronome were masked (see also Supplemental Table S6); the distance from the 5′SS/BS/3′SS is relative to the center of the sliding window; sliding window size is 40 nt.

changes in GC content levels, which remain fairly even. Correlation between GC and LFE on intronic intervals confirms that GC content only partially explains the LFE pattern observed (e.g., in *S. cerevisiae*: $r = -0.40$, $P = 0.016$ for the intronic donor domain, and $r = -0.11$, $P = 0.392$ for the intronic acceptor domain). These results are robust to the size of the sliding window used (analogous tendency can be seen in Supplemental Fig. S1 for window size of 50 nt; 30 nt not shown; additional correlation results can be found in Supplemental Table S1).

Evidence of a strong preference for weak secondary structure adjacent to intronic splice and branch sites

The reported results may suggest that the observed pattern of weak pre-mRNA folding near splice sites is preferred, probably in order to promote the exposure of the splice and branch sites consensus sequences, which allows for easier intronic recognition by the splicing machinery. In order to support our hypothesis and provide evidence of selection, we need to show that genomic features such as amino acid (AA) bias, codon bias, and intronic/exonic/UTR nucleotide distribution (i.e., GC content) cannot explain the observed pattern, and that the pattern is stronger than expected under random evolution of synonymous codons and intronic/UTR nucleotides that preserve the aforementioned features. In order to control for GC content in the various regions surrounding the splice sites, we devised several randomized models which preserve protein encoding information and major intronic and genomic features (see Materials and Methods and Fig. 1C–E); specifically, these models preserve the organism's intronic/exonic GC content and exonic codon-usage bias (CUB). Intron nucleotides were uniformly permuted, maintaining the 5′SS/BS/3′SS consensus sequences and GC content. Random codons were assigned to exons according to their genomic distribution, while maintaining the proteins they encode. Untranslated regions such as 5′ UTR and 3′ UTR were also randomized, maintaining their GC content properties. We analyzed a set of randomization models based on the aforementioned rules (see additional details in the Materials and Methods section) and compared them to the original intronome. Finally, and in order to provide evidence of selection, *Z*-score profiles were calculated: These profiles include deviation of the actual LFE and GC content from what is expected by the randomized model in standard deviation units (see Materials and Methods) such that a more extreme *Z*-score value suggests that there is stronger evidence of selection for weak/strong pre-mRNA folding.

LFE and *Z*-score profiles for several randomized models in *S. cerevisiae* can be seen in Figure 3 (sliding window size 50 nt; analogous tendency can be seen in Supplemental Fig. S2 for window size of 40 nt; the real and randomized LFE profiles look similar due to the fact that the randomized models are very strict and maintain many of the features of the actual genome, as previously described). As can be seen, there is clear evidence of weak folding preference around the intronic splice and branch sites for the real intronome in comparison to the randomized models applied, which cannot be explained by the global GC content gradient near intronic splice sites (e.g., a relatively low correlation for the intron randomization model in the donor aligned domain: $r = -0.24$, $P = 2.24 \times 10^{-4}$; see Materials and Methods and Supplemental Table S1; see Supplemental Figs. S3, S4 for detailed randomization profiles). Importantly, as can be seen in Figure 3C, the most extreme *Z*-scores (largest deviations from the randomized models) appear near the intronic boundaries and surrounding the BS, supporting the conjecture that in *S. cerevisiae* the reported signals are indeed related to splicing. In the rest of the organisms, the BS is located very close to the 3′SS (<25 nt; see Supplemental Fig. S5; Supplemental Table S6) resulting in coalescence of this signal with the signal of weak folding surrounding the 3′SS; specifically, in *S. pombe* the peak in LFE/*Z*-score is located 15–20-nt upstream of the 3′SS (see Fig. 2A; Supplemental Figs. S6, S7). A similar pattern was found in the other fungi examined (*A. nidulans* and *C. albicans*; LFE/GC/*Z*-score profiles can be seen in Supplemental Figs. S8, S9 and S10, S11, respectively). Finally, the observed pattern still occurs when introns, exons, or UTRs are randomized separately and specifically when introns were randomized in *S. cerevisiae* based on the exonic codon frequencies or separately for different intronic parts (see Supplemental Fig. S12), supporting the conjecture that each part of the transcript contributes to the signal and that there is co-evolution between the nucleotide composition in these parts to maintain the observed LFE pattern; this outcome is also relevant for results reported in the following sections.

Evidence of a strong preference for strong secondary structure further away from intronic splice and branch sites

In addition to the exon–intron boundary preference, and as can be seen in Figure 3, there is also clear evidence (FDR with $q = 0.03$) of strong folding selection in both of the exonic domains (further upstream and downstream from the 5′SS and BS/3′SS, respectively) in comparison to the randomized models applied; e.g., on the acceptor side of Figure 3C we can clearly see a preference for low LFE between nucleotides 44–60 downstream from the 3′SS. As before, the GC content gradient cannot fully explain the preference in all the fungi examined (see also the profiles for *S. pombe*, *A. nidulans*, and

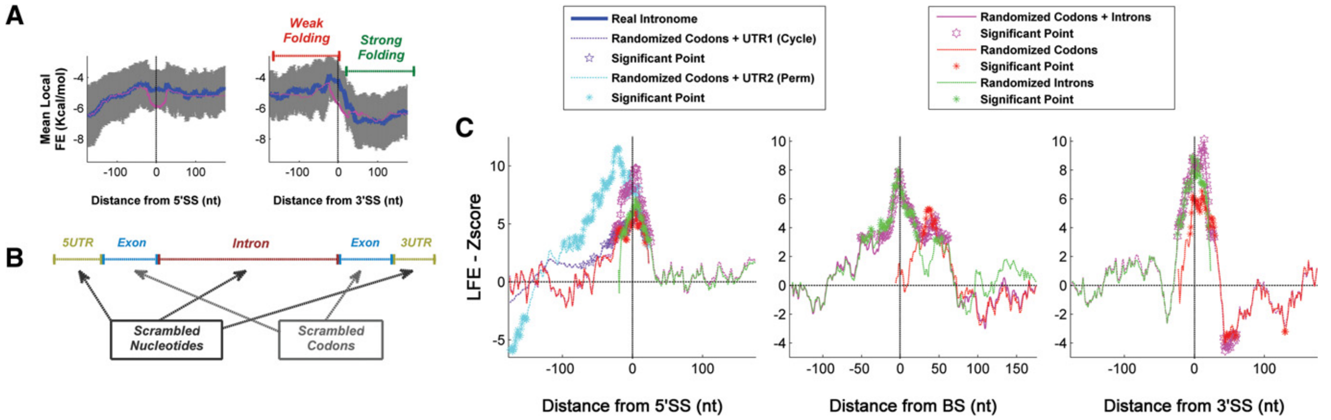

FIGURE 3. LFE and *Z*-score profiles for the complete *S. cerevisiae* intronome. The profiles correspond to the mean local pre-mRNA folding energy (LFE, *A*,*C*) of the intronome in comparison to the randomized ones, using several sequence randomization models of the real intronome that maintain consensus sequences and control for codon usage bias (CUB) and GC content in various regions (*B*; including codon, intron, and UTR randomizations; see also Fig. 1C–E and Materials and Methods). (*A*) LFE profiles and confidence intervals (gray; see Supplemental Methods) as well as randomized model of scrambled codon and scrambled intron; the randomized and real profiles are similar due to the fact that the randomized models maintain many of the features of the real intronome. The marked positions indicate significantly weak/strong folding controlled for false discovery rate (FDR with $q = 0.03$; see Materials and Methods; *C*); see detailed random profiles in Supplemental Figure S4. The most significant differences in the folding strength relative to the randomized models (i.e., the most extreme *Z*-score values) appear near the 5′SS/BS/3′SS, and the differences decrease for regions further away from the 5′SS/BS/3′SS. The distance from the 5′SS/BS/3′SS is relative to the center of the sliding window; sliding window size is 50 nt; positions surrounding the BS were masked in the 3′SS aligned profiles.

C. albicans in Supplemental Figs. S6–S11, respectively). This pattern is in line with previous work in *S. cerevisiae* showing a tendency for strong folding between the BS and 3′SS, possibly to decrease the distance between these regions (Plass et al. 2012). Together they imply that a complementary mechanism may exist, which promotes strong folding via protecting undesired splice site sequences from being recognized by the splicing machinery (see also PPT randomization in Supplemental Fig. S12).

A summary of the folding preference intervals for weak and strong folding in all the examined fungi is displayed in Figure 4 (FDR with $q = 0.03$; combined window sizes of 40 and 50 nt; please refer to Supplemental Table S2 for more details). In *S. cerevisiae* and *C. albicans* many introns are located very close to the beginning of the ORF (e.g., in *S. cerevisiae* >75% of them are found up to 100 nt from it) where it is known that there is a preference for weak folding related to translation initiation (Gu et al. 2010; Tuller et al. 2010). This most likely is the cause for the relatively long interval of weak folding in the exonic donor domain of those organisms.

Higher preference levels on LFE adjacent to intronic splice sites of ribosomal introns and on introns located in highly expressed genes in *S. cerevisiae*

In budding yeast, almost all introns were lost during evolution (Jeffares et al. 2006; Cohen et al. 2012). However, the ones that were preserved are known to be generally found in highly expressed genes; moreover, about one-third of its intronome originates from genes encoding ribosomal proteins (Ares et al. 1999). In order to determine whether there is a difference in LFE preference around the splice sites between ribosomal and nonribosomal introns, we divided intron-containing genes in *S. cerevisiae* into two subgroups: 93 introns originating from the very highly expressed ribosomal genes and 93 introns originating from nonribosomal genes (arbitrarily selected out of 187 in total to control for the effect of different group sizes; see Supplemental Methods). The LFE profiles shown in Figure 5A demonstrate that the reported signal appears both in ribosomal and in nonribosomal introns. However, ribosomal introns tend to exhibit a stronger signal as they have weaker folding in the vicinity of both the 5′SS and 3′SS locations in comparison to nonribosomal introns (up to 100 nt from each splice site; $\overline{\Delta G} = 0.83$[kcal/mol] and $\overline{\Delta G} = 1.11$[kcal/mol], $P = 6 \times 10^{-4}$ and $P = 2.7 \times 10^{-2}$, respectively; Wilcoxon rank-sum test, window size is 50 nt; see additional details in the Materials and Methods and Supplemental Methods sections). Further away toward the exonic donor and acceptor domains, the folding of the ribosomal introns is stronger than in the case of nonribosomal introns (100–200 nt from each splice site; $\overline{\Delta G} = -0.56$[kcal/mol] and $\overline{\Delta G} = -1.35$[kcal/mol], $P = 3 \times 10^{-2}$ and $P = 5.8 \times 10^{-8}$, respectively; Wilcoxon rank-sum test; additional data can be found in Supplemental Table S3). This result is consistent with previous studies showing a preference for strong mRNA structure in the beginning of the ORFs that is stronger in highly expressed genes, probably to promote translation efficiency and prevent aggregation of mRNA molecules (Tuller et al. 2011; Zur and Tuller 2012).

We further analyzed the model that combines randomizations of codons and introns for the aforementioned subgroups to identify additional evidence of selection; results indeed show the existence of a preference for weaker folding at the splice and branch sites in both subgroups, as can be seen in Figure 5B (for 40-nt profiles please refer to Supplemental Fig. S13). However, ribosomal introns show higher preference levels (i.e., more extreme *Z*-score values; $P < 4.7 \times 10^{-3}$, Wilcoxon rank-sum test) for weak folding at the splice and branch sites; e.g., on the acceptor site the LFE *Z*-score for ribosomal introns is 6.4 compared to 2.9

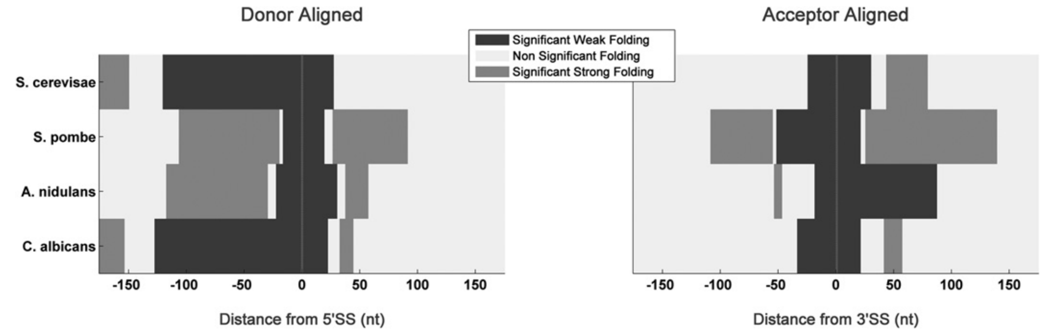

FIGURE 4. A summary of the folding preference intervals for all the examined organisms (*S. cerevisiae, S. pombe, A. nidulans,* and *C. albicans*). Positions with evidence of selection for weak pre-mRNA folding are marked in black while positions that exhibit evidence of selection for strong pre-mRNA folding are marked in silver. The preference for weak pre-mRNA folding is positioned around the donor and acceptor splice sites (*right* and *left*, respectively), and further away toward the exonic and intronic domains in most cases there is a preference for strong pre-mRNA folding (silver; *right* and *left*, respectively). The relatively long interval of weak folding in *S. cerevisiae* and *C. albicans (left)* is probably due to the fact that in those organisms many introns are located relatively close to the beginning of the ORF where it is known that there is a preference for weak folding which is related to translation initiation (Gu et al. 2010; Tuller et al. 2010); the distance from the 5′SS/3′SS is relative to the center of the sliding window; FDR with $q = 0.03$; combined window sizes of 40 and 50 nt; BS preference intervals were similar/identical (very close) to 3′SS intervals and therefore are not shown.

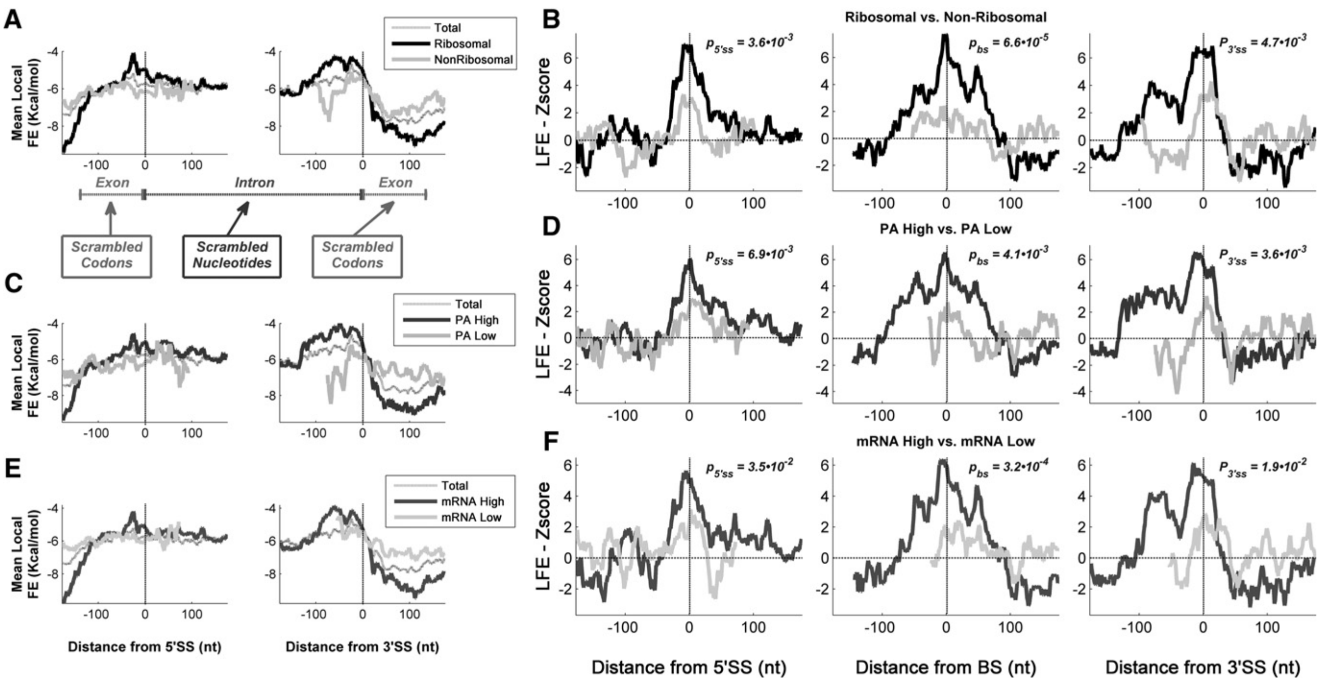

FIGURE 5. LFE and *Z*-score profiles for ribosomal/nonribosomal and highly/lowly expressed introns in *S. cerevisiae*. (*A*) LFE analysis of ribosomal versus nonribosomal intronome shows that ribosomal introns have a tendency for weaker folding in the intronic domains of the donor and acceptor splice sites, in comparison to nonribosomal introns (black and silver; *left* and *right*, respectively; intronome in dotted gray). (*B*) *Z*-score profiles correspond to the LFE preference of the real ribosomal and nonribosomal intronome (i.e., intronome with only ribosomal or nonribosomal genes; black and silver, respectively) in comparison to the scrambled ones using the combined codon and intron randomization model. (*C–F*) LFE profiles for introns originating from highly/lowly expressed genes (*C*,*E*; PA and mRNA, respectively) show similar results, as do the *Z*-score profiles (*D*,*F*; PA and mRNA, respectively). Profiles are aligned around the donor, branch, and acceptor sites (*left*/*middle*/*right*, respectively); the distance from the 5′SS/BS/3′SS is relative to the center of the sliding window; sliding window size is 50 nt; locations with <20% of the subgroup intronome were masked; positions surrounding the BS were masked in the 3′SS aligned profiles.

for nonribosomal introns (Fig. 5B, right). This result may be related to the higher protein abundance (PA) and mRNA levels of ribosomal genes in *S. cerevisiae* (PA average levels are 1804 and 127; mRNA average levels are 13.27 and 0.99; ribosomal and nonribosomal genes, respectively). Considering that since ribosomal genes are highly expressed (both in terms of mRNA levels and in terms of protein levels) and thus are better adapted to the transcription and translation process (e.g., highly expressed genes tend to have higher codon usage bias [Ikemura 1985; Kurland 1991]), we presume that they are also better adapted to other gene expression steps such as splicing. Hence, and in order to further consolidate our findings, we examined the LFE and *Z*-score profiles of two additional subgroups of the *S. cerevisiae* intronome based on their PA and mRNA levels (271 and 273 introns, respectively); this was accomplished via sorting introns by their expression levels, and analyzing their upper and lower quartiles (i.e., 68 introns in each subgroup). LFE profiles show similar results, i.e., weaker folding around splice sites and stronger folding in the exonic domains for highly expressed genes in comparison to lowly expressed genes (Fig. 5C,E; the stronger folding noticed in the highly expressed profiles is explained by this area being a part of the *S. cerevisiae* 5′ UTR/initiation site). The levels of preference for weak folding at the splice and branch sites were also shown to be stronger for highly expressed genes in comparison to lowly expressed genes, both based on mRNA and PA differentiations (Fig. 5D, F; $P < 3.5 \times 10^{-2}$, Wilcoxon rank-sum test; more details in Materials and Methods and in Supplemental Fig. S14; energy characteristics and additional statistical data can be found in Supplemental Table S3).

The connection between pre-mRNA secondary structure preference and splicing efficiency measured via a synthetic intron library system

According to the aforementioned results, introns with strong pre-mRNA folding strength surrounding their splice sites should be spliced less efficiently than introns with weak folding strength. Thus, we expect that exonic/intronic regions which have a higher folding effect on splicing efficiency (SE) and highly spliced introns will also be under stronger pre-mRNA folding preference. One way to validate these two hypotheses is by comparing the inferred pre-mRNA folding signal (and folding selection signal) with SE estimations. We did this using a previously reported synthetic library (Yofe et al. 2014). Briefly, *S. cerevisiae* is transformed with a library of DNA transformation cassettes, each containing a different native yeast intron. The cassettes are assembled using the Y-operation by which introns are embedded in yellow fluorescence protein (YFP) fragment and concatenated to a common selection marker in high throughput. Consequently, the sole difference between all strains is the native *S. cerevisiae* intron separating the YFP gene. The average expression level of each intron-containing strain (termed YiFP strain) is compared to that of an intron-less reference strain to provide a measure of its relative expression level that is associated with its intronic SE, i.e., YiFP expression level. YiFP strains whose expression levels did not pass the detection limit were considered as nonspliced. Summary of the system can be seen in Figure 6A; for more details see Materials and Methods and Yofe et al. (2014).

First, we extracted the intronic SE library related measurements of 215 introns. Next, we looked at the LFE/*Z*-score profiles of two subgroups of endogenous introns in correspondence to their SE in the synthetic system: 93 genes with the highest SE and 93 genes with the lowest SE; each group accounts for ∼33% of the *S. cerevisiae* intronome. As can be seen in Figure 6B,C (sliding window size 50 nt; for 40-nt profiles please refer to Supplemental Fig. S15), there is evidence of selection and a higher preference for weaker folding at the splice sites for highly spliced genes in comparison to lowly spliced ones ($P < 2.1 \times 10^{-2}$, Wilcoxon rank-sum test). In order to assess the combined contribution of LFE to the YiFP expression levels, we then constructed a linear regression function that optimizes combination of folding features that accurately account for the measured YiFP expression levels. The prediction function was built by iteratively adding LFE features that yield the highest correlation to expression, considering only features in the vicinity of each splice site separately (up to 50-nt upstream/downstream from each splice site). Our model prediction near the donor site yielded an adjusted *R* correlation (i.e., a correlation that considers the number of features and thus overfitting) of 0.28 with the YiFP measurements using a combination of five features, and up to 0.34 using a combination of eight features; near the acceptor site the adjusted *R* was 0.34 using five features, and up to 0.36 using eight features (Fig. 6D; $P < 2.63 \times 10^{-6}$). Additionally, we performed cross validations using a 50%/50% train/test scheme; results gave $r = 0.36$ and $r = 0.24$ for the donor and acceptor site, respectively (Fig. 6E; $P < 1.24 \times 10^{-2}$; see Materials and Methods and Supplemental Table S8). Finally, we correlated SE with the LFE profiles of the synthetic introns (which was named Synthetic LFE) and compared the received correlation profiles with the randomization models' *Z*-score profiles of the endogenous introns. As can be seen in Figure 6F,G, the randomized models' *Z*-score and the SE correlation show a clear association (e.g., $r = 0.77$, $P = 2.98 \times 10^{-21}$ for LFE *Z*-score based on the combined codon and intron randomized model in the donor domain; $r = 0.67$, $P = 4.93 \times 10^{-14}$ based on a randomized model that maintains the GC content of the intron in the acceptor domain; and $r = 0.46$, $P = 2.63 \times 10^{-6}$ for partial correlation between SE and LFE *Z*-score of the intron randomized model when controlling for the GC content in the donor domain); these results support our conjecture that regions with a higher folding effect on SE will be under stronger pre-mRNA folding preference. Further validations when looking at a subgroup of only 167 spliced YiFP introns or when excluding ribosomal introns showed similar results (but less significant, as can be seen in Supplemental Fig. S16; see also Materials and Methods and Supplemental Methods; complete correlation and partial correlation results can be found in Supplemental Table S4). These results show that there is a strong relation between the strength/significance level of the LFE preference and the effect of LFE on splicing in different positions near the intronic splice sites, which is consistent with the aforementioned results.

Higher preference levels on LFE adjacent to intronic splice sites of conserved introns and on introns located in highly expressed genes in *S. pombe*

In fission yeast, 1500–2000 introns were lost during evolution while ∼1800 introns demonstrated a high degree of conservation (Roy and Gilbert 2005; Carmel et al. 2007; Rhind et al. 2011; Cohen et al. 2012). In order to compare conserved and nonconserved genes in *S. pombe*, we looked at 1718 conserved and 3029 nonconserved introns found in four diverse fission yeasts reported by Zhu and Niu (2013) (*S. cryophilus*, *S. octosporus*, *S. pombe*, and *S. japonicus*); we selected 1718 conserved and 1718 nonconserved introns and analyzed their folding preference by generating *Z*-score profiles (see explanation in the Supplemental Methods). The LFE profiles showed minor differences between conserved and nonconserved introns surrounding the splice sites (Fig. 7A; e.g., $\overline{\Delta G} = -0.025[\text{kcal/mol}]$ on the acceptor side; sliding window size of 50 nt). However, conserved introns demonstrated higher preference for weak folding (Fig. 7B; $P < 4.8 \times 10^{-2}$ at the splice sites, Wilcoxon rank-sum test). Further downstream in the intronic/exonic domains, the LFE and preference levels for strong folding were higher for conserved introns in comparison to nonconserved ones (e.g., $\overline{\Delta G} = -0.39[\text{kcal/mol}]$ on the acceptor side; $P = 9.8 \times 10^{-3}$ in the BS/3′SS, Wilcoxon rank-sum test); these results differ from those of *S. cerevisiae* since most of the introns in *S. pombe* are not found near the beginning of the ORF (Wood et al. 2002). In addition, we analyzed 2000 introns originating in highly and lowly expressed genes, based on their PA. As can be seen in Figure 7C,D, highly expressed genes likewise have a preference for weaker folding surrounding the splice sites and for stronger folding downstream at their intronic/exonic domains in comparison to lowly expressed genes (e.g., $P = 1.8 \times 10^{-2}$ in the 5′SS; Wilcoxon rank-sum test). These results support our previous results and hypotheses (additional data can be found in Supplemental Table S3; for 40-nt profiles please refer to Supplemental Fig. S17).

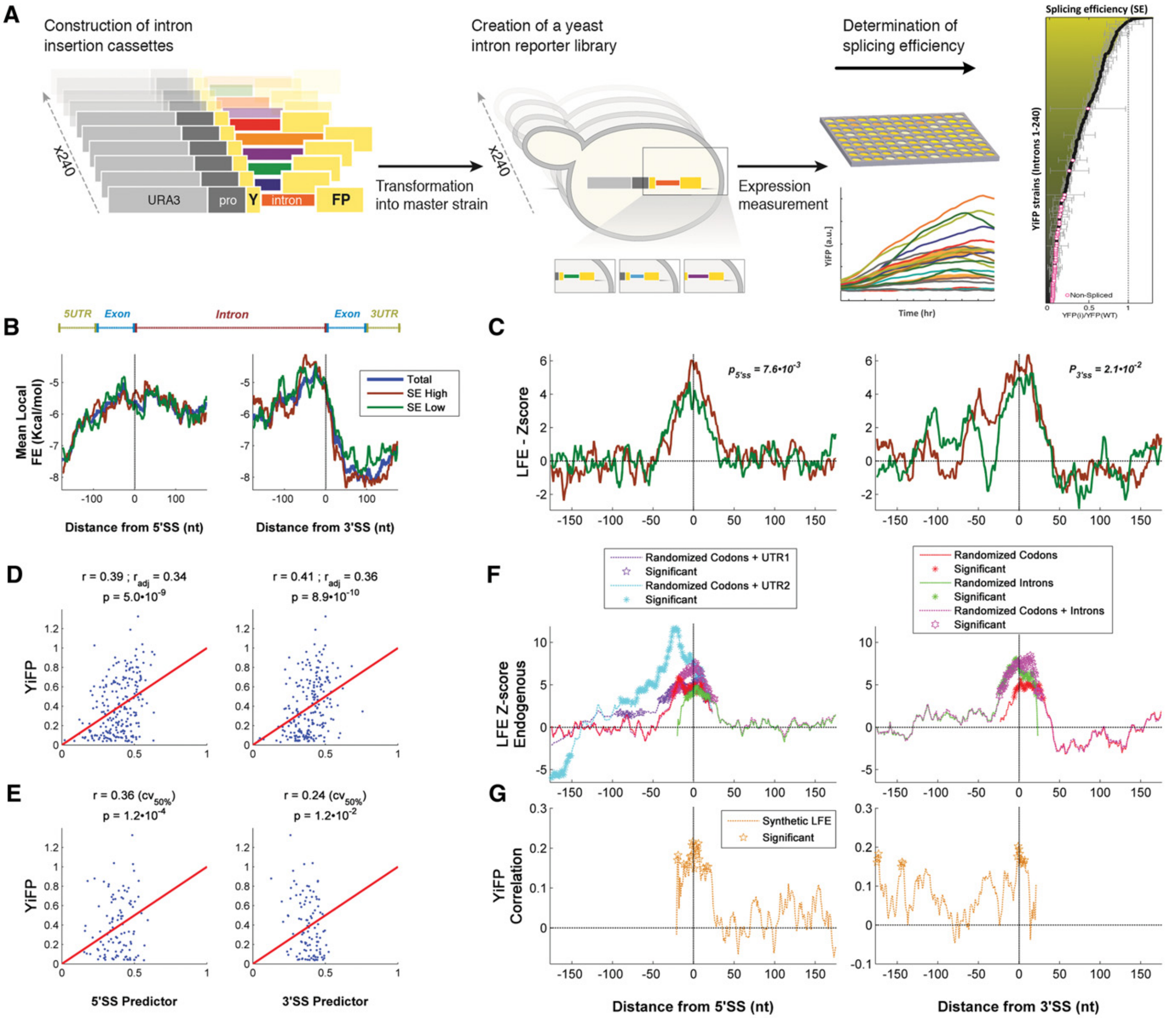

FIGURE 6. Analysis of YiFP expression levels measured via a synthetic system demonstrates the connection between folding strength and preference to splicing efficiency. (*A*) Overview of the reporter approach for studying splicing mediated gene expression regulation: Intron insertion cassettes were constructed in vitro, each comprised of a selection marker (URA3), a constitutive promoter, the first 195 nt of the YFP gene, and one of 240 native *S. cerevisiae* introns followed by an additional 600 nt of the YFP gene. Each insertion cassette was transformed into the genome of a master strain which contained a promoter-less YFP gene, thus creating an in vivo intron-reporter yeast library (YiFP); each strain's average expression levels were compared with that of an intron-less reference strain, to get an assessment of splicing efficiency (SE); YiFP strains whose YFP levels did not pass the detection limit were considered as nonspliced (marked with circles; error bars represent STD from four independent experiments; see Materials and Methods and Yofe et al. 2014). (*B*) LFE analysis of endogenous introns corresponding to highly/lowly expressed YiFP genes (i.e., SE) shows that both groups have a similar tendency for weaker folding surrounding their splice sites (brown and green, respectively; intronome in blue). (*C*) *Z*-score profiles correspond to the LFE preference of the real highly/lowly spliced intronome in comparison to the scrambled ones (using the combined codon and intron randomization model) which show higher preference levels for highly spliced introns at the splice sites. (*D*) LFE-based prediction of the YiFP expression levels yielded correlation of 0.39 surrounding the 5′SS using a combination of eight features; surrounding the 3′SS the correlation was 0.41 when using a combination of eight features ($P < 4.98 \times 10^{-9}$; see also Supplemental Table S8); adjusted *R* correlation values were 0.34 and 0.36 for the donor and acceptor sites, respectively. (*E*) LFE-based predictions of the YiFP expression levels using 50%/50% train/test cross validation yielded correlations of 0.36 and 0.24, respectively ($P < 1.24 \times 10^{-2}$). (*F*,*G*) *Z*-score profiles correspond to the LFE of the endogenous introns (*F*) in comparison to the correlation profiles between all introns synthetic LFE to YiFP expression levels (*G*) which show high resemblance. The randomization models include codon, intron, and UTR randomizations: the randomized codon model includes scrambled exonic sequences; the randomized intron model includes scrambled intronic sequences; the randomized UTR1/UTR2 (cycle/permutation) models include scrambled untranslated sequences. Profiles are aligned around the donor and acceptor domains (*left* and *right*, respectively); the distance from the 5′SS/3′SS is relative to the center of the sliding window; sliding window size is 50 nt.

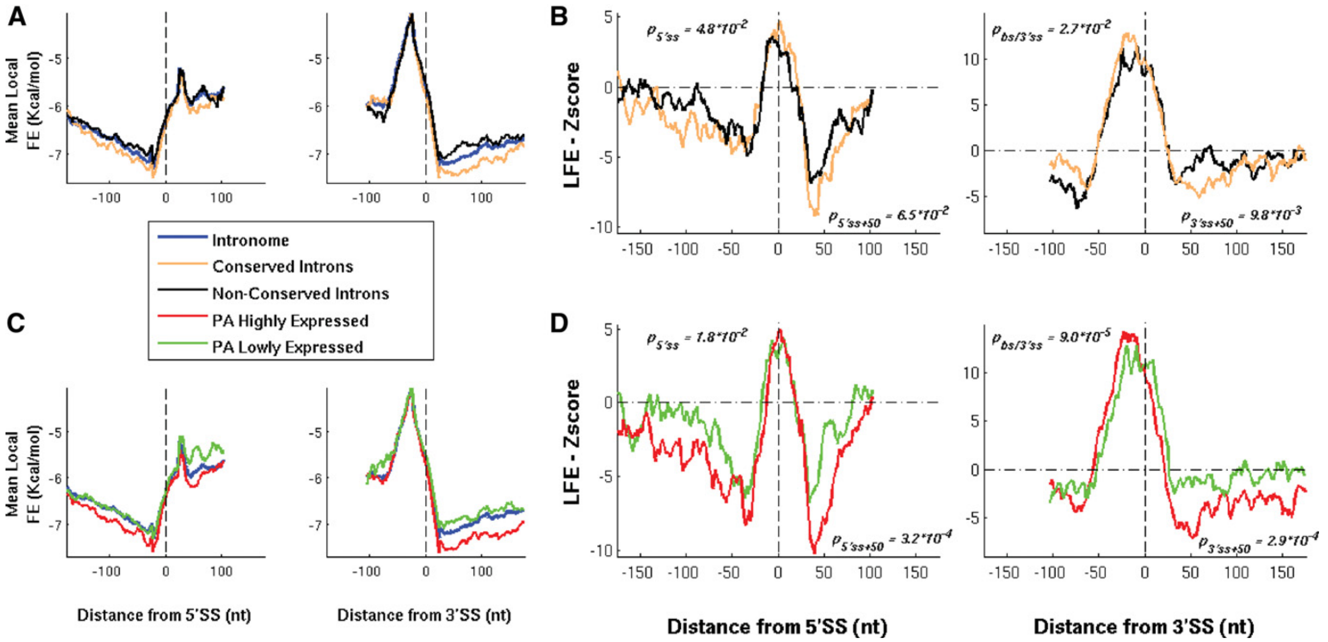

FIGURE 7. *Z*-score profiles for conserved/nonconserved and highly/lowly expressed introns in *S. pombe*. (*A*) LFE analysis of conserved versus nonconserved introns shows that both groups have a similar tendency for weaker folding surrounding their splice sites (apricot versus black; intronome in blue). However, further away from both splice sites and toward the exonic domains, conserved introns exhibit a tendency for stronger folding. (*B*) *Z*-score profiles correspond to the LFE of the real conserved and nonconserved intronome in comparison to the randomized ones using the combined codon and intron randomization model, and show higher preference levels for the conserved introns at the splice sites and 50-nt downstream from both splice sites. (*C*) LFE analysis of highly expressed versus lowly expressed introns shows that both groups have a similar tendency for weaker folding surrounding their splice sites (red versus green; intronome in blue). However, further away from both splice sites and toward the exonic domains, highly expressed introns exhibit a tendency for stronger folding in comparison to lowly expressed introns (e.g., $\overline{\Delta G} = -0.51$[kcal/mol] on the acceptor side). (*D*) *Z*-score profiles correspond to the LFE of the real highly and lowly expressed intronome (red and green, respectively; PA based) in comparison to the randomized ones using the combined codon and intron randomization model, show higher preference levels for the highly expressed introns in and 50-nt downstream from both splice sites (see also Supplemental Table S3). Profiles are aligned around the donor and acceptor domains (*left* and *right*, respectively); the distance from the 5′SS/3′SS is relative to the center of the sliding window; sliding window size is 50 nt; locations with <20% of the intronome were masked; acceptor site *P*-values were calculated in the combined location of the BS/3′SS.

Preference on LFE adjacent to intronic splice sites in *S. pombe* is not associated with a specific function

In order to show that the preference for weak/strong folding, which was previously shown, appears in many cellular functions and is not function-specific or found in a small set of introns related to very specific function(s), we used gene ontology (GO) and performed the folding preference analyses mentioned above for 90 different GO terms separately (see Supplemental Methods; see also list of the included and excluded terms in Supplemental Table S7). As can be seen in Figure 8 and Supplemental Figure S18, there is a preference for weak folding adjacent to 5′ sites further on in the exonic/intronic domains in 49%/46%/67% of the gene functions for biological process, molecular function, and cellular component, respectively. Similarly, there is a preference for weak folding adjacent to 3′ sites or further in the exonic/intronic domains in 85%/79%/81% of the gene functions for biological process, molecular function, and cellular component, respectively. These results suggest that the signal reported in previous sections appears in many gene groups/functions and strengths, and is not related to very specific function(s); we do not claim, however, that the strength of the signal is identical in all gene groups.

DISCUSSION

In this study, we provide lines of evidence that support the conjecture that evolution shapes fungal pre-mRNA folding near intronic splice sites at the genomic level to improve splicing efficiency (SE). Specifically, we report the following major results: (A) The intron–exon boundaries exhibit weaker local pre-mRNA folding; (B) there is a *preference* for low folding strength in these regions, presumably to improve SE; (C) there is a *preference* for strong pre-mRNA folding 30–140 nt further away from the intron–exon boundaries; (D) the reported signal of pre-mRNA folding preference is correlative with the effect of pre-mRNA folding on SE measurements based on a synthetic system in yeast; (E) the reported signals are stronger for highly expressed genes and genes found in conserved introns; (F) the reported signals are not associated with a specific function; (G) all the different parts of the transcript (exons, introns, and UTRs)

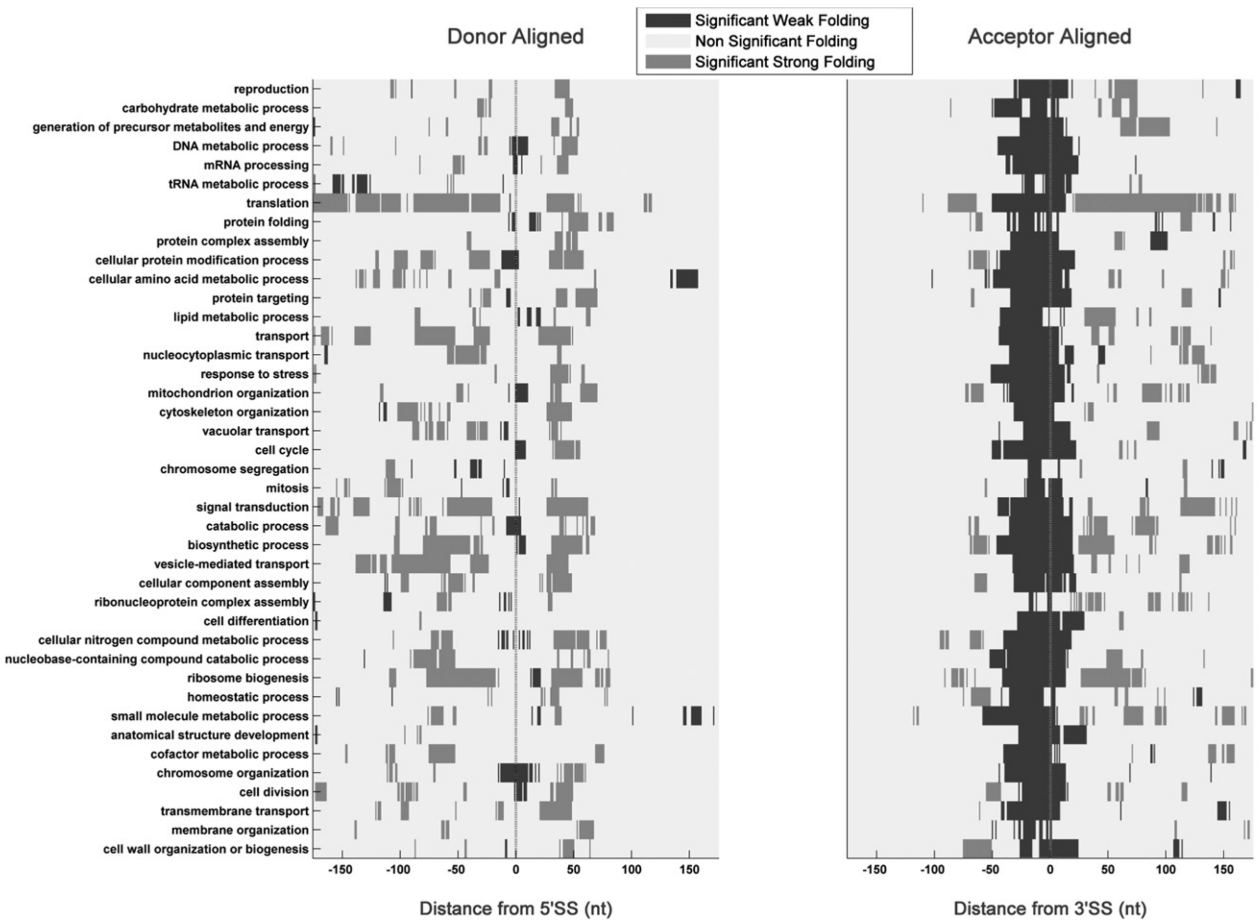

FIGURE 8. GO terms analysis in *S. pombe* reveals that pre-mRNA folding preference is not associated with a specific function. GO summary of folding preference intervals for biological processes demonstrates that in various processes the preference for weak folding is positioned around the donor and acceptor splice sites (dim gray; *right* and *left*, respectively), and further away toward the exonic/intronic domains for strong folding (silver); thus, the reported signal seems to appear in many cellular functions and not only in a small set of introns related to very specific function(s). Combined window sizes of 40 and 50 nt; the distance from the 5′SS/3′SS is relative to the center of the sliding window; significant positions are marked ($P < 0.01$); locations with <20% of the intronome were masked; the reported signal is usually weaker than the genome signal due to the smaller number of genes in functional groups (relative to the total number of genes).

undergo adaptation to maintain the observed local pre-mRNA folding signal.

While previous work has already studied specific structure regions and their conservation (Warf and Berglund 2010; McManus and Graveley 2011) and suggested that the GC content of introns and exons is an important feature related to alternative splice site recognition through RNA structure (Zhang et al. 2011; Amit et al. 2012), the evolution of local folding strength surrounding the intronic boundaries has not been studied before at a wide genomic level and in a single nucleotide resolution. It is also important to emphasize that in our analyses the null models maintained and controlled for the GC content of introns, exons, and UTRs; thus, the reported results cannot be explained only by GC content. As demonstrated, the ascent in LFE coincides with a decrease in GC content, which is consistent with weaker folding (Shepard and Hertel 2008). However, variations and GC content preference can only partially explain the LFE pattern observed. Therefore, we assert that the LFE preference is not only due to changes in GC content near the intronic boundaries.

Previous studies have demonstrated that transcript folding affects the efficiency of translation initiation and elongation, and may prevent aggregation of mRNA molecules (de Smit and van Duin 1990; Gu et al. 2010; Tuller et al. 2010, 2011; Zur and Tuller 2012). Here we show that preference for pre-mRNA folding may also be related to pre-translational processes such as splicing. In addition, the fact that the strength of the reported signals is highest near the splice sites and is correlated with splicing measurements suggests that they are indeed strongly related to splicing and not to other stages of gene expression, such as translation initiation and elongation (de Smit and van Duin 1990; Gu et al. 2010; Tuller et al. 2010, 2011; Zur and Tuller 2012).

The studied organisms differ in many of their genomic/intronic aspects: GC content, number and length of introns (as well as the number of introns per gene), distance of the introns from the UTRs, and function and expression levels of intron-containing genes; e.g., in *C. albicans* introns are not randomly distributed and are overrepresented in genes involved in specific cellular processes such as splicing, translation, and mitochondrial respiration (Mitrovich et al. 2007). In addition, it is easy to see that most of the introns analyzed here do not have homologs in more than one organism. Nonetheless, the reported findings lead to similar conclusions in all analyzed organisms and could be the result of convergent evolution preference for weak/strong folding surrounding the splice and branch sites, thus emphasizing their importance and universality.

In conclusion, we believe that these new discoveries should have an important contribution to various biomedical disciplines including molecular evolution, functional genomics, and biotechnology. Specifically, they are contributory steps toward a broader understanding of intron–exon boundary evolution and can be used for developing future models of intronic evolution that will consider the effect of mutations on SE. In addition, they clearly elucidate how SE is encoded in gene sequences. Furthermore, they can help us understand how silent mutations are related to human diseases via their possible effect on pre-mRNA folding near intronic splice sites (Sauna and Kimchi-Sarfaty 2011). Finally, the results suggest new methods for engineering gene expression of synthetic genes via the manipulation of pre-mRNA folding near intronic splice sites (Yofe et al. 2014).

MATERIALS AND METHODS

The analyzed organisms

The four fungi analyzed here (*S. cerevisiae, S. pombe, A. nidulans,* and *C. albicans*) were chosen based on the following considerations: First, *S. cerevisiae* and *S. pombe* are well-studied organisms with well-established databases that are known to have diverged 350–1000 million years ago (Berbee and Taylor 2001). *A. nidulans* and *C. albicans* are two additional fungi known to have diverged from *S. pombe* ∼650 million years ago (Berbee and Taylor 2010) and from *S. cerevisiae* ∼235 million years ago (Taylor and Berbee 2006). The genomes of these organisms are also fully sequenced and their introns are well annotated. In addition, *C. albicans* is a dimorphic fungus which can be a significant pathogen in humans.

Intronic sequence information

S. cerevisiae ORFs and intron-containing gene sequences (strain 288c) were taken from the *Saccharomyces* Genome Database (SGD) (Cherry et al. 1998); BS location information was obtained from the Ares laboratory database (Grate and Ares 2002). *S. pombe* genome information (Assembly 16) was taken from the PomBase database (Wood et al. 2012); BS locations were calculated based on the position specific scoring matrix (PSSM) information extracted from the Sanger Institute, which is based on the original full genome sequencing from (Wood et al. 2002). *A. nidulans* (FGSC A4) and *C. albicans* (SC5314 Assembly 21) genome information was taken from the *Aspergillus* Genome Database (AspGD) (Arnaud et al. 2012) and *Candida* Genome Database (CGD) (Inglis et al. 2012), respectively; BS locations were calculated based on the fungal BS consensus sequence (*CURAY*). We used only introns taken from coding sequence genes and excluded 5′ UTR and 3′ UTR introns. Introns associated with alternatively spliced genes were also excluded. The full intron exclusion list can be found in Supplemental Table S5. Additional GC content and intron information can be found in Supplemental Table S6; exon–intron GC content was calculated using intron sequences and their flanking exon sequences (150 nt upstream and downstream).

Protein abundance and mRNA levels

Protein abundance (PA) information for *S. cerevisiae* and *S. pombe* was taken from the PaxDb, which integrates information from various resources (Wang et al. 2012). Levels of mRNA for *S. cerevisiae* are based on RNA-seq and DNA chip and were obtained by integration of three data sets (Wang et al. 2002; Nagalakshmi et al. 2008; Ingolia et al. 2009) as follows: First, we normalized each data set by its average mRNA levels; next, for each gene we averaged all its normalized measurements. PA levels are measured in parts per millions. The mRNA is based on RNA-seq and DNA chip and thus is proportional to the mRNA levels (number of molecules in the cell). Since our analysis is based on Spearman correlation, ranking of genes according to their expression levels is enough to provide the required results (i.e., the actual levels will not change the results).

Construction of secondary structure and GC content profiles

The local pre-mRNA folding and GC content profiles were computed as follows: We used three sliding window sizes (30, 40, and 50 nt, corresponding to the approximated splicing factors and spliceosome footprints size in fungi) focusing on the intronic region and its flanking exonic sequences; for every intron we computed local folding energy (LFE) and GC content for all sliding windows, with a single nucleotide shift. Let LFE_WL(i) denote the folding energy of a window size of WL nucleotides, centered on the i-th nucleotide of the gene's pre-mRNA transcript. The intronic profile of gene j was defined as the vector of the LFE values assigned to n sliding windows of size WL, i.e., $\mathrm{LFE\_WL}_{\mathrm{Gene}_j} = (\mathrm{LFE\_WL}^j(1),\ \mathrm{LFE\_WL}^j(2), \ldots, \mathrm{LFE\_WL}^j(n))$. For each organism, all the intron-containing genes were aligned once according to their donor site (5′SS location), and once according to their acceptor site (3′SS location). Let $i_{5'\mathrm{ss}}$ and $i_{3'\mathrm{ss}}$ denote the positions of the 5′SS start and 3′SS end of the introns (for each of the analyzed introns these indices correspond to its own 5′SS and 3′SS, respectively). The profiles of mean LFE were calculated as

$$\overline{\mathrm{LFE\_WL}_{5'\mathrm{ss}}} = \left(\overline{\mathrm{LFE}\left(i_{5'\mathrm{ss}} - \left(n - \frac{1}{2}\right)\cdot \mathrm{WL} + 1\right)}, \ldots, \overline{\mathrm{LFE}(i_{5'\mathrm{ss}})}, \ldots, \overline{\mathrm{LFE}\left(i_{5'\mathrm{ss}} + \left(n - \frac{1}{2}\right)\cdot \mathrm{WL}\right)}\right)$$

$$\overline{\mathrm{LFE\_WL}_{3'\mathrm{ss}}} = \left(\overline{\mathrm{LFE}\left(i_{3'\mathrm{ss}} - \left(n - \frac{1}{2}\right)\cdot \mathrm{WL} + 1\right)}, \ldots, \overline{\mathrm{LFE}(i_{3'\mathrm{ss}})}, \ldots, \overline{\mathrm{LFE}\left(i_{3'\mathrm{ss}} + \left(n - \frac{1}{2}\right)\cdot \mathrm{WL}\right)}\right),$$

where $\overline{\mathrm{LFE\_WL}(i)}$ is the average LFE in position i when considering all genes with introns long enough to have a value in this position, and $[n-(1/2]\cdot\mathrm{WL}$ is the number of nucleotides in the complete analyzed exonic and intronic regions (we used $n = 4$). Additionally, downstream 3′SS/exons in 5′SS profiles and upstream 5′SS/exons in 3′SS profiles were masked. For calculation simplicity, genes containing $n > 1$ introns were duplicated n times. Thus, for each duplicate, a different intron was retained while the other introns were extracted. In the case of the profiles generated in Figure 2, for each position we only considered intronic/exonic sequences and masked UTR sequences (this is relevant only in the case of short introns/exons). The profiles for GC content were calculated in the same manner. A scheme of the procedure with some possible gene examples is presented in Supplemental Figure S19.

Secondary structure randomization models

The randomized models were designed to conserve encoded protein information, in addition to the intronic and UTR properties. Specifically, we maintained synonymous codon frequencies, canonical splicing signals, and GC content. To this end, the distribution of synonymous codons was calculated for each organism, and the coding sequences were then scrambled while maintaining these codon frequencies, i.e., for any individual AA the probability of a certain codon was determined by its distribution in the genomic coding sequences. Intronic nucleotides were uniformly permutated, maintaining consensus sequences (5′SS/BS/3′SS; PPT was not maintained since it is not essential in yeast [Ruby and Abelson 1991; Birney et al. 1992]) and GC content. 5′ UTR and 3′ UTR sequences (upstream and downstream flanking 200 nt, respectively) were also randomized while maintaining their GC content using the following methods: (1) cyclic shift and (2) uniform permutation. Thus, the randomized profiles shown in Figure 3A and Supplemental Figs. S2–S4, S7, S9, S11 are similar to the real profiles due to the fact that the randomized models are very strict and maintain many of the features of the actual genome, such as the protein encoded in the exons, CUB, GC content of introns/exons, and consensus sequences in the introns. We used the following randomization schemes to generate the random sets: (a) Codon only, (b) Intron only, (c) UTR cyclic shift, (d) UTR permutation. In addition, a combination of the basic schemes was applied: $(a) + (b)$, $(a) + (c)$, and $(a) + (d)$. Details of the various randomization models are illustrated in Figure 1C–E.

The level of significance, i.e., the P-value that controls for multiple hypotheses testing, was determined in the following manner: For each genome and for each scheme, 100 random intron sets were generated and an empirical P-value was calculated; we also evaluated the distribution characteristics for each position in the average profile using the Kolmogorov–Smirnov test; positions with LFE distributions that were not significantly different from a normal distribution were treated as belonging to a normal distribution, and based on the parameters of the normal distribution an analytical significant level was computed; positions that did not pass remained with their empirical significance level. In case the empirical P-value was <0.01, a new one was generated based on additional permutations until we were able to estimate a $P\text{-value} > 0$ (200 in this case); further, we controlled for false discovery rate (FDR) with $q = 0.03$ as a cutoff (Benjamini and Yekutieli 2001).

Secondary structure predictions

RNA secondary structure and LFE predictions were calculated using *rnafold* (Vienna) (Mathews et al. 1999; Wuchty et al. 1999), which predicts the secondary structure associated with the minimum free energy for the sequence using the thermodynamic nearest-neighbor approach.

Computing *Z*-scores based on the randomized models

Standard Z-score values were calculated according to the following equation:

$$Z_{\text{score}} = \frac{\mu_{\text{real}} - \mu_{\text{rand}}}{\sigma_{\text{rand}}},$$

where μ_{real} is the mean of the intronome, μ_{rand} is the randomized model mean, and σ_{rand} is the randomized model standard deviation. For each random profile a vector of Z-score values was generated, i.e., a Z-score was generated for each randomized position; Z-score values for nonrandomized positions were masked and are not shown. Detailed example of the Z-score calculations (as well as the mean difference and standard deviation) can be seen in Supplemental Figures S14, S20.

Partial correlation analysis

Partial correlation analysis is aimed at finding the correlation between two variables after removing the effects of other variables; the partial correlation coefficient $\rho_{xy,z}$ between X and Y given a set of n controlling variables $Z = \{Z1, Z2 \dots, Zn\}$ is the correlation between the residuals R_X and R_Y resulting from the linear regression of X with Z and of Y with Z, respectively; the approach can be generalized to deal with Spearman correlation (Kendall and Stuart 1973).

Synthetic YiFP reporter library building and analysis

The synthetic YiFP reporter library was downloaded from Yofe et al. (2014). To create the synthetic intron-reporter library, *S. cerevisiae* was transformed with a library of DNA transformation cassettes, each containing a different native yeast intron. Intron insertion cassettes were constructed in vitro, using the Y-operation; each comprised of a selection marker (URA3), a constitutive promoter, the first 195 nt of a yellow fluorescent protein (YFP) gene, and one of 240 native *S. cerevisiae* introns followed by an additional 600 nt of the YFP gene. Each insertion cassette was transformed into the genome of a master strain which contained a promoter-less YFP gene, thus creating an in vivo intron-reporter yeast library (YiFP).

In this manner 240 strains were created, termed YiFP strains, where the sole difference between all strains is the native *S. cerevisiae* intron intervening the YFP gene. The contribution of introns to the regulation of gene expression was assessed by dynamic measurements of YFP expression. Following normalization, each strain's average expression level was compared with that of an intron-less reference strain to get an assessment of its SE; YiFP strains that had a signal to noise ratio (SNR) of ≥5 were classified as spliced (178 introns), while the others (SNR < 5; 62 introns) were classified as nonspliced (see Fig. 6A; marked with circles).

Linear regression prediction of YiFP expression

We constructed predictors for each sliding window scheme. All the LFE features were put into a linear regression model to assemble an expression prediction for the donor and acceptor splice sites separately. For each prediction a feature assembly list was calculated and the accumulation of features was done using a greedy algorithm; in each feature assembly iteration k, Spearman correlation was calculated. In order to control for overfitting we used two schemes: (1) An adjusted correlation (which controls for the number of features) value was calculated according to the following formula:

$$R^2_{\mathrm{adj}}(k) = R^2 - (1 - R^2)\frac{k}{n - (k + 1)},$$

where n is the number of measurement features and R is the Spearman correlation in the k-th iteration; the predictor stops when $R^2_{\mathrm{adj}}(k + 1) < R^2_{\mathrm{adj}}(k)$. (2) Train/test cross-validation was performed using 108 introns as a training set and 107 as a test set (randomly chosen from a total of 215 introns). The results were matched up to the value of R^2 and are presented in Supplemental Table S8.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. We thank Ms. Hadas Zur and Ms. Talia Shrem for their helpful comments.

Received February 4, 2015; accepted June 25, 2015.

REFERENCES

Alexander K, Anatoliy I, Alexander B. 2011. Statistical analysis of exon lengths in various eukaryotes. *Open Access Bioinformatics* **2011:** 1–15.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, et al. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1:** 543–556.

Ares M Jr, Grate L, Pauling MH. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5:** 1138–1139.

Arnaud MB, Cerqueira GC, Inglis DO, Skrzypek MS, Binkley J, Chibucos MC, Crabtree J, Howarth C, Orvis J, Shah P, et al. 2012. The *Aspergillus* Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Res* **40:** D653–D659.

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5:** 773–782.

Barbazuk WB, Fu Y, McGinnis KM. 2008. Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res* **18:** 1381–1392.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* **29:** 1165–1188.

Berbee ML, Taylor JW. 2001. Fungal molecular evolution: gene trees and geologic time. In *Systematics and evolution*, pp. 229–245. Springer, New York.

Berbee ML, Taylor JW. 2010. Dating the molecular clock in fungi—how close are we? *Fungal Biol Rev* **24:** 1–16.

Bergkessel M, Whitworth GB, Guthrie C. 2011. Diverse environmental stresses elicit distinct responses at the level of pre-mRNA processing in yeast. *RNA* **17:** 1461–1478.

Birney E, Kumar S, Krainer AR. 1992. A putative homolog of U2AF65 in *S. cerevisiae*. *Nucleic Acids Res* **20:** 4663.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72:** 291–336.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol Biol* **7:** 192.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26:** 73–79.

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

Cohen NE, Shen R, Carmel L. 2012. The role of reverse transcriptase in intron gain and loss mechanisms. *Mol Biol Evol* **29:** 179–186.

Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* **22:** 1053–1066.

de Smit MH, van Duin J. 1990. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci* **87:** 7668–7672.

Gahura O, Hammann C, Valentová A, Půta F, Folk P. 2011. Secondary structure is required for 3′ splice site recognition in yeast. *Nucleic Acids Res* **39:** 9759–9767.

Grate L, Ares M Jr. 2002. Searching yeast intron data at ares lab web site. In *Methods in enzymology* (ed. Christine G, Gerald RF), Vol. 350, pp. 380–392. Academic Press, New York.

Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6:** e1000664.

Habara Y, Urushiyama S, Tani T, Ohshima Y. 1998. The fission yeast *prp10*$^+$ gene involved in pre-mRNA splicing encodes a homologue of highly conserved splicing factor, SAP155. *Nucleic Acids Res* **26:** 5662–5669.

Hoskins AA, Moore MJ. 2012. The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem Sci* **37:** 179–188.

Hoskins AA, Friedman LJ, Gallagher SS, Crawford DJ, Anderson EG, Wombacher R, Ramirez N, Cornish VW, Gelles J, Moore MJ. 2011. Ordered and dynamic assembly of single spliceosomes. *Science* **331:** 1289–1295.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2:** 13–34.

Inglis DO, Arnaud MB, Binkley J, Shah P, Skrzypek MS, Wymore F, Binkley G, Miyasato SR, Simison M, Sherlock G. 2012. The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res* **40:** D667–D674.

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324:** 218–223.

Jeffares DC, Mourier T, Penny D. 2006. The biology of intron gain and loss. *Trends Genet* **22:** 16–22.

Juneau K, Miranda M, Hillenmeyer ME, Nislow C, Davis RW. 2006. Introns regulate RNA and protein abundance in yeast. *Genetics* **174:** 511–518.

Kendall MG, Stuart A. 1973. *The advanced theory of statistics*. Hafner Publishing, New York.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11:** 345–355.

Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA* **18:** 2174–2186.

Kim E, Goren A, Ast G. 2008. Alternative splicing: current perspectives. *Bioessays* **30:** 38–47.

Krämer A. 1996. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu Rev Biochem* **65:** 367–409.

Kriventseva EV, Gelfand MS. 1999. Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn* **17:** 281–288.

Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW. 2004. Introns and splicing elements of five diverse fungi. *Eukaryot Cell* **3:** 1088–1100.

Kurland CG. 1991. Codon bias and gene expression. *FEBS Lett* **285:** 165–169.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 2007. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. *Proc Natl Acad Sci* **104:** 19908–19913.

Le Hir H, Izaurralde E, Maquat LE, Moore MJ. 2000. The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions. *EMBO J* **19:** 6860–6869.

Le Hir H, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. *Trends Biochem Sci* **28:** 215–220.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci* **98:** 11193–11198.

Maniatis T, Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236–243.

Marshall AN, Montealegre MC, Jiménez-López C, Lorenz MC, van Hoof A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet* **9:** e1003376.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.

McKee AE, Silver PA. 2007. Systems perspectives on mRNA processing. *Cell Res* **17:** 581–590.

McManus CJ, Graveley BR. 2011. RNA structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev* **21:** 373–379.

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

Mishra SK, Ammon T, Popowicz GM, Krajewski M, Nagel RJ, Ares M Jr, Holak TA, Jentsch S. 2011. Role of the ubiquitin-like protein Hub1 in splice-site usage and alternative splicing. *Nature* **474:** 173–178.

Mitrovich QM, Tuch BB, Guthrie C, Johnson AD. 2007. Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*. *Genome Res* **17:** 492–502.

Mougin A, Grégoire A, Banroques J, Ségault V, Fournier R, Brulé F, Chevrier-Miller M, Branlant C. 1996. Secondary structure of the yeast *Saccharomyces cerevisiae* pre-U3A snoRNA and its implication for splicing efficiency. *RNA* **2:** 1079–1093.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Nasim MT, Eperon IC. 2006. A double-reporter splicing assay for determining splicing efficiency in mammalian cells. *Nat Protoc* **1:** 1022–1028.

Nedelcheva-Veleva MN, Sarov M, Yanakiev I, Mihailovska E, Ivanov MP, Panova GC, Stoynov SS. 2013. The thermodynamic patterns of eukaryotic genes suggest a mechanism for intron–exon recognition. *Nat Commun* **4:** 2101.

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, Fluhr R. 2004. Intron retention is a major phenomenon in alternative splicing in *Arabidopsis*. *Plant J* **39:** 877–885.

Nilsen TW. 2003. The spliceosome: the most complex macromolecular machine in the cell? *BioEssays* **25:** 1147–1149.

Okazaki K, Niwa O. 2000. mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res* **7:** 27–30.

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

Pérez-Valle J, Vilardell J. 2012. Intronic features that determine the selection of the 3′ splice site. *Wiley Interdiscip Rev RNA* **3:** 707–717.

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

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol* **5:** e90.

Reichert VL, Le Hir H, Jurica MS, Moore MJ. 2002. 5′ exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev* **16:** 2778–2791.

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI, et al. 2011. Comparative functional genomics of the fission yeasts. *Science* **332:** 930–936.

Rodríguez-Trelles F, Tarrío R, Ayala FJ. 2006. Origin and evolution of spliceosomal introns. *Annu Rev Genet* **40:** 47–76.

Rogozin I, Carmel L, Csuros M, Koonin E. 2012. Origin and evolution of spliceosomal introns. *Biol Direct* **7:** 11.

Roy SW, Gilbert W. 2005. The pattern of intron loss. *Proc Natl Acad Sci* **102:** 713–718.

Roy SW, Gilbert W. 2006. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet* **7:** 211–221.

Ruby SW, Abelson J. 1991. Pre-mRNA splicing in yeast. *Trends Genet* **7:** 79–85.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12:** 683–691.

Shepard PJ, Hertel KJ. 2008. Conserved RNA secondary structures promote alternative splicing. *RNA* **14:** 1463–1469.

Spingola M, Grate L, Haussler D, Ares M Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5:** 221–234.

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

Taylor JW, Berbee ML. 2006. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia* **98:** 838–849.

Toor N, Keating KS, Taylor SD, Pyle AM. 2008. Crystal structure of a self-spliced group II intron. *Science* **320:** 77–82.

Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* **107:** 3645–3650.

Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. 2011. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol* **12:** R110.

Wahl MC, Will CL, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136:** 701–718.

Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8:** 749–761.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci* **99:** 5860–5865.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics* **11:** 492–500.

Warf MB, Berglund JA. 2010. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35:** 169–178.

Wood V, Gwilliam R, Rajandream MA Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, et al. 2002. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415:** 871–880.

Wood V, Harris MA, McDowall MD, Rutherford K, Vaughan BW, Staines DM, Aslett M, Lock A, Bähler J, Kersey PJ, et al. 2012. PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res* **40:** D695–D699.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.

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

Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK. 2002. Minimal introns are not "junk". *Genome Res* **12:** 1185–1189.

Zhang J, Kuo CC, Chen L. 2011. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12:** 1–11.

Zhu T, Niu DK. 2013. Mechanisms of intron loss and gain in the fission yeast *Schizosaccharomyces*. *PLoS One* **8:** e61683.

Zur H, Tuller T. 2012. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep* **13:** 272–277.