Hindawi Publishing Corporation
Comparative and Functional Genomics
Volume 2011, Article ID 212146, 13 pages
doi:10.1155/2011/212146

Research Article

Factors Affecting Splicing Strength of Yeast Genes

Pinchao Ma[1] and Xuhua Xia[1, 2]

[1] *Department of Biology, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, ON, Canada K1N 6N5*
[2] *Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, ON, Canada K1H 8M5*

Correspondence should be addressed to Xuhua Xia, xxia@uottawa.ca

Received 30 July 2011; Accepted 6 September 2011

Academic Editor: John Parkinson

Accurate and efficient splicing is of crucial importance for highly-transcribed intron-containing genes (ICGs) in rapidly replicating unicellular eukaryotes such as the budding yeast *Saccharomyces cerevisiae*. We characterize the 5′ and 3′ splice sites (ss) by position weight matrix scores (PWMSs), which is the highest for the consensus sequence and the lowest for splice sites differing most from the consensus sequence and used PWMS as a proxy for splicing strength. *HAC1*, which is known to be spliced by a nonspliceosomal mechanism, has the most negative PWMS for both its 5′ ss and 3′ ss. Several genes under strong splicing regulation and requiring additional splicing factors for their splicing also have small or negative PWMS values. Splicing strength is higher for highly transcribed ICGs than for lowly transcribed ICGs and higher for transcripts that bind strongly to spliceosomes than those that bind weakly. The 3′ splice site features a prominent poly-U tract before the 3′AG. Our results suggest the potential of using PWMS as a screening tool for ICGs that are either spliced by a nonspliceosome mechanism or under strong splicing regulation in yeast and other fungal species.

1. Introduction

Introns in eukaryotic genes are spliced out mainly by the spliceosome [1] through a multitude of RNA-RNA, RNA-protein, and protein-protein interactions involving the 5′ splice site (ss), the 3′ ss, and the branch point sequence [2–10]. The yeast, *S. cerevisiae,* appears to have only U2-type introns [11, 12], with the consensus sequences of 5′ ss and 3′ ss being 5′-|GUAUGU and YAG|-3′, respectively [11, 13, 14, page 428]. 5′ ss is strongly constrained by base pairing with U1 and U6 snRNAs [15–17], leading to an overwhelming majority of 5′ ss having the consensus of GUAUGU in the yeast. In multicellular eukaryotes and in fission yeast, 3′ ss is strongly constrained by U2AF^{35} proteins [18]. However, no *U2AF35* homologue has been found in the budding yeast [19, 20], although 3′ ss is known to be partially constrained by the PRP8p protein [21].

A gene whose protein needs to be mass produced would need not only to have a high transcription rate, but also to possess features allowing it to be spliced efficiently and accurately. Thus, splicing is a major component of the quality control process in mRNA production in eukaryotes [22]. Highly expressed genes should evolve to have efficient 5′ ss and 3′ ss to avoid aberrant splicing which is not only wasteful but can also produce wrong proteins that perturb the normal cellular processes. In contrast, the selection for high splicing strength should be relatively weak in lowly expressed genes whose ss may drift to low splice efficiency through mutation. This has two implications concerning splicing strength. First, splicing strength of highly transcribed genes should, on average, be higher than that of lowly transcribed genes. Second, the variance of splicing strength should be larger for lowly transcribed genes (whose splicing strength could be high but may also drift to low values through mutation) than that for highly transcribed genes (whose splicing strength should all be high). This paper presents a first systematic analysis of the relationship between splicing strength and the level of gene expression.

A comprehensive assessment of the relationship between intron splicing strength and gene expression requires accurate characterization of introns and reliable large-scale measurement of gene expression. The yeast (*S. cerevisiae*) is the first species with accurate characterization of its introns and gene expression at mRNA and protein levels. Two powerful methods have recently been developed to characterize yeast introns. The first is to use high-density yeast tiling arrays in conjunction with a yeast mutant deficient for degradation of processed intron lariats [23]. The accumulation of lariats in the RNA pool is detected by the high-density tiling array which allows not only intron validation but also detection of new introns. The second approach involves designing microarray probes specific for exon-intron junctions and exon-exon junctions to quantitatively characterize unspliced and spliced mRNA [24–26]. Furthermore, *S. cerevisiae* is one of the few species with large-scale genome-wide characterization of both mRNA transcripts [27, 28] and protein abundance [29] or protein synthesis rate [30]. While transcripts and proteins have now been characterized for other species as well, there is no species in which introns have been characterized as accurately and thoroughly as the yeast. In this study, we use the yeast data to investigate the relationship between gene expression and splicing strength.

Other than the availability of high-quality molecular data, there are additional advantages in using *S. cerevisiae* for such a study. First, the yeast cells need to replicate rapidly and natural selection should act strongly against highly expressed yeast genes with poorly spliced introns. Second, the yeast genome has few introns, and most of them have been correctly annotated [23, 24, 31, 32]. Third, the splicing mechanism in the yeast is relatively simple compared to higher eukaryotes [24], with key spliceosome proteins better characterized than any other organisms [33]. Fourth, except for a few genes [31, 34], alternative splicing observed in multicellular eukaryotes is rare in the yeast [35]. The splicing mechanism in *S. cerevisiae* appears to be simple even among fungal species, for example, its genome does not have homologs of the U2AF^{35} spliceosomal protein which is present in other fungal species such as the fission yeast (*Schizosaccharomyces pombe*) as well as multicellular eukaryotes with sequenced genomes [19, 20]. *S. cerevisiae* also lacks the serine-arginine proteins serving as essential splicing factors in metazoans [36].

It is difficult to measure splicing strength directly, and previous publications have used the position weight matrix (PWM, [37, 38, pages 83–92] for detailed numerical illustration) derived from the ss and the resulting PWM score (PWMS) as a proxy for splicing strength [39, 40]. If an overwhelming majority of introns are spliced by the spliceosome mechanism, if there is an optimal state of the ss strongly preferred by the spliceosome mechanism (i.e., introns with ss in their optimal state are most efficiently spliced), and if there is strong selection pressure to maintain such an optimal state for most of the genes (i.e., if mutations leading to deviation from such an optimal state are deleterious), then we should expect that most ss should converge towards the optimal state, that the ss with the optimal state will have the highest PWMS, and that those deviating from the optimal state should have low PWMS. In short, PWMS may be used as a proxy for splicing strength by the spliceosome mechanism if the three conditions are satisfied. Hereafter, splicing strength refers specifically to splicing strength by the spliceosome mechanism.

2. Materials and Methods

2.1. 5′ and 3′ Splice Sites. The genomic sequences of all 16 chromosomes of *Saccharomyces cerevisiae* were retrieved from ftp.ncbi.nlm.nih.gov/genomes/Fungi/Saccharomyces_cerevisiae_uid128/ (assembly date: 14-JUL-2011). There are 279 annotated introns breaking the coding region in 270 genes, with 261 genes each containing a single intron and nine genes (SUS1, VMA9, HMRA1, DYN2, YOS1, RPL7A, AML1, TAD3, and RPL7B) each containing two introns. Some introns from paralogous genes are identical. Genes YBL111C, YHR218W, YLL067C, YLL066C, and YML133C are paralogous and contain the same intron, so are the genes YIL177C and YJL225C and the genes YRF1-3, YRF1-6, and YRF1-7. This creates two problems. The first involves the lack of data independence in statistical analysis. The second involves the quantification of mRNA and protein production. Take genes YIL177C and YJL225C, for example. It is difficult to know if the mRNA and protein abundance is contributed by only one of the two genes or by both. However, paralogues are few among yeast ICGs, and excluding these genes from analysis does not alter the conclusions reached in this paper.

There are 24 genes with introns in the 5′-UTR (Table 1). We originally thought that they might have weaker ss than those located within coding sequences because the failure to splice such introns seems to have little functional consequence as long as translation machinery can find the proper translation initiation site. However, there is no detectable difference between the two. Excluding or including these 24 yeast ICGs does not alter the conclusion in the paper.

For each intron, we originally extracted 10 bases from the exon side and 12 bases from the intron side by using DAMBE [41, 42]. This 10 + 12 configuration excluded some ss because the first exon in some yeast genes is shorter than 10 bases (note that the term "first exon" refers to the coding part of the first exon in this paper). For example, the first exon of the two-exon *MUD1* gene is only eight bases long. Because our extraction requires 10 bases on the exon side, 5′ ss of such genes would be missed. The most extreme cases of this are the *RPL20A* and *RPL20B* genes which have a single nucleotide as their first exon, that is, in the configuration of 5′-A|intron|TG-3′. With the requirement of 10 + 12 configuration, the total number of 5′ ss is only 223 in the yeast genome. As a preliminary analysis revealed that only five sites on the exon side of 5′ ss showed significant sequence conservation, we defined our 5′ ss to consist of 5 nucleotides on the exon site and 12 nucleotides on the intron side (referred to hereafter as the 5 + 12 configuration). Similarly, a 3′ ss consists of 12 nucleotides on the intron side and 5 nucleotides on the exon side. This results in 275 5′ ss and 301 3′ ss that have the 5 + 12 configuration, including the 24 introns in 5′UTR.

Some researchers (e.g., [39, 43]) have taken 5′ ss to span from the last 3 nucleotides of the exon to the first 6 or 7 nucleotides of the intron. 5′ ss defined in this way may produce spurious site patterns in the yeast. For example, as shown in Table 2 which lists genes with their 5′ ss excluded due to too short upstream exon, 20 *S. cerevisiae*

Table 1: The names and intron positions of 24 yeast protein-coding genes which have introns in their $5'$-UTRs.

Syst. name[(1)]	Std name[(2)]	Chr	Position[(3)]	Genome position	Strand[(4)]
YBL072C	RPS8A	2	-315..-8	89440..89133	C
YBL092W	RPL32	2	-333..-1	45645..45977	W
YBR089C-A	NHP6B	2	-384..-28	426873..426517	C
YDL061C	RPS29B	4	-421..-13	341219..340811	C
YDL137W	ARF2	4	-371..-40	216158..216489	W
YDL189W	RBS1	4	-138..-40	122078..122176	W
YDR099W	BMH2	4	-826..-84	652781..653523	W
YER102W	RPS8B	5	-367..-8	362733..363092	W
YER131W	RPS26B	5	-361..-1	423591..423951	W
YFR032C-A	RPL29	6	-334..-4	223771..223441	C
YGL031C	RPL24A	7	-463..-8	438397..437942	C
YGL187C	COX4	7	-354..-13	150525..150184	C
YGL189C	RPS26A	7	-378..-11	148966..148599	C
YGR027C	RPS25A	7	-327..-16	534785..534474	C
YGR148C	RPL24B	7	-399..-8	788178..787787	C
YIL123W	SIM1	9	-489..-3	127662..128148	W
YJL130C	URA2	10	-385..-66	172752..172433	C
YKL150W	MCR1	11	-144..-57	166400..166487	W
YKL186C	MTR2	11	-167..-14	93465..93312	C
YLR333C	RPS25B	12	-436..-14	796335..795913	C
YLR367W	RPS22B	12	-564..-8	855878..856434	W
YLR388W	RPS29A	12	-493..-6	898158..898645	W
YNL066W	SUN4	14	-358..-13	501157..501502	W
YPL230W	USV1	16	-93..-19	115219..115293	W

[(1)] Systematic name.
[(2)] Standard name.
[(3)] Site numbering relative to start codon.
[(4)] C—Crick strand (reverse complement), W—Watson strand.

genes have first exons with exactly three nucleotides (i.e., containing only the initiation codon). Defining $5'$ ss with three nucleotides in the exon side will substantially increase the representation of A, U, and G at the three nucleotide sites (i.e., the -3, -2, and -1 sites) in $5'$ ss (where the first nucleotide of the intron is labeled 1).

Some yeast introns might have been annotated incorrectly. The annotated intron in the *YJR112W-A* gene is the shortest intron in yeast (49 bp) and does not end with AG. It is possible that the intron is in fact longer with the real $3'$ ss further downstream. According to SGD annotation [44], *YJR112W-A* is described as "putative protein of unknown function, identified based on homology to *Ashbya gossypii.*" So, we excluded its $3'$ ss from our analysis. This reduces 303 $3'$ ss to 302.

2.2. Characterizing the Efficiency of Splicing Sites (ss) by Position Weight Matrix (PMW) and Sequence Logos. The consensus $5'$ ss on the intron side in the yeast is GUAUGU. Thus, a simple approach to characterize $5'$ ss splicing strength would be to give $5'$ ss a high splicing strength value if it is similar to the consensus but a low value if it is entirely different from the consensus. A more formal approach is to characterize the ss by a position weight matrix (PWM, [37, 38, pages 83–92] for detailed numerical illustration) and use the PWM score (PWMS) for each ss as its index of splicing strength [39, 40]. We used DAMBE [41, 42] to compute PWMS.

The nucleotide frequencies of entire transcripts (i.e., including both exons and introns) were used as background frequencies for computing PWM, with A = 0.3279, C = 0.1915, G = 0.2043, and U = 0.2763. Because some site-specific frequencies are 0, a pseudocount with $\alpha = 0.0001$ is added to all frequencies to avoid taking Log_2 of 0 [38, pages 83–92]. An alternative is to specify the nucleotide frequencies of all exons as the background frequencies for the exon part of the ss and nucleotide frequencies of all introns as the background frequencies for the intron part of the ss. However, results thus obtained are similar to those using the first approach. We have also obtained results by using nucleotide frequencies of the extracted ss as background frequencies. The results are again similar.

Several studies [43, 45] assumed equal background frequencies in characterizing ss with PWM. This is not a good approach because it confounds the site-specific nucleotide bias at the ss with the genomic nucleotide bias. For example, the yeast genome is AT rich, and sequence segments assembled randomly from an AT-rich nucleotide

Table 2: Yeast genes whose first exon (i.e., the coding part of the first exon) is shorter than five nucleotides.

Gene	PWM*	1st Exon len	Sequence
BET4	4.2977	3	AUG
BOS1	3.5760	3	AUG
DCN1	6.5363	3	AUG
MND1	8.1685	3	AUG
MPT5	8.5055	3	AUG
PSP2	8.4546	4	AUGG
QCR9	5.7592	3	AUG
RPL13A	6.9991	4	AUGG
RPL13B	8.6752	4	AUGG
RPL19A	6.9298	2	AU
RPL19B	11.7762	2	AU
RPL20A	9.7145	1	A
RPL20B	8.0214	1	A
RPL2A	12.0769	4	AUGG
RPL2B	9.7326	4	AUGG
RPL30	8.1799	3	AUG
RPL35A	7.3834	3	AUG
RPL35B	7.8326	3	AUG
RPL42A	9.2392	4	AUGG
RPL42B	7.0558	4	AUGG
RPL43A	9.9976	2	AU
RPL43B	12.0547	2	AU
RPS17A	9.1269	3	AUG
RPS17B	10.1283	3	AUG
RPS24A	9.4227	3	AUG
RPS24B	11.3548	3	AUG
RPS27A	6.3612	3	AUG
RPS27B	10.3823	3	AUG
RPS30A	10.8845	3	AUG
RPS30B	6.2290	3	AUG
UBC12	8.4505	3	AUG
VMA10	8.2722	3	AUG
YSF3	7.1596	3	AUG

* Position weight matrix score at 3′ ss.

pool will also be AT rich. Such random segments, when characterized by PWM with equal background frequencies, will appear informative and lead to false discovery of site patterns.

Another commonly used method for graphically displaying site-specific nucleotide patterns is the sequence logo which has been used to characterize intron ss [19]. The original method [46] does not take background nucleotide bias into consideration, and the resulting sequence logo is equivalent to a PWM assuming equal nucleotide frequencies. For example, AT-biased background frequencies in the yeast imply that the sequence logo will display A and T more prominently than C and G even when the sequences of interest contain no site-specific information. However, this problem has been eliminated by a recent improvement [47] which allows one to specify background (prior) frequencies just as in PWM. The sequence logographs in this paper are generated from the RNA Structure Logo website at http://www.cbs.dtu.dk/~gorodkin/appl/slogo.html.

2.3. Gene Expression. We used three measures of gene expression. The first is codon adaptation index [48] with its improved implementation in DAMBE [49], computed with the reference set of highly expressed yeast genes whose codon usage table is compiled in the Eysc_h.cut file distributed with EMBOSS [50]. The coding sequences (CDSs) for computing CAI were extracted by using DAMBE. CAI is intended to measure the efficiency of translation elongation but is highly and positively correlated with gene expression at the protein and mRNA level [51–53]. The advantage of using CAI is that it can be computed for all coding sequences, whereas empirical quantification of gene expression may be limited to relatively highly expressed genes which will not give us a whole picture of the relationship between splicing strength and gene expression.

The second measure of gene expression is the relative mRNA abundance of yeast genes from two previous studies that characterizes genome-wide RNA abundance in yeast [27, 28]. The microarray data [27] were downloaded from http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt. The data set includes mRNA levels for 5460 yeast genes. The absolute quantification data [28] is downloaded from the online supplementary material. Only the average expression in the YPD medium for 4817 genes was analyzed in this paper.

The third measure of gene expression is the protein production of yeast genes characterized in two previous studies. The protein abundance data [29] were downloaded from http://www.nature.com/nature/journal/v425/n6959/extref/nature02046-s2.xls. The predicted protein synthesis rate in two experimental conditions (mating pheromone treatment and control) was reliably measured for 3916 genes (Supplemental Table II in [30]), and we used the average of the two experiments.

In the mRNA and protein characterization, YAR044W is synonymous to YAR042W in the GenBank file, so is YDR474C to YDR475C, YJL018W to YJL019W, YJL021C to YJL020C, YPR090W to YPR089W, and YFR024C to YFR024C-A. Some genes (YEL068C, YER084W, YHR173C, YIL054W, YJR146W, YLR358C, YNL140C, YNL143C, YNL184C, and YOR105W) were annotated in SGD as “dubious open reading frame unlikely to encode a protein”, and are not annotated at all in the *S. cerevisiae* genome in NCBI. However, they were found to be expressed at both mRNA [27] and protein levels [29] and are therefore included in our analysis. YFL006W and YFL007W have been merged into YFL007W, YJL017W and YJL016W into YJL016W, and YOR087W and YOR088W into YOR087W in the most recent yeast genome annotation.

Two compiled data files are attached as supplementary materials. One (PWM-All.xls) includes all introns, mRNA abundance from the GATC-PCR method [28], and protein synthesis rate based on ribosomal loading and mRNA [30]. The other (PWM-No5UTRintrno.xls) excludes 5′ UTR

Table 3: Site-specific frequencies and position weight matrix (PWM) for 275 5′ ss. The consensus sequence (UA***AAG*** |***GUAUGUU*** UAAUU) can be obtained from those large site-specific PWM entries, with the most important sites in ***bold italics*** . The χ^2 test is performed for each site against the background frequencies (A = 0.3279, C = 0.1915, G = 0.2043, and U = 0.2763). The nucleotide sites are labeled with the five exon nucleotides as −5 to −1 and the 12 intron nucleotides as 1 to 12. The PWM is nearly identical when the introns in 5′ UTR were excluded.

Site	A	C	G	U	χ^2	P	A	C	G	U
−5	94	32	57	92	11.798	0.0081088	0.0641	−0.7117	0.0245	0.2792
−4	119	47	48	61	14.117	0.0027505	0.4032	−0.1599	−0.2225	−0.3115
−3	139	38	43	55	39.672	0.0000001	***0.6268***	−0.4651	−0.3805	−0.4601
−2	138	40	36	61	38.899	0.0000001	***0.6164***	−0.3915	−0.6355	−0.3115
−1	91	45	88	51	27.270	0.0000052	0.0174	−0.2223	***0.6492***	−0.5685
1	0	1	274	0	1060.426	0.0000004	−8.1042	−5.4675	***2.2855***	−8.1044
2	0	9	0	266	658.096	0.0000003	−8.1042	−2.5200	−8.1048	***1.8081***
3	268	1	2	4	522.754	0.0000003	***1.5723***	−5.4675	−4.6732	−4.1523
4	17	29	1	228	428.607	0.0000002	−2.3805	−0.8528	−5.5454	***1.5859***
5	2	0	272	1	1041.047	0.0000004	−5.2765	−8.1049	***2.2750***	−5.8967
6	10	8	2	255	583.545	0.0000003	−3.1271	−2.6862	−4.6732	***1.7472***
7	97	18	39	121	55.570	0.0000001	0.1092	−1.5351	−0.5206	***0.6734***
8	95	54	35	91	11.363	0.0099180	0.0793	0.0397	−0.6759	0.2635
9	123	45	34	73	22.172	0.0000601	0.4508	−0.2223	−0.7175	−0.0534
10	118	41	38	78	17.334	0.0006034	0.3911	−0.3560	−0.5579	0.0418
11	105	33	43	94	17.367	0.0005940	0.2232	−0.6676	−0.3805	0.3101
12	90	44	42	99	12.109	0.0070180	0.0015	−0.2546	−0.4142	0.3847

introns and includes mRNA abundance from microarray [27] and protein abundance data [29].

3. Results and Discussion

3.1. Position Weight Matrix (PWM) and Its Statistical Significance. Consistent with previous experimental studies on *S. cerevisiae*, the position weight matrices (Tables 3 and 4) and the sequence logos (Figure 1) not only confirmed but also expanded the consensus sequence of yeast splice sites, with UAAAG|GUAUGUUUAAUU as the strongest 5′ ss and UUUUUUUUAYAG|GCUUC as the strongest 3′ ss. Whether a PWM contains significant site-specific information can be tested by using the F statistic [37] defined as

$$F = \sum_{i=1}^{4}\sum_{j=1}^{L} p_{ij} \ln \frac{p_{ij}}{p_i}, \tag{1}$$

where L is the sequence length (motif width equal to 17 in our study), $i = 1,\ 2,\ 3$, and 4 corresponding to A, C, G, and U, p_i is the background nucleotide frequency for nucleotide i, and p_{ij} is the frequency of nucleotide i at position $j(= 1, 2, \ldots, 17)$. A straightforward method for evaluating the significance of PWM is by resampling. With the tetranomial distribution defined by $(p_\text{A} + p_\text{C} + p_\text{G} + p_\text{T})^L$, we can obtain a new set of sequences (e.g., 246 sequences of 17 nt each) and compute F. This is repeated for, say, 5000 times to obtain 5000 F values. The 95th or 99th percentile of the F values can be taken as critical F values at 0.05 and 0.01 significance levels, respectively. An observed F for the PWM is significant if it is greater than the critical F. Based on this criterion, the PWM from the 275 5′ ss and that from the 301 3′ ss are both highly significant ($P < 0.0001$). It is also highly significant ($P < 0.0001$) when the 24 introns in 5′ UTR are excluded.

Given the significant PWM for 5′ ss and 3′ ss, we want to know which individual nucleotide sites (out of 17 in total) contribute to the significance. All 17 nucleotide sites of 5′ ss and 16 nucleotide sites of 3′ ss are significant at 0.05 level when experimentwise error rate is not controlled for (Tables 3 and 4). One popular statistical method for controlling experimentwise error rate is the method of false discovery rate (FDR) [54, 55]. The classical FDR approach [54], commonly referred to as the Benjamini-Hochberg procedure or simply the BH procedure, sorts p values in ascending order and computes $p_{\text{critical.BH}.i}$ (where the subscript BH stands for the BH procedure) for the ith p value as

$$p_{\text{critical.BH}.i} = \frac{q \cdot i}{N}, \tag{2}$$

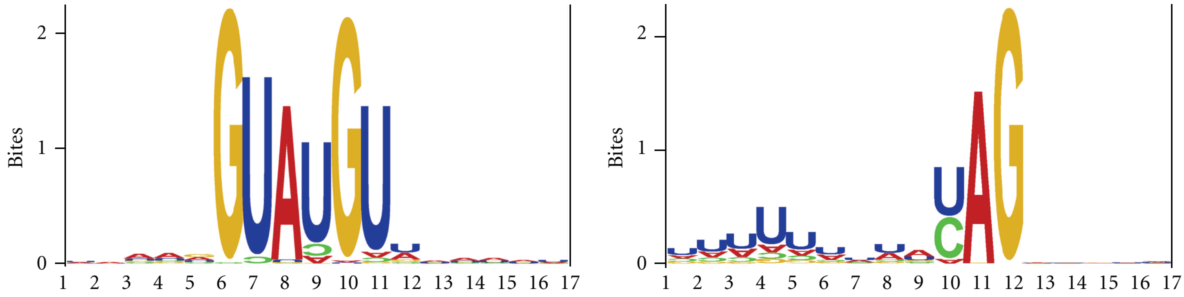
where q is FDR (e.g., 0.05), i is the rank of the p value in the sorted array of p values, and N is the number of tests (i.e., the number of p values, 17 in our case). If k is the largest i satisfying the condition of $p_i \le p_{\text{critical.BH}.i}$, then we reject hypotheses from H_1 to H_k. In our case, all the 17 nucleotide sites are statistically significant based on $p_{\text{critical.BH}.i}$ (Table 5).

The FDR procedure above assumes that the test statistics are independent or positively dependent and a more conservative FDR procedure has been developed that relaxes the assumption [55]. This method, commonly referred to as the Benjamini-Yekutieli or simply the BY procedure, computes $p_{\text{critical.BY}.i}$ for the ith hypothesis as

$$p_{\text{critical.BY}.i} = \frac{q \cdot i}{N \sum_{i=1}^{N} 1/i} = \frac{p_{\text{critical.BH}.i}}{\sum_{i=1}^{N} 1/i}. \tag{3}$$

Table 4: Site-specific frequencies and position weight matrix (PWM) for 301 3′ ss. The consensus sequence (***UUUUUUUUAYAG*** |GCUUC) can be obtained from those large site-specific PWM entries, with the most important sites in ***bold italics*** . The χ^2 test is performed for each site against the expected background frequencies. The sites are labeled with first-exon site as 1. The PWM is nearly identical when the introns in 5′ UTR were excluded.

Site	A	C	G	U	χ^2	P	A	C	G	U
−12	70	58	37	136	51.729	0.0000001	−0.4898	0.0122	−0.7264	***0.7114***
−11	79	51	23	148	79.511	0.0000001	−0.3161	−0.1727	−1.4074	***0.8332***
−10	86	45	14	156	105.131	0.0000001	−0.1941	−0.3525	−2.1155	***0.9090***
−9	43	33	23	202	236.063	0.0000001	−1.1886	−0.7978	−1.4074	***1.2812***
−8	56	43	31	171	130.216	0.0000001	−0.8100	−0.4178	−0.9801	***1.0412***
−7	102	35	31	133	54.256	0.0000001	0.0512	−0.7134	−0.9801	***0.6793***
−6	103	46	38	114	23.130	0.0000380	0.0653	−0.3210	−0.6881	***0.4574***
−5	100	36	25	140	68.925	0.0000000	0.0228	−0.6729	−1.2882	***0.7532***
−4	145	27	41	88	45.473	0.0000001	***0.5574***	−1.0854	−0.5790	0.0850
−3	15	127	0	159	284.824	0.0000002	−2.6877	***1.1404***	−8.2350	***0.9364***
−2	299	1	1	0	605.789	0.0000003	***1.5998***	−5.5977	−5.6756	−8.2346
−1	0	0	301	0	1171.443	0.0000004	−8.2345	−8.2351	***2.2908***	−8.2346
1	109	39	74	79	9.936	0.0191208	0.1467	−0.5580	0.2697	−0.0701
2	84	66	55	96	6.036	0.1098600	−0.2279	0.1981	−0.1571	0.2102
3	103	58	50	90	2.969	0.3964877	0.0653	0.0122	−0.2940	0.1173
4	96	45	56	104	8.655	0.0342400	−0.0359	−0.3525	−0.1312	0.3253
5	100	69	39	93	11.698	0.0084938	0.0228	0.2620	−0.6508	0.1645



(a)

(b)

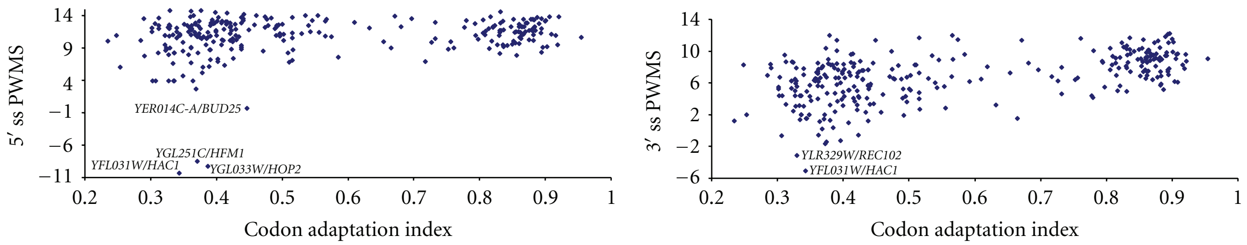
Figure 1: Sequence logos of 5′ ss (a) and 3′ ss (b), produced with the background frequencies specified as A = 0.3279, C = 0.1915, G = 0.2043, and U = 0.2763. The nucleotides whose frequencies are lower than expected are plotted upside down. The vertical bar is the information index computed as $-[\sum P_i \log_2(P_i)]$, where P_i is the frequency of nucleotide i (= A, C, G or U) at each site.

With $N = 17$ in our case, $\sum 1/k = 3.439552523$. Based on $p_{\text{critical.BY.}i}$, nucleotide sites −5 and −4 in 5′ ss are not statistically significant (Table 5).

All 17 nucleotide sites of 3′ ss are also significant at the 0.05 level based on the criterion of $p_{\text{critical.BH.}i}$. However, with the more conservative criterion of $p_{\text{critical.BY.}i}$, the five nucleotide sites on the exon side are not significant.

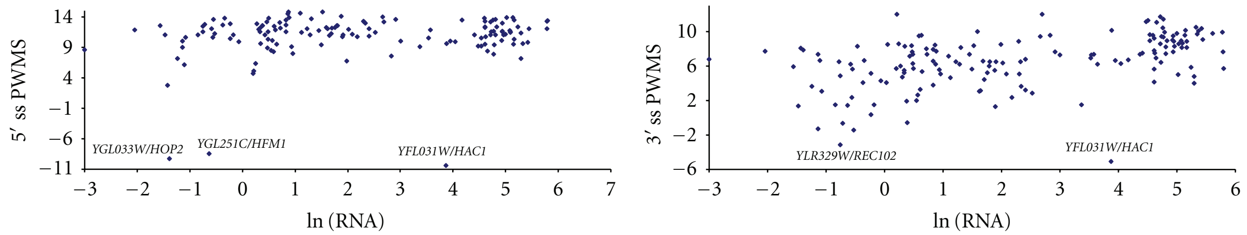
There is no significant difference in 5′ and 3′ ss PWMS between the 24 introns in 5′ UTR and those in the coding regions ($P = 0.1606$ for 5′ ss PWMS and $P = 0.3182$ for 3′ ss PWMS). The two sets are pooled in the rest of the analysis.

3.2. Gene Expression and Splicing Strength. We have argued previously that lowly expressed genes will, on average, have introns with lower splicing strength (as measured by PWMS) but greater variance in PWMS than highly expressed genes. The splicing strength characterized by PWMS exhibited expected relationship with gene expression when the latter is measured by either CAI (Figure 2), mRNA abundance (Figure 3), or protein production (Figure 4). In addition, lowly expressed genes have greater variation in PWMS values than highly expressed genes. To statistically test the differences in mean and variance, we have ranked genes by gene expression, that is, ranked separately by CAI, mRNA abundance, or protein production. For each ranking, we designate 1/3 of the genes with the highest expression values (i.e., highest CAI, mRNA, or protein production, resp.) as the high-expression group and another 1/3 of the genes with the lowest expression values as the low-expression group and tested the differences in mean PWMS and the variance of PWMS between the two groups. As shown in Table 7, the two predictions are consistently supported, that is, (1) the highly expressed genes have significantly greater mean PWMS values than lowly expressed genes and (2) the highly



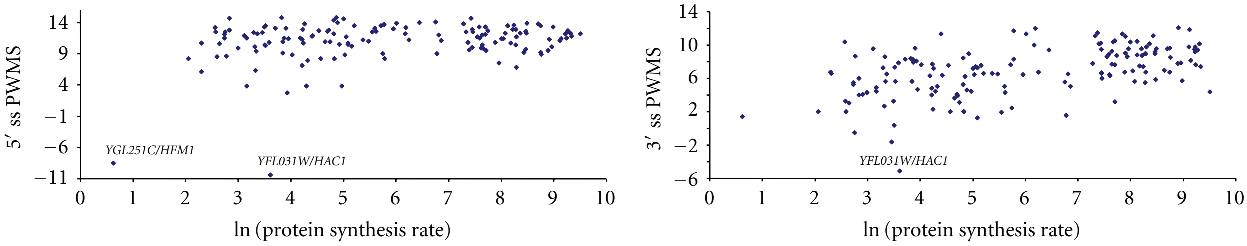
(a) (b)

Figure 2: Relationship between splicing strength measured by position weight matrix score (PWMS) at 5′ (a) and 3′ (b) splice sites (5′ ss and 3′ ss) and gene expression measured by codon adaptation index.



(a) (b)

Figure 3: Relationship between splicing strength measured by position weight matrix score (PWMS) at 5′ (a) and 3′ (b) splice sites (5′ ss and 3′ ss) and gene expression measured by mRNA abundance [28]. The mRNA abundance is log transformed. A similar pattern is observed when the mRNA abundance from Holstege et al. [27] is used.



(a) (b)

Figure 4: Relationship between splicing strength measured by position weight matrix score (PWMS) at 5′ (a) and 3′ (b) splice sites (5′ ss and 3′ ss) and gene expression measured by protein synthesis rate [30] which is log transformed. A similar pattern is observed when the protein synthesis rate is replaced by protein abundance from Ghaemmaghami et al. [29].

expressed genes have significantly smaller variance in PWMS than the lowly expressed genes (Table 7). The t-tests used assume unequal variances between the two groups. The tests for differences in variance between the two groups are regular variance ratio F-test [56, pages 136–139].

The cluster of points in Figure 2 with CAI greater 0.8 and that in Figure 3 with ln(mRNA) greater than 3 are almost all ribosomal protein-coding genes which are highly transcribed [57] and have strong splicing sites (high PWMS). For these genes, mutations that weaken the splicing strength of their splicing sites are expected to be deleterious. Our result suggests that natural selection may be involved in maintaining high splicing strength in the splice sites of highly expressed genes.

3.3. Introns with the Poorest PWMSs for Their ss Are Spliced by Nonspliceosome Mechanisms or Require Additional Splicing Factors. 5′ ss in three genes (*HAC1*, *HFM1*, and *HOP2*) have the most negative PWMSs (−10.3544, −9.2192, and −8.4717, resp.). Such PWMSs imply that 5′ ss of these genes have evolved to avoid being spliced by the spliceosomal mechanism because a random 17mer assembled from the nucleotide pool with the nucleotide frequencies of yeast protein-coding genes (A = 0.3279, C = 0.1915, G = 0.2043, and U = 0.2763) would have an expected PWMS of zero.

It is now known that the splicing of the pre-mRNA of these genes requires either a nonspliceosomal mechanism or additional protein factors for intron removal. *HAC1*, which plays a key role in the unfolded protein response (UPR)

Table 5: Evaluating statistical significance of individual nucleotide sites (site, with 5 nucleotides on the exon side labelled −5 to −1 and 12 on the intron side labeled 1 to 12) of 5′ ss by two types of false discovery rate.

Site	*P*	pBH(1)	pBY(2)
1	*0.0000000000†	0.002941	0.000855
5	*0.0000000000†	0.005882	0.001710
2	*0.0000000000†	0.008824	0.002565
6	*0.0000000000†	0.011765	0.003420
3	*0.0000000000†	0.014706	0.004276
4	*0.0000000000†	0.017647	0.005131
7	*0.0000000000†	0.020588	0.005986
−2	*0.0000004842†	0.023529	0.006841
−3	*0.0000013734†	0.026471	0.007696
−1	*0.0000030965†	0.029412	0.008551
9	*0.0002619304†	0.032353	0.009406
10	*0.0006307900†	0.035294	0.010261
12	*0.0025004071†	0.038235	0.011116
11	*0.0033589734†	0.041176	0.011971
8	*0.0084455695†	0.044118	0.012827
−5	*0.0177349476	0.047059	0.013682
−4	*0.0182291629	0.050000	0.014537

(1) Critical *P* based on [54].
(2) Critical *P* based on [55].
*Significant by the criterion in [54].
†Significant by the criterion in [55].

by binding to the UPR element [58–62], is one of the few yeast genes whose first exon is much longer than the second (661 bp and 56 bp, resp.), and its transcript is processed by an unconventional mechanism (i.e., nonspliceosomal splicing), with the intron cleaved by the protein kinase Ire1p, which possesses endonuclease activity and tRNA ligase [61, 63–65].

The *HFM1/MER3* and *HOP2* are both meiosis-specific genes, with *HFM1* coding for a meiosis-specific DNA helicase [66, 67] that participates in crossover control and unwinding of Holliday junctions [66–70] and *HOP2* coding for a protein essential for forming meiotic synapsis between homologous chromosomes [71, 72]. The splicing of their transcripts is not constitutive but strictly regulated. The splicing of the *HFM1/MER3* transcripts is regulated by the Mer1p and Bud13p proteins [73–75]. Unspliced *HOP2* transcripts accumulate when the cell is not in meiosis [33]. The splicing of the *HOP2* transcripts depends heavily on the nuclear exosome component Rrp6 protein, with the loss of *RRP6* dramatically decreasing the splicing of *HOP2* transcripts [33].

Other than the three genes above, the gene with the smallest 5′ ss PWMS is *BUD25* with its PWMS equal to 0.4267. Yeast spliceosome does not bind to *BUD25* transcripts during transcription [33]. The *BUD25* gene is also implicated in chromosome segregation and meiosis [76]. Most yeast introns can be deleted with no effect, but deletion of *BUD25* intron causes defective growth [76], suggesting that splicing is important for its function and that its ss may be under additional constraints other than splicing strength. In other words, the ss of *BUD25* may not be free to evolve towards high splicing strength.

3′ ss of several genes also have negative PWMSs. The intron of the *HAC1* gene, which is spliced by a nonspliceosome mechanism [61, 63–65], has a 3′ ss with the smallest PWMS (−5.1038). The intron whose 3′ ss has the second smallest PWMS (−3.1252) belongs to *REC102* which is also a meiosis-specific gene, required for chromosome synapsis [77–79]. The splicing of its intron also makes use of a nonspliceosome mechanism [31]. Based on *in vitro* experiments, splicing of *REC102* message by yeast spliceosome is both inefficient [80] and inaccurate [80, 81], leading to many unspliced and wrongly spliced mRNAs. However, in a large-scale characterization of yeast total transcripts, only a single correctly spliced mRNA is found (file S03052-07_G10.seq in the online supplementary material in [82]). The amino acid sequence from the correctly spliced mRNA is highly conserved among different *Saccharomyces* species [80, 83–85]. Taken together, these results suggest that the *in vivo* correct splicing of *REC102* pre-mRNA requires additional factors at the meiosis stage. These yeast ICGs whose intron splicing requires a nonspliceosome mechanism or additional splicing factors are poor in recruiting U1 snRPNs [33]. The result that such genes are strongly regulated and have low or negative PWMS indicates the potential of using bioinformatic methods to identify these strongly regulated genes.

3.4. A Prominent Poly-U Upstream of the AG Dinucleotide in 3′ ss.

Efficiently spliced introns in the yeast are characterized by a poly-U tract upstream of the 3′AG (Table 4 and Figure 1). This trend is stronger when we exclude the yeast ICGs whose transcripts bind poorly to spliceosomes (result not shown). Such a poly-U tract can increase the efficiency of 3′ ss that has previously been demonstrated in *S. cerevisiae* [86], especially in introns with a long distance between the branch point site and 3′ ss [87]. A recent study of intron splicing of mammalian genes in YACs (yeast artificial chromosome) is consistent with the proposed importance of the poly-U tract upstream of 3′ ss in *S. cerevisiae* [88].

Previous compilations of yeast introns [11, 13] have missed the poly-U tract upstream of 3′ ss. Thus, the poly-U tract upstream of 3′ ss has not been included as a feature of *S. cerevisiae* intron in molecular biology textbooks (e.g., [14, page 428]).

The poly-U tract upstream of the yeast 3′ ss is different from the polypyrimidine tract (where both U and C are overrepresented) that is often present upstream of 3′ ss in multicellular eukaryotes as well as in *Schizosaccharomyces pombe.* In *S. cerevisiae,* only U is overrepresented and C is underrepresented (shown backwards in the sequence logo for 3′ ss in Figure 1(b)). In multicellular eukaryotes and *S. pombe,* the polypyrimidine tract upstream of 3′ ss is important for splicing strength [89] and is recognized by the essential U2AF^{65} splicing factor [90]. However, while U2AF^{65} is highly conserved from *S. pombe* to multicellular eukaryotes, the U2AF^{65} homologue in the budding yeast, MUD2p, is highly diverged and not essential for survival

Table 6: Position weight matrix scores (PWMSs, as a proxy for splicing strength) is significantly smaller for splice sites from intron-containing genes (ICGs) whose transcripts failed to recruit U1 snRNPs (NRG for nonrecruiting group) than for those from ICGs whose transcripts bind well to U1 snRNPs (RG for recruiting group). The pattern is consistent for both 5′ ss and 3′ ss, based on two-sample t-tests assuming equal variances. Mann-Whitney tests yield the same conclusion.

	5′ ss		3′ ss	
	NRG	RG	NRG	RG
PWMS mean	8.8138	11.1978	5.3129	7.1762
PWMS Var.	31.5069	4.8646	13.3017	8.2077
N	44	231	49	252
t	−4.6346		−3.9257	
P	0.0000		0.0001	

[20]. This may have contributed to the evolutionary origin of the poly-U tract in the budding yeast.

The presence of poly-U implies that the recognition of 3′ ss may involve more than simple scanning for the first AG after the branchpoint site. In fact, it has previously been shown that a proximal PyAG without poly-U is often skipped if a more distal PyAG occurs with a poly-U [86].

3.5. Yeast ICGs Weakly Bound to U1 snRNPs Have Smaller PWMS Than Those Bound Strongly. A recent study [33] documented 50 yeast ICGs whose mRNA failed to recruit U1 snRNPs to the site of transcription in detectable amount. We tested the possibility that these genes may have weak ss by comparing PWMS between these 50 genes and other yeast ICGs. Three of these 50 genes (YOR074C, YOR221C, and YLR312W-A) actually do not have introns and should not be included as yeast ICGs. In addition, YOR318C is a dubious gene with no *in vivo* evidence, that it is, putative intron is spliced. The remaining 46 genes have 48 introns, with YCL005W-A and YCR097W each having two introns. The mean PWMS is 8.8138 for the 5′ ss of these 48 introns and 11.1978 for the rest of introns. The difference is highly significant based on a two-sample t-test (DF = 273, t = −4.6346, P < 0.0001, two-tailed test, Table 6). The same pattern is seen for 3′ ss (Table 6). Thus, Yeast ICGs weakly bound to U1 snRNPs during transcription have weaker 5′ and 3′ ss than those bound strongly to U1 snRNPs.

It is not clear why Yeast ICGs weakly bound to U1 snRNPs during transcription should have weak 3′ ss because 3′ ss is not expected to be involved in recruiting U1 snRNPs during transcription. One possible explanation is that a weak 5′ ss that does not recruit U1 efficiently tends to be associated with a weak 3′ ss.

4. Discussion

There has been no large-scale experimental characterization of splicing efficiency, so it is difficult to relate PWMS as a proxy of splicing strength to splicing efficiency. However, the three yeast ICGs whose splicing depends on Mer1p have experimentally measured splicing efficiency expressed as percentage of transcripts spliced in the wild type [73]. When *Mer1* is not expressed, these percentages are 32% and 31% for *AMA1*, 14% and 13% for *REC107/Mer2*, and 3% and 4% for *HFM1/Mer3*. The corresponding values when *Mer1* is expressed are 71% and 72% for *AMA1*, 53% and 59% for *REC107/Mer2*, and 42% and 42% for *HFM1/Mer3*. Consistent with this ranking of splicing efficiency of *AMA1 > REC107/Mer2 > HFM1/Mer3,* 5′ ss and 3′ ss PWMS values are 8.7838 and 11.4573 for *AMA1,* 4.6995, and 5.6668 for *REC107/Mer2*, and −7.3825 and 1.3488 for *HFM1/Mer3*. Thus, in this limited case, the experimentally measured splicing efficiency shows excellent concordance with PWMS.

Three additional but indirect lines of evidence suggest that PWMS is an appropriate proxy for splicing strength. First, PWMS for both 5′ ss and 3′ ss is positively correlated with gene expression. Second, introns spliced by nonspliceosomal mechanisms or requiring additional protein factors for splicing generally have low PWMS. Third, yeast ICGs that recruit splicing factors poorly tend to have lower PWMS than those that bound well to splicing factors. These results suggest the potential of using PWMS as a screening tool for ICGs not spliced by the spliceosome mechanism or requiring additional regulatory factors for splicing in other fungal species.

The characterized 5′ ss (UA***AAG***|***GUAUGUU***UAAUU, where significant sites are in bold italic) and 3′ ss (***UUUUUUUUAYAG***|GCUUC) in the yeast expanded the conventional yeast consensus splice sites. While the +1 site in the 3′ ss site is not statistically significant after adjusting for experimentwise error rate, a recent experimental study suggests that it does affect splicing efficiency. The overused nucleotide at this site is G, followed by A, but C is strongly avoided (Table 4). Changing the +1A to +1C in the *LSM7* mRNA resulted in splicing at a downstream AG|A site [81]. In *REC102* gene, the splicing at the normal 3′ ss site UGAAG|A site is reduced when the +1A is changed to C, especially when nucleotide C in an upstream AG|C site is changed to A [81]. Our results (Table 4) showing the preference of +1R and avoidance of −1C corroborate these experimental studies and suggest that the preference of +1R and avoidance of −1C may be a general feature of splicing by yeast spliceosome.

Selection for increased splicing strength is expected to be stronger in highly transcribed genes than in lowly transcribed genes. Consistent with this expectation, splicing strength is higher for highly transcribed ICGs than for lowly transcribed ICGs and is higher for ICGs whose mRNAs are efficiently translated than those whose mRNAs are not efficiently translated. It has long been known that highly expressed yeast genes exhibit a high degree of codon-anticodon adaptation [91]. Our result here suggests that natural selection is also operating on the splicing machinery.

The presence of poly-U immediately before the 3′ AG in 3′ ss, instead of a polypyrimidine tract in other eukaryotes, may arise from the following evolutionary process. The polypyrimidine is recognized by, and bound to, U2AF^{65} in other eukaryotes including *S. pompe*. The U2AF^{65} homologue in the budding yeast, MUD2p, is highly diverged and not essential for survival [20]. This may have contributed

Table 7: Testing the predictions that introns in highly expressed genes have higher PWMS and smaller variance in PWMS than in lowly expressed genes, with gene expression measured by CAI, mRNA, and protein abundance. Introns spliced by nonspliceosomal mechanisms are excluded. Mann-Whitney tests generate similar results. All tests are two tailed. The results are nearly identical when mRNA abundance from microarray [27] is used instead of that from GATC-PCR [28] or when protein abundance [29] is used instead of the protein synthesis rate [30].

	5′ ss			3′ ss		
	CAI	lnMRNA[1]	lnPROT[2]	CAI	lnMRNA	lnPROT
N[3]	91	48	55	100	53	67
MeanH[4]	11.4927	11.3128	11.4879	8.7188	8.5447	8.6004
MeanL[5]	9.4135	9.7143	9.7581	5.1109	4.9729	5.3359
DF[6]	113	59	55	155	83	82
T	4.1635	2.2501	2.2411	9.7833	6.9719	5.7687
P	0.0001	0.0282	0.0291	<0.0001	<0.0001	<0.0001
VarH[7]	3.2053	2.8657	2.7345	2.6326	3.4395	3.4220
VarL[8]	10.3934	21.3602	24.6692	20.0609	10.4712	9.9223
F	3.2429	7.4537	9.0214	7.6211	3.0444	2.9000
P	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0001

[1] Natural logarithm of mRNA abundance [28].
[2] Natural logarithm of protein synthesis rate [30].
[3] Number of ss in the highly expressed and lowly expressed groups (note that $N_1 = N_2 = N$).
[4] Mean PWMS in highly expressed group.
[5] Mean PWMS in lowly expressed group.
[6] The t-test assuming unequal variance is used. SoDF is not equal to $(N_1 + N_2 - 2)$.
[7] Variance in the highly expressed group.
[8] Variance in the lowly expressed group.

to a weakened selection constraint on the evolution of the polypyrimidine tract in the budding yeast. Because the budding yeast genome is AT rich, nucleotide C may be progressively replaced by nucleotide T, leading to the transition of the polypyrimidine tract to the poly-U tract. However, this mutationist hypothesis cannot explain why the poly(U) can increase splicing efficiency. It is likely that both mutation and selection participated in the evolution of poly(U) in the yeast intron before the 3′ AG.

We should mention that splicing efficiency of yeast introns depends not only on 5′ ss and 3′ ss, but also on the branchpoint sequence (BPS, [92–94]) as well as the spacing between BPS and 3′ ss [95–97]. For example, both 5′ and 3′ ss of *Yra1* are strong with high PWMS, but the intron splicing is regulated, possibly through its unconventional branchpoint site GACUAAC (in contrast to the consensus UACUAAC).

Acknowledgments

This study is supported by NSERC's Discovery and Strategic Grants. The author thanks S. Aris-Brosou, G. Palidwor, M. Ragonnet, E. Rocha, A. van Weringh, H. Xiong, Q. Yang, and X. Yao for discussion and comments.

References

[1] P. A. Sharp, "The discovery of split genes and RNA splicing," *Trends in Biochemical Sciences*, vol. 30, no. 6, pp. 279–281, 2005.

[2] R. A. Padgett, P. J. Grabowski, and M. M. Konarska, "Splicing of messenger RNA precursors," *Annual Review of Biochemistry*, vol. 55, pp. 1119–1150, 1986.

[3] H. Du and M. Rosbash, "The U1 snRNP protein U1C recognizes the 5′ splice site in the absence of base pairing," *Nature*, vol. 419, no. 6902, pp. 86–90, 2002.

[4] M. Freund, M. J. Hicks, C. Konermann, M. Otte, K. J. Hertel, and H. Schaal, "Extended base pair complementarity between U1 snRNA and the 5′ splice site does not inhibit splicing in higher eukaryotes, but rather increases 5′ splice site recognition," *Nucleic Acids Research*, vol. 33, no. 16, pp. 5112–5119, 2005.

[5] O. A. Kent, D. B. Ritchie, and A. M. MacMillan, "Characterization of a U2AF-independent commitment complex (E′) in the mammalian spliceosome assembly pathway," *Molecular and Cellular Biology*, vol. 25, no. 1, pp. 233–240, 2005.

[6] M. Lund and J. Kjems, "Defining a 5′ splice site by functional selection in the presence and absence of U1 snRNA 5′ end," *RNA*, vol. 8, no. 2, pp. 166–179, 2002.

[7] P. A. Sharp, "Split genes and RNA splicing," *Cell*, vol. 77, no. 6, pp. 805–815, 1994.

[8] C. G. Simpson, G. Thow, G. P. Clark, S. N. Jennings, J. A. Watters, and J. W. S. Brown, "Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon," *RNA*, vol. 8, no. 1, pp. 47–56, 2002.

[9] T. W. Nilsen, "Spliceosome assembly in yeast: one ChIP at a time?" *Nature Structural and Molecular Biology*, vol. 12, no. 7, pp. 571–573, 2005.

[10] M. S. Jurica and M. J. Moore, "Pre-mRNA splicing: awash in a sea of proteins," *Molecular Cell*, vol. 12, no. 1, pp. 5–14, 2003.

[11] P. A. Sharp and C. B. Burge, "Classification of introns: U2-type or U12-type," *Cell*, vol. 91, no. 7, pp. 875–879, 1997.

[12] C. B. Burge, R. A. Padgett, and P. A. Sharp, "Evolutionary fates and origins of U12-type introns," *Molecular Cell*, vol. 2, no. 6, pp. 773–785, 1998.

[13] I. J. Jackson, "A reappraisal of non-consensus mRNA splice sites," *Nucleic Acids Research*, vol. 19, no. 14, pp. 3795–3798, 1991.

[14] R. F. Weaver, *Molecular Biology*, McGraw-Hill Higher Education, Boston, UK, 3rd edition, 2005.

[15] C. F. Lesser and C. Guthrie, "Mutational analysis of pre-mRNA splicing in *Saccharomyces cerevisiae* using a sensitive new reporter gene, CUP1," *Genetics*, vol. 133, no. 4, pp. 851–863, 1993.

[16] C. F. Lesser and C. Guthrie, "Mutations in U6 snRNA that alter splice site specificity: implications for the active site," *Science*, vol. 262, no. 5142, pp. 1982–1988, 1993.

[17] E. J. Sontheimer and J. A. Steitz, "The U5 and U6 small nuclear RNAs as active site components of the spliceosome," *Science*, vol. 262, no. 5142, pp. 1989–1996, 1993.

[18] C. J. Webb and J. A. Wise, "The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans," *Molecular and Cellular Biology*, vol. 24, no. 10, pp. 4229–4240, 2004.

[19] D. M. Kupfer, S. D. Drabenstot, K. L. Buchanan et al., "Introns and splicing elements of five diverse fungi," *Eukaryotic Cell*, vol. 3, no. 5, pp. 1088–1100, 2004.

[20] N. Abovich, X. C. Liao, and M. Rosbash, "The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition," *Genes and Development*, vol. 8, no. 7, pp. 843–854, 1994.

[21] C. A. Collins and C. Guthrie, "Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome," *Genes and Development*, vol. 13, no. 15, pp. 1970–1982, 1999.

[22] L. E. Maquat and G. G. Carmichael, "Quality control of mRNA function," *Cell*, vol. 104, no. 2, pp. 173–176, 2001.

[23] K. Juneau, C. Palm, M. Miranda, and R. W. Davis, "High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 5, pp. 1522–1527, 2007.

[24] T. A. Clark, C. W. Sugnet, and M. Ares Jr., "Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays," *Science*, vol. 296, no. 5569, pp. 907–910, 2002.

[25] J. A. Pleiss, G. B. Whitworth, M. Bergkessel, and C. Guthrie, "Rapid, transcript-specific changes in splicing in response to environmental stress," *Molecular Cell*, vol. 27, no. 6, pp. 928–937, 2007.

[26] J. A. Pleiss, G. B. Whitworth, M. Bergkessel, and C. Guthrie, "Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components.," *PLoS biology*, vol. 5, article e90, 2007.

[27] F. C. P. Holstege, E. G. Jennings, J. J. Wyrick et al., "Dissecting the regulatory circuitry of a eukaryotic genome," *Cell*, vol. 95, no. 5, pp. 717–728, 1998.

[28] F. Miura, N. Kawaguchi, M. Yoshida et al., "Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs," *BMC Genomics*, vol. 9, article 574, 2008.

[29] S. Ghaemmaghami, W. K. Huh, K. Bower et al., "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003.

[30] V. L. MacKay, X. Li, M. R. Flory et al., "Gene expression analyzed by high-resolution state array analysis and quantitative proteomics," *Molecular and Cellular Proteomics*, vol. 3, no. 5, pp. 478–489, 2004.

[31] C. A. Davis, L. Grate, M. Spingola, and M. Ares, "Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in meiotically regulated genes of yeast," *Nucleic Acids Research*, vol. 28, no. 8, pp. 1700–1706, 2000.

[32] M. Spingola, L. Grate, D. Haussler, and A. Manuel Jr., "Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*," *RNA*, vol. 5, no. 2, pp. 221–234, 1999.

[33] M. J. Moore, E. M. Schwartzfarb, P. Silver, and M. C. Yu, "Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mRNA splicing," *Molecular Cell*, vol. 24, no. 6, pp. 903–915, 2006.

[34] M. Irimia, J. L. Rukov, D. Penny, and S. W. Roy, "Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing," *BMC Evolutionary Biology*, vol. 7, article 188, 2007.

[35] G. Ast, "How did alternative splicing evolve?" *Nature Reviews Genetics*, vol. 5, no. 10, pp. 773–782, 2004.

[36] H. Shen and M. R. Green, "RS domains contact splicing signals and promote splicing by a common mechanism in yeast through humans," *Genes and Development*, vol. 20, no. 13, pp. 1755–1765, 2006.

[37] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, no. 7-8, pp. 563–577, 1999.

[38] X. Xia, *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*, Springer, New York, NY, USA, 2007.

[39] C. L. Zheng, F. U. Xiang-Dong, and M. Gribskov, "Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse," *RNA*, vol. 11, no. 12, pp. 1777–1787, 2005.

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

[41] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[42] X. Xia, *Data analysis in Molecular Biology and Evolution*, Kluwer Academic Publishers, Boston, UK, 2001.

[43] N. Sheth, X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer, and R. Sachidanandam, "Comprehensive splice-site analysis using comparative genomics," *Nucleic Acids Research*, vol. 34, no. 14, pp. 3955–3967, 2006.

[44] S. S. Dwight, R. Balakrishnan, K. R. Christie et al., "Saccharomyces genome database: underlying principles and organisation," *Briefings in Bioinformatics*, vol. 5, no. 1, pp. 9–22, 2004.

[45] S. H. Schwartz, J. Silva, D. Burstein, T. Pupko, E. Eyras, and G. Ast, "Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes," *Genome Research*, vol. 18, no. 1, pp. 88–103, 2008.

[46] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[47] J. Gorodkin, L. J. Heyer, S. Brunak, and G. D. Stormo, "Displaying the information contents of structural RNA alignments: the structure logos," *Computer Applications in the Biosciences*, vol. 13, no. 6, pp. 583–586, 1997.

[48] P. M. Sharp and W. H. Li, "The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[49] X. Xia, "An improved implementation of codon adaptation index," *Evolutionary Bioinformatics*, vol. 3, pp. 53–58, 2007.

[50] P. Rice, L. Longden, and A. Bleasby, "EMBOSS: the European molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[51] A. Coghlan and K. H. Wolfe, "Relationship of codon bias to mRNA and concentration protein length in *Saccharomyces cerevisiae*," *Yeast*, vol. 16, no. 12, pp. 1131–1145, 2000.

[52] L. Duret and D. Mouchiroud, "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4482–4487, 1999.

[53] X. Xia, V. MacKay, X. Yao et al., "Translation initiation: a regulatory role for poly(A) tracts in front of the AUG codon in *Saccharomyces cerevisiae*," *Genetics*, vol. 189, no. 2, pp. 469–478, 2011.

[54] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journals of the Royal Statistical Society B*, vol. 57, pp. 289–300, 1995.

[55] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[56] J. H. Zar, *Biostatistical Analysis*, Prentice Hall, Upper Saddle River, NJ, USA, 4th edition, 1999.

[57] M. Ares Jr., L. Grate, and M. H. Pauling, "A handful of intron-containing genes produces the lion's share of yeast mRNA," *RNA*, vol. 5, no. 9, pp. 1138–1139, 1999.

[58] J. S. Cox, R. E. Chapman, and P. Walter, "The unfolded protein response coordinates the production of endoplasmic reticulum protein and endoplasmic reticulum membrane," *Molecular Biology of the Cell*, vol. 8, no. 9, pp. 1805–1814, 1997.

[59] T. Kawahara, H. Yanagi, T. Yura, and K. Mori, "Endoplasmic reticulum stress-induced mRNA splicing permits synthesis of transcription factor Hac1p/Ern4p that activates the unfolded protein response," *Molecular Biology of the Cell*, vol. 8, no. 10, pp. 1845–1862, 1997.

[60] T. Kawahara, H. Yanagi, T. Yura, and K. Mori, "Unconventional splicing of HAC1/ERN4 mRNA required for the unfolded protein response. Sequence-specific and non-sequential cleavage of the splice sites," *Journal of Biological Chemistry*, vol. 273, no. 3, pp. 1802–1807, 1998.

[61] R. J. Kaufman, "Stress signaling from the lumen of the endoplasmic reticulum: coordination of gene transcriptional and translational controls," *Genes and Development*, vol. 13, no. 10, pp. 1211–1233, 1999.

[62] R. E. Chapman and P. Walter, "Translational attenuation mediated by an mRNA intron," *Current Biology*, vol. 7, no. 11, pp. 850–859, 1997.

[63] C. Sidrauski, J. S. Cox, and P. Walter, "tRNA ligase is required for regulated mRNA splicing in the unfolded protein response," *Cell*, vol. 87, no. 3, pp. 405–413, 1996.

[64] C. Sidrauski and P. Walter, "The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response," *Cell*, vol. 90, no. 6, pp. 1031–1039, 1997.

[65] T. N. Gonzalez, C. Sidrauski, S. Dörfler, and P. Walter, "Mechanism of non-spliceosomal mRNA splicing in the unfolded protein response pathway," *EMBO Journal*, vol. 18, no. 11, pp. 3119–3132, 1999.

[66] T. Nakagawa and R. D. Kolodner, "*Saccharomyces cerevisiae* Mer3 is a DNA helicase involved in meiotic crossing over," *Molecular and Cellular Biology*, vol. 22, no. 10, pp. 3281–3291, 2002.

[67] T. Nakagawa and H. Ogawa, "The *Saccharomyces cerevisiae* MER3 gene, encoding a novel helicase-like protein, is required for crossover control in meiosis," *EMBO Journal*, vol. 18, no. 20, pp. 5714–5723, 1999.

[68] O. M. Mazina, A. V. Mazin, T. Nakagawa, R. D. Kolodner, and S. C. Kowalczykowski, "*Saccharomyces cerevisiae* Mer3 helicase stimulates $3'$-$5'$ heteroduplex extension by Rad51: implications for crossover control in meiotic recombination," *Cell*, vol. 117, no. 1, pp. 47–56, 2004.

[69] T. Nakagawa and R. D. Kolodner, "The MER3 DNA helicase catalyzes the unwinding of holliday junctions," *Journal of Biological Chemistry*, vol. 277, no. 31, pp. 28019–28024, 2002.

[70] T. Nakagawa and H. Ogawa, "Involvement of the MRE2 gene of yeast in formation of meiosis-specific double-strand breaks and crossover recombination through RNA splicing," *Genes to Cells*, vol. 2, no. 1, pp. 65–79, 1997.

[71] H. Tsubouchi and G. S. Roeder, "The Mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair," *Molecular and Cellular Biology*, vol. 22, no. 9, pp. 3078–3088, 2002.

[72] J. Y. Leu, P. R. Chua, and G. S. Roeder, "The meiosis-specific Hop2 protein of *S. cerevisiae* ensures synapsis between homologous chromosomes," *Cell*, vol. 94, no. 3, pp. 375–386, 1998.

[73] F. W. Scherrer Jr. and M. Spingola, "A subset of Mer1p-dependent introns requires Bud13p for splicing activation and nuclear retention," *RNA*, vol. 12, no. 7, pp. 1361–1372, 2006.

[74] M. Spingola and M. Ares Jr., "A yeast intronic splicing enhancer and Nam8p are required for Mer1p-activated splicing," *Molecular Cell*, vol. 6, no. 2, pp. 329–338, 2000.

[75] M. Spingola, J. Armisen, and M. Ares Jr., "Mer1p is a modular splicing factor whose function depends on the conserved U2 snRNP protein Snu17p," *Nucleic Acids Research*, vol. 32, no. 3, pp. 1242–1250, 2004.

[76] J. Parenteau, M. Durand, S. Véronneau et al., "Deletion of many yeast introns reveals a minority of genes that require splicing for function," *Molecular Biology of the Cell*, vol. 19, no. 5, pp. 1932–1941, 2008.

[77] J. Bhargava, J. Engebrecht, and G. S. Roeder, "The rec102 mutant of yeast is defective in meiotic recombination and chromosome synapsis," *Genetics*, vol. 130, no. 1, pp. 59–69, 1992.

[78] K. Jiao, L. Salem, and R. Malone, "Support for a meiotic recombination initiation complex: interactions among Rec102p, Rec104p, and Spo11p," *Molecular and Cellular Biology*, vol. 23, no. 16, pp. 5928–5938, 2003.

[79] R. E. Malone, S. Bullard, M. Hermiston, R. Rieger, M. Cool, and A. Galbraith, "Isolation of mutants defective in early steps of meiotic recombination in the yeast *Saccharomyces cerevisiae*," *Genetics*, vol. 128, no. 1, pp. 79–88, 1991.

[80] S. Maleki, M. J. Neale, C. Arora, K. A. Henderson, and S. Keeney, "Interactions between Mei4, Rec114, and other proteins required for meiotic DNA double-strand break formation in *Saccharomyces cerevisiae*," *Chromosoma*, vol. 116, no. 5, pp. 471–486, 2007.

[81] L. B. Crotti and D. S. Horowitz, "Exon sequences at the splice junctions affect splicing fidelity and alternative splicing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 45, pp. 18954–18959, 2009.

[82] F. Miura, N. Kawaguchi, J. Sese et al., "A large-scale full-length cDNA analysis to explore the budding yeast transcriptome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 47, pp. 17846–17851, 2006.

[83] P. Cliften, P. Sudarsanam, A. Desikan et al., “Finding functional features in Saccharomyces genomes by phylogenetic footprinting,” *Science*, vol. 301, no. 5629, pp. 71–76, 2003.

[84] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, “Sequencing and comparison of yeast species to identify genes and regulatory elements,” *Nature*, vol. 423, no. 6937, pp. 241–254, 2003.

[85] P. F. Cliften, L. W. Hillier, L. Fulton et al., “Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis,” *Genome Research*, vol. 11, no. 7, pp. 1175–1186, 2001.

[86] B. Patterson and C. Guthrie, “A U-rich tract enhances usage of an alternative 3′ splice site in yeast,” *Cell*, vol. 64, no. 1, pp. 181–187, 1991.

[87] R. Parker and B. Patterson, “Architecture of fungal introns: implications for spliceosome assembly,” in *Molecular Biology of RNA, New Perspectives*, M. Inouye and B. Dudock, Eds., pp. 133–149, Academic Press, New York, NY, USA, 1987.

[88] B. Kunze, T. Hellwig-Bürgel, D. Weichenhan, and W. Traut, “Transcription and proper splicing of a mammalian gene in yeast,” *Gene*, vol. 246, no. 1-2, pp. 93–102, 2000.

[89] C. M. Romfo and J. A. Wise, “Both the polypyrimidine tract and the 3’ splice site function prior to the first step of splicing in fission yeast,” *Nucleic Acids Research*, vol. 25, no. 22, pp. 4658–4665, 1997.

[90] V. Sridharan and R. Singh, “A conditional role of U2AF in splicing of introns with unconventional polypyrimidine tracts,” *Molecular and Cellular Biology*, vol. 27, no. 20, pp. 7334–7344, 2007.

[91] X. Xia, “How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*?” *Genetics*, vol. 149, no. 1, pp. 37–44, 1998.

[92] C. G. Simpson, G. Clark, D. Davidson, P. Smith, and J. W. S. Brown, “Mutation of putative branchpoint consensus sequences in plant introns reduces splicing efficiency,” *Plant Journal*, vol. 9, no. 3, pp. 369–380, 1996.

[93] M. D. Chiara, O. R. Gozani, M. Bennett, P. Champion-Arnaud, L. Palandjian, and R. Reed, “Identification of proteins that interact with exon sequences, splice sites, and the branchpoint sequence during each stage of spliceosome assembly,” *Molecular and Cellular Biology*, vol. 16, no. 7, pp. 3317–3326, 1996.

[94] R. Reed and T. Maniatis, “The role of the mammalian branchpoint sequence in pre-mRNA splicing,” *Genes and development*, vol. 2, no. 10, pp. 1268–1276, 1988.

[95] A. Cellini, E. Felder, and J. J. Rossi, “Yeast pre-messenger RNA splicing efficiency depends on critical spacing requirements between the branch point and 3’ splice site,” *EMBO journal*, vol. 5, no. 5, pp. 1023–1030, 1986.

[96] A. Brys and B. Schwer, “Requirement for SLU7 in yeast pre-mRNA splicing is dictated by the distance between the branchpoint and the 3’ splice site,” *RNA*, vol. 2, no. 7, pp. 707–717, 1996.

[97] B. G. M. Luukkonen and B. Séraphin, “The role of branchpoint-3’ splice site spacing and interaction between intron terminal nucleotides in 3’ splice site selection in *Saccharomyces cerevisiae*,” *EMBO Journal*, vol. 16, no. 4, pp. 779–792, 1997.